Università di Torino

QUADERNI
DIDATTICI
del
Dipartimento di Matematica

MARIA GARETTO

STATISTICA
Lezioni ed esercizi

Corso di Laurea in Biotecnologie
A.A. 2002/2003

Quaderno # 13 – Novembre 2002

Prefazione

In questo quaderno sono state raccolte le lezioni del corso di Metodi Matematici e Statistici per il primo anno del Corso di Studi in Biotecnologie dell'Università di Torino. Una parte del materiale è stata anche utilizzata per un corso di Statistica per il primo anno del Corso di Studi in Ingegneria del Politecnico di Torino.
Entrambi i corsi sopra citati si inquadrano nei nuovi corsi di studi triennali, nei quali le nuove esigenze didattiche richiedono di privilegiare l'aspetto operativo piuttosto che l'eccessivo approfondimento teorico; si è scelto quindi di fornire un'introduzione elementare e abbastanza sintetica ai principali argomenti di un corso di statistica di base, accompagnando ogni argomento con numerosi esempi, ma sacrificando sia la maggior parte delle dimostrazioni dei risultati teorici, sia alcuni argomenti, pur di rilevante importanza.
La statistica descrittiva è trattata come primo argomento; lo scopo è quello di introdurre i metodi di analisi dei dati, i principali tipi di grafici, il concetto di variabile, che sarà poi sviluppato con la definizione di variabile aleatoria, le definizioni delle più importanti statistiche e le nozioni di correlazione e regressione da un punto di vista elementare.
Vengono poi introdotti i concetti di base del calcolo delle probabilità, con un breve cenno al calcolo combinatorio.
Molti fra gli esercizi riguardanti il calcolo delle probabilità possono essere risolti senza ricorrere alle tecniche del calcolo combinatorio; questo argomento può perciò essere considerato facoltativo e gli esercizi che lo richiedono sono indicati con un asterisco.
Particolare importanza viene data allo studio delle distribuzioni di probabilità discrete e continue e dei loro parametri e vengono introdotti i modelli fondamentali: la distribuzione binomiale, la distribuzione di Poisson e la distribuzione normale; nell'ambito della statistica inferenziale vengono anche introdotte le distribuzioni *t*, χ^2 e *F*.
La parte dedicata alla statistica inferenziale è preceduta da una breve trattazione delle distribuzioni di campionamento; anche in questo caso si è scelto di non dedicare troppo spazio ai risultati teorici e di concentrare invece l'attenzione sugli intervalli di confidenza e sui test di ipotesi in numerosi casi importanti; sono trattati i vari tipi di test di uso più comune, accompagnati da molte applicazioni.
Vengono infine descritti il test chi-quadro di adattamento e il test chi-quadro di indipendenza, frequentemente utilizzati nelle applicazioni.
Il testo, come i corsi a cui è destinato, è costruito come una successione di lezioni ed esercitazioni e gli argomenti teorici sono sempre seguiti da numerosi esempi, che illustrano la teoria esposta; gli esempi sono sviluppati nei dettagli, riportando tutti i calcoli, le tabelle e i grafici: lo svolgimento a volte un po' noioso e ripetitivo può aiutare lo studente ad acquisire la capacità di risolvere correttamente i problemi.
Il corso di Metodi Matematici e Statistici è accompagnato da un ciclo di esercitazioni di laboratorio in aula informatica, nelle quali viene illustrato l'utilizzo del foglio elettronico Excel; anche se Excel non è un software specificamente destinato alla statistica, tuttavia contiene molte funzioni e strumenti che consentono di effettuare analisi e calcoli statistici e la sua grande diffusione ha motivato la scelta di questo software.
Il materiale utilizzato per lo svolgimento del laboratorio farà parte di un altro quaderno di questa collana.
Per la realizzazione dei grafici presentati in questo testo e per la stesura delle tavole riportate in Appendice è stato utilizzato il software scientifico Matlab, che dispone di un toolbox specificamente destinato alla statistica; questo software offre potenzialità grafiche e di calcolo numerico e simbolico molto superiori a Excel, ma non si presta a un immediato utilizzo per un'attività di laboratorio di breve durata.

Indice

Introduzione

Per **statistica** si intendeva in origine la raccolta di dati demografici ed economici di vitale interesse per lo stato. Da quel modesto inizio essa si è sviluppata in un metodo scientifico di analisi ora applicato a molte scienze, sociali, naturali, mediche, ingegneristiche, ed è uno dei rami più importanti della matematica.
Come esempio di indagine statistica si consideri il seguente problema.
Prima di ogni elezione gli exit poll tentano di individuare quale sarà la proporzione della popolazione che voterà per ciascuna lista: ovviamente non è possibile intervistare tutti i votanti e quindi si sceglie come unica alternativa un **campione** di qualche migliaia di unità, nella speranza che la proporzione campionaria sia una buona stima della proporzione relativa alla popolazione totale.
Per ottenere un risultato sicuro sulla popolazione si dovrebbe aspettare fino alla conclusione dell'elezione, quando siano stati computati tutti i voti, ma questo non costituirebbe più una previsione.
Però, se il campionamento è compiuto correttamente e con metodi adeguati, si possono avere forti speranze che la proporzione campionaria sarà circa uguale alla corrispondente proporzione della popolazione.
Questo ci consente di stimare la proporzione incognita P dell'intera popolazione mediante la proporzione p del campione osservato

$$P = p \pm e$$

dove e indica un errore.
La stima non è fatta con certezza; si deve cioè ammettere la possibilità di essere incorsi in un errore, poiché può essere stato scelto un campione non rappresentativo, eventualità possibile, anche se improbabile: in tale circostanza la conclusione potrebbe essere errata; si può perciò avere soltanto un certo grado di fiducia nelle conclusioni.
Le conclusioni statistiche dunque sono sempre accompagnate da un certo grado di incertezza.
Si noti che l'affermazione che la proporzione della popolazione può essere indotta dalla proporzione del campione, si basa su una deduzione a priori, cioè che la proporzione campionaria molto probabilmente è vicina alla proporzione della popolazione.
L'esempio dell'exit poll rappresenta un tipico esempio di **statistica inferenziale**: le caratteristiche della popolazione complessiva sono indotte da quelle osservate su un campione estratto dalla popolazione stessa.
Altri esempi di indagine statistica possono essere: il censimento della popolazione italiana fatto dall'ISTAT, lo studio di campioni di pezzi prodotti da un'azienda per il controllo della qualità media del prodotto, la sperimentazione di un nuovo farmaco su un gruppo di persone volontarie.
La statistica si può dunque vedere come lo studio delle popolazioni, lo studio della variazione fra gli individui della popolazione, lo studio dei metodi di riduzione dei dati.
Le **popolazioni** di cui si occupa la statistica non sono solo le popolazioni umane, come l'esempio precedente potrebbe far pensare. Le popolazioni sono intese come aggregati di individui non necessariamente viventi o materiali: ad esempio, se si effettua un certo numero di misure, l'insieme dei risultati costituisce una popolazione di misure.
Le popolazioni che sono oggetto di studio statistico evidenziano sempre delle variazioni al loro interno, ossia gli individui che le costituiscono non sono tutti identici: compito della statistica è lo studio di tali variazioni.
All'origine di queste variazioni sono spesso fenomeni aleatori, dove per aleatorio si intende un fenomeno in cui è presente in modo essenziale un elemento di casualità. Questo significa che il fenomeno non è completamente prevedibile a priori, il che non vuol dire che sia completamente imprevedibile. Ad esempio se si estrae una pallina da un'urna che contiene 30 palline bianche e 20 nere, non siamo certi del risultato, ma abbiamo una certa aspettativa.

Occorre quindi studiare il **calcolo delle probabilità**, che, oltre a essere utile per se stesso, ad esempio nella teoria dei giochi, costituisce anche una base per l'inferenza statistica.
Per mezzo del calcolo delle probabilità si può fare una trattazione matematica dell'incertezza, ossia delle regole con cui si può dare un certo grado di fiducia al realizzarsi di un dato evento; in molte situazioni concrete si può formulare un modello probabilistico in base al quale calcolare la probabilità di un certo evento.
Ad esempio, riferendosi al caso dell'urna contenente palline bianche e nere, si potrà calcolare la probabilità che, estraendo 5 palline, 3 siano bianche.
Le conclusioni che la **statistica inferenziale** ci permette di trarre sulla popolazione complessiva a partire dall'indagine sul campione, non sono certezze, come già osservato, ma asserzioni formulate con i metodi, precisi e quantitativi, del calcolo delle probabilità.
La **statistica descrittiva** si occupa invece dell'analisi dei dati osservati, prescindendo sia da qualsiasi modello probabilistico che descriva il fenomeno in esame, sia dal fatto che l'insieme dei dati sia un campione estratto da una popolazione più vasta o sia invece l'intera popolazione.
Lo scopo basilare della statistica descrittiva è di ridurre il volume dei dati osservati, esprimendo l'informazione rilevante contenuta in tali dati per mezzo di grafici e indicatori numerici che li descrivono; inoltre possono essere fatte indagini di tipo comparativo e si può verificare l'adattarsi dei dati sperimentali a un certo modello teorico.

1. Statistica descrittiva

1.1 Distribuzioni di frequenza

Quando si raccolgono dei dati su una popolazione o su un campione, i valori ottenuti si presentano allo statistico come un insieme di dati disordinati; i dati che non sono stati organizzati, sintetizzati o elaborati in qualche modo sono chiamati **dati grezzi**. A meno che il numero delle osservazioni sia piccolo, è improbabile che i dati grezzi forniscano qualche informazione finché non siano stati ordinati in qualche modo.
In questo capitolo verranno descritte alcune tecniche per organizzare e sintetizzare i dati in modo da poter evidenziare le loro caratteristiche importanti e individuare le informazioni da essi fornite. In questo contesto non è importante se tali dati costituiscono l'intera popolazione o un campione estratto da essa. Consideriamo i seguenti esempi.

Esempio 1
Rilevando con uno strumento di misurazione il numero di particelle cosmiche in 40 periodi consecutivi di un minuto si ottengono i seguenti dati

0	2	1	4	3	1	2	3	8	2	5	2	1	3	3	1	3	2	2	5
4	4	4	2	3	5	5	1	1	2	4	4	2	3	3	3	3	3	3	2

Tabella 1

Esempio 2
I seguenti dati sono il risultato di 80 determinazioni, in una data unità di misura, dell'emissione giornaliera di un gas inquinante da un impianto industriale

15.8	26.4	17.3	11.2	23.9	24.8	18.7	13.9	9.0	13.2
22.7	9.8	6.2	14.7	17.5	26.1	12.8	28.6	17.6	23.7
26.8	22.7	18.0	20.5	11.0	20.9	15.5	19.4	16.7	10.7
19.1	15.2	22.9	26.6	20.4	21.4	19.2	21.6	16.9	19.0
18.5	23.0	24.6	20.1	16.2	18.0	7.7	13.5	23.5	14.5
14.4	29.6	19.4	17.0	20.8	24.3	22.5	24.6	18.4	18.1
8.3	21.9	12.3	22.3	13.3	11.8	19.3	20.0	25.7	31.8
25.9	10.5	15.9	27.5	18.1	17.9	9.4	24.1	20.1	28.5

Tabella 2

Esempio 3
In uno stabilimento vengono registrati i casi di malfunzionamento di una macchina utensile controllata dal computer, e le loro cause. I dati relativi a un certo mese sono i seguenti

fluttuazioni di tensione	6
instabilità del sistema di controllo	22
errore dell'operatore	13
strumento usurato e non sostituito	2
altre cause	5
Totale	48

Tabella 3

In ciascuno degli esempi si osserva una variabile, che è rispettivamente
1 – il numero di particelle rilevate in un intervallo di un minuto;
2 – la quantità di gas inquinante emesso in un giorno;
3 – la causa di un guasto verificato.

Della variabile in questione abbiamo un insieme di *n* osservazioni registrate (negli esempi *n* vale, rispettivamente, 40, 80, 48), che costituiscono i dati da analizzare.
Le variabili oggetto di rilevazioni statistiche si classificano in più tipi diversi, a seconda del tipo di valori che assumono

$$\text{variabili} \begin{cases} \text{numeriche (quantitative)} \begin{cases} \text{discrete} \\ \text{continue} \end{cases} \\ \text{non numeriche (qualitative)} \end{cases}$$

Una **variabile** si dice **numerica** se i valori che essa assume sono numeri, **non numerica** altrimenti; una variabile numerica si dice **discreta** se l'insieme dei valori che essa a priori può assumere è finito o numerabile[1], **continua** se l'insieme dei valori che essa a priori può assumere è l'insieme **R** dei numeri reali o un intervallo I di numeri reali.
Le variabili degli esempi 1 e 2 sono numeriche, la variabile dell'esempio 3 è non numerica. La variabile dell'esempio 1 è discreta, perché il numero di particelle osservate è sempre un numero intero maggiore o uguale a 0, e l'insieme dei numeri interi è infinito ma numerabile; la variabile dell'esempio 2 è invece continua, perché la misura della quantità di gas emesso può essere un numero reale positivo qualunque (in un certo intervallo). Molto spesso i valori assunti da una variabile continua sono risultati di misure.
Si osservi che, per decidere se una variabile è discreta o continua, occorre ragionare su quali sono i valori che a priori la variabile può assumere e non sui valori effettivamente assunti: è evidente infatti che i valori assunti in *n* osservazioni saranno al più *n*, quindi sempre in numero finito.
Per studiare i dati degli esempi precedenti dividiamo i dati stessi in **classi** e determiniamo il numero di individui appartenenti a ciascuna classe, detto **frequenza della classe**. Costruiamo poi la **tabella di distribuzione di frequenza**, ossia una tabella che raccoglie i dati secondo le classi e le corrispondenti frequenze.
I dati ordinati e riassunti nella tabella di distribuzione di frequenza sono detti **dati raggruppati**.

Esempio 4 – Variabili numeriche discrete
Nell'esempio 1 la variabile *x* osservata è una variabile numerica discreta, che può assumere solo valori interi; poiché i valori assunti sono i numeri interi 0, 1, 2, 3, 4, 5, 8, è naturale scegliere come classi i numeri $k = 0, 1, 2, 3, 4, 5, 6, 7, 8$ e contare per ogni classe il numero di osservazioni in cui sono state rilevate esattamente *k* particelle. In questo modo si costruisce la seguente tabella di distribuzione di frequenza.
Nella tabella la prima colonna indica la **classe**; la seconda la **frequenza assoluta**, detta anche semplicemente **frequenza di classe**, ossia il numero di osservazioni che cadono in ciascuna classe; la terza colonna la **frequenza relativa**, ossia il rapporto tra frequenza assoluta e numero totale di osservazioni (in questo caso 40); la quarta è la **frequenza percentuale**, ossia la frequenza relativa moltiplicata per 100.

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
0	1	0.025	2.5%
1	6	0.15	15%
2	10	0.25	25%
3	12	0.3	30%
4	6	0.15	15%
5	4	0.1	10%
6	0	0	0%
7	0	0	0%
8	1	0.025	2.5%
Totale	40	1	100%

Tabella 4

[1] Ricordiamo che un **insieme numerabile** è un insieme che si può mettere in corrispondenza biunivoca con l'insieme **N** dei numeri naturali.

Osservazione
Si osservino le seguenti proprietà dei numeri riportati nella tabella di distribuzione di frequenza (tabella 4): la frequenza assoluta è un numero intero compreso tra 0 e il numero totale di osservazioni; la frequenza relativa è un numero reale compreso tra 0 e 1; la frequenza percentuale è un numero reale compreso tra 0 e 100.
La somma delle frequenze assolute è sempre uguale al numero totale di osservazioni; la somma delle frequenze relative è sempre uguale a 1; la somma delle frequenze percentuali è uguale a 100; i valori ottenuti come quozienti devono essere spesso arrotondati e questo fatto comporta che la somma di tutte le percentuali può non essere esattamente uguale a 100.

Esempio 5 – Variabili numeriche continue
Nell'esempio 2 la variabile osservata è continua. I valori dei dati sono compresi tra 6.2 e 31.8; il **campo di variazione *R*** o **range** dei dati, cioè la differenza tra il più grande e il più piccolo, vale

$$R = 31.8 - 6.2 = 25.6$$

Scegliamo come classi i 7 intervalli

$$5.0 \le x \le 8.9$$
$$9.0 \le x \le 12.9$$
$$13.0 \le x \le 16.9$$
$$17.0 \le x \le 20.9$$
$$21.0 \le x \le 24.9$$
$$25.0 \le x \le 28.9$$
$$29.0 \le x \le 32.9$$

Il modo di scegliere le classi non è unico: potremmo scegliere un numero differente di classi, o classi con estremi diversi; in ogni caso le classi non devono sovrapporsi e devono contenere tutti i dati. Di solito le classi hanno tutte la stessa ampiezza, ma questa caratteristica in generale non è obbligatoria e in certi casi il tipo di dati può suggerire la scelta di classi di ampiezza diversa (si vedano gli esempi 8 e 9); inoltre, per dati continui, è necessario specificare se le classi sono chiuse a destra e/o a sinistra, ossia se i dati coincidenti con gli estremi della classe devono essere raggruppati nella classe stessa o in una delle classi adiacenti.
Troppe classi rendono la tabella poco leggibile; troppo poche classi la rendono poco significativa: il numero delle classi è normalmente compreso fra 5 e 15; se i dati sono molto numerosi si può arrivare a usare un massimo di 20 classi. Una semplice regola pratica che si rivela a volte utile consiste nello scegliere un numero di classi approssimativamente uguale alla radice quadrata del numero dei dati

$$k \cong \sqrt{n}\,.$$

Un'altra regola consiste nell'applicare la seguente formula

$$k \cong 1 + 3.322 \cdot log_{10}\, n$$

dove n rappresenta il numero dei dati presi in considerazione e k il numero delle classi da usare.
L'ampiezza delle classi (nel caso di classi di uguale ampiezza) può essere determinata applicando la formula

$$a \cong \frac{R}{k}$$

dove R è il campo di variazione dei dati.
Le risposte ottenute applicando queste formule devono essere comunque interpretate come indicazioni di massima, da valutare caso per caso, a seconda dei dati da trattare.
Nell'esempio che stiamo esaminando si ha

$$k = 1 + 3.322 \cdot log_{10}\, 80 \cong 7$$

$$a = \frac{25.6}{7} \cong 3.7$$

Si giustifica così la scelta di 7 classi di ampiezza 4.

Una scrittura del tipo $5.0 \le x \le 8.9$, definente una classe, è detta **intervallo della classe**; i numeri 5.0 e 8.9 sono detti **limiti inferiore e superiore della classe**.
Con la scelta delle 7 classi indicate si ottiene la tabella seguente

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$5.0 \le x \le 8.9$	3	0.0375	3.75%
$9.0 \le x \le 12.9$	10	0.1250	12.5%
$13.0 \le x \le 16.9$	14	0.1750	17.5%
$17.0 \le x \le 20.9$	25	0.3125	31.25%
$21.0 \le x \le 24.9$	17	0.2125	21.25%
$25.0 \le x \le 28.9$	9	0.1125	11.25%
$29.0 \le x \le 32.9$	2	0.0250	2.5%
Totale	80	1	100%

Tabella 5

Si noti che le **classi** sono **chiuse** e che i limiti delle classi utilizzate per la tabella precedente sono assegnati con tanti decimali quanti ne possiedono i dati.
Le classi hanno uno "stacco" per evitare ambiguità. Infatti se si scegliessero ad esempio le classi

$5.0 \le x \le 9.0$
$9.0 \le x \le 13.0$
.....

il dato 9.0 potrebbe andare nella prima classe o nella seconda, e così via.
Per evitare questa difficoltà si potrebbero scegliere le classi

$4.95 \le x \le 8.95$
$8.95 \le x \le 12.95$
$12.95 \le x \le 16.95$
$16.95 \le x \le 20.95$
$20.95 \le x \le 24.95$
$24.95 \le x \le 28.95$
$28.95 \le x \le 32.95$

Si può notare che anche se i limiti delle classi si sovrappongono, non ci sono ambiguità, perché questi limiti sono valori che i dati non assumono, dal momento che i dati hanno un solo decimale. Questa scelta però non è particolarmente felice, in quanto l'uso di più decimali appesantisce la scrittura delle classi.
E ' più consigliabile scegliere **classi chiuse a sinistra (aperte a destra)**, ad esempio

$5 \le x < 9$
$9 \le x < 13$
.....
$29 \le x < 33$

oppure **classi chiuse a destra (aperte a sinistra)**, ad esempio

$5 < x \le 9$
$9 < x \le 13$
.....
$29 < x \le 33$

Si noti che queste classi non presentano "stacchi".
Con la scelta delle classi chiuse a sinistra sopra indicate, per la distribuzione di frequenza si ottiene una distribuzione di frequenza uguale a quella della tabella 5 (cambiano solo gli estremi delle classi, ma non le frequenze assolute). Invece con la scelta delle classi chiuse a destra si ottiene la distribuzione di frequenza della tabella 5b.

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$5 < x \le 9$	4	0.05	5%
$9 < x \le 13$	9	0.1125	11.25%
$13 < x \le 17$	15	0.1875	18.75%
$17 < x \le 21$	24	0.3	30%
$21 < x \le 25$	17	0.2125	21.25%
$25 < x \le 29$	9	0.1125	11.25%
$29 < x \le 33$	2	0.0250	2.5%
Totale	80	1	100%

Tabella 5b

Una volta che i dati sono stati raggruppati, ciascun valore esatto dei dati non è più utilizzato: si rappresentano tutti i dati appartenenti a una certa classe con il suo punto medio, detto **valore centrale della classe**.
Per ciascuna delle scelte proposte per le classi in questo esempio, le classi hanno la stessa **ampiezza**, uguale a 4; tale ampiezza è in generale uguale alla differenza tra due valori centrali successivi; nel caso delle classi senza stacchi, chiuse da un lato, l'ampiezza è più semplicemente uguale alla differenza tra gli estremi di ogni classe.
Con i dati dell'esempio 5 e con la scelta delle classi chiuse a destra (tabella 5b) si ottiene
a – valori centrali delle classi

$$\frac{5+9}{2}=7, \qquad \frac{9+13}{2}=11,$$
$$15, \quad 19, \quad 23, \quad 27, \quad 31$$

b – ampiezza di classe

$$a = 9 - 5 = 4 \quad \text{oppure} \quad a = 11 - 7 = 4.$$

Il procedimento di raggruppamento dei dati fa perdere alcune delle informazioni che provengono dai dati: ad esempio invece di conoscere l'esatto valore di un'osservazione, si sa solo che cade in un certo intervallo. Ciò accade per la distribuzione di frequenza di ogni variabile continua. Tuttavia si trae un importante vantaggio dalla "leggibilità" che si ottiene e dalle relazioni fra i dati che si rendono evidenti.
Nel caso della variabile discreta dell'esempio 4 non vi è perdita di informazione, in quanto le classi tengono conto di ogni valore assunto. Talvolta però anche per una variabile discreta è conveniente utilizzare come classi degli intervalli, anziché distinguere tutti i valori assunti, soprattutto quando i dati sono numerosi (si veda anche l'esempio 8).
Con i dati dell'esempio 1 si possono usare classi comprendenti due possibili valori della variabile osservata, ottenendo la seguente tabella di distribuzione di frequenza

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$0 \le x \le 1$	7	0.175	17.5%
$2 \le x \le 3$	22	0.55	55%
$4 \le x \le 5$	10	0.25	25%
$6 \le x \le 7$	0	0.0	0%
$8 \le x \le 9$	1	0.025	2.5%
Totale	40	1	100%

Tabella 6

Esempio 6 – Variabili non numeriche

Nell'esempio 3 la variabile "tipo di guasto verificato" è non numerica; i dati sono già raggruppati in classi e si ottiene la seguente tabella di distribuzione di frequenza

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
fluttuazioni tensione	6	0.125	12.5%
instabilità	22	0.458	45.8%
errore operatore	13	0.271	27.1%
strumento	2	0.042	4.2%
altro	5	0.104	10.4%
Totale	48	1	100%

Tabella 7

Ci sono altri modi di raggruppare i dati: ad esempio dati "minori di", "maggiori di"; si ottengono in questo modo le **distribuzioni cumulative**.

La frequenza totale di tutti i valori minori del limite superiore di una data classe è detta **frequenza cumulativa**. Una tabella che presenti frequenze cumulative è detta **tabella di distribuzione cumulativa di frequenza**.

Si possono cumulare frequenze assolute, relative e percentuali; l'ultimo valore che compare nella tabella sarà uguale al numero totale di dati per le frequenze assolute, uguale a 1 per le frequenze relative e uguale a 100 per quelle percentuali.

Nelle tabelle 8 e 9 si riportano le distribuzioni cumulative che si possono ricavare rispettivamente dalle tabelle 4 e 6 (dati dell'esempio 1).

Classe	*Freq. cumul. assoluta*
$x \le 0$	1
$x \le 1$	7
$x \le 2$	17
$x \le 3$	29
$x \le 4$	35
$x \le 5$	39
$x \le 6$	39
$x \le 7$	39
$x \le 8$	40

Tabella 8

Classe	*Freq. cumul. assoluta*
$x \le 1$	7
$x \le 3$	29
$x \le 5$	39
$x \le 7$	39
$x \le 9$	40

Tabella 9

La distribuzione cumulativa ottenibile dalla tabella 5 (esempio 5), è riportata nella tabella 10; se si usano le classi chiuse a destra (tabella 5b) si ottiene la tabella 11.

Classe	*Freq. cumul. assoluta*
$x \le 4.9$	0
$x \le 8.9$	3
$x \le 12.9$	13
$x \le 16.9$	27
$x \le 20.9$	52
$x \le 24.9$	69
$x \le 28.9$	78
$x \le 32.9$	80

Tabella 10

Classe	*Freq. cumul. assoluta*
$x \le 5$	0
$x \le 9$	4
$x \le 13$	13
$x \le 17$	28
$x \le 21$	52
$x \le 25$	69
$x \le 29$	78
$x \le 33$	80

Tabella 11

Esempio 7
Sono date 150 misurazioni del valore di una variabile; la più piccola è 5.18 e la più grande è 7.44. Determinare delle classi adatte per raggruppare i dati in una distribuzione di frequenza, e i corrispondenti valori centrali.

Campo di variazione dei dati

$$R = 7.44 - 5.18 = 2.26$$

Numero di classi e ampiezza delle classi

$$k = 1 + 3.322 \cdot log_{10} 150 \cong 8.23 \qquad a = \frac{2.26}{8} \cong 0.28$$

Si possono utilizzare 8 classi di ampiezza $a = 0.3$. Nella tabella 12 sono indicate le classi scelte e i relativi valori centrali (questa scelta ovviamente non è l'unica possibile).

Classi	*Val. centrali*
$5.1 < x \leq 5.4$	5.25
$5.4 < x \leq 5.7$	5.55
$5.7 < x \leq 6.0$	5.85
$6.0 < x \leq 6.3$	6.15
$6.3 < x \leq 6.6$	6.45
$6.6 < x \leq 6.9$	6.75
$6.9 < x \leq 7.2$	7.05
$7.2 < x \leq 7.5$	7.35

Tabella 12

Esempio 8
I seguenti sono i numeri di lavoratori assenti da un'azienda in 50 giorni lavorativi

13	5	13	37	10	16	2	11	6	12
8	21	12	11	7	7	9	16	49	18
3	11	19	6	15	10	14	10	7	24
11	3	6	10	4	6	32	9	12	7
29	12	9	19	8	20	15	5	17	10

Tabella 13

Per costruire la tabella della distribuzione di frequenza si utilizzano 6 classi; infatti

$$k = 1 + 3.322 \cdot log_{10} 50 \cong 6.6$$

Si noti che in questa tabella è stata usata come ultima classe una classe senza limite superiore, detta **classe aperta**: questo evita di avere classi vuote o con frequenze molto basse.

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$0 < x \leq 4$	4	0.08	8%
$4 < x \leq 9$	15	0.30	30%
$9 < x \leq 14$	16	0.32	32%
$14 < x \leq 19$	8	0.16	16%
$19 < x \leq 24$	3	0.06	6%
$x > 24$	4	0.08	8%
Totale	50	1	100%

Tabella 14

Esempio 9
Nella tabella seguente sono riportati i pesi alla nascita di 100 bambini nati in un ospedale in un dato periodo di tempo.

1640	3340	2600	3060	3740	900	3980	3900	2720	4560
2340	2440	3260	3340	2700	2360	3180	3620	3600	2300
3480	1800	2660	1900	3500	4380	2960	2840	1200	1980
2940	3740	2780	4120	1740	2640	2400	2660	3280	3200
3440	1940	3040	2360	3580	2480	2520	3060	3260	2400
940	2200	3500	2960	3540	2880	3460	3880	2120	2860
2580	3460	4100	2800	3260	2940	2760	2520	2380	1080
2940	2260	1900	2980	4080	2460	2480	2920	3060	980
3620	3000	3540	3060	2780	3760	2940	2360	3500	3100
3780	3260	3600	3820	2520	3440	3180	4100	3260	1800

Tabella 15

Per costruire una distribuzione di frequenza in questo caso si possono usare 8 classi, in base al fatto che

$$k = 1 + 3.322 \cdot log_{10} 100 \cong 7.64$$

e le classi possono essere di ampiezza diversa, per tener conto della natura dei dati.
Il campo di variazione dei dati è

$$R = 4560 - 900 = 3660.$$

I dati possono essere raggruppati nella seguente distribuzione di frequenza

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$800 < x \leq 1400$	5	0.05	5%
$1400 < x \leq 2000$	8	0.08	8%
$2000 < x \leq 2400$	11	0.11	11%
$2400 < x \leq 2800$	18	0.18	18%
$2800 < x \leq 3200$	21	0.21	21%
$3200 < x \leq 3600$	21	0.21	21%
$3600 < x \leq 4000$	10	0.10	10%
$4000 < x \leq 4600$	6	0.06	6%
Totale	100	1	100%

Tabella 16

1.2 Grafici delle distribuzioni di frequenza

Introduciamo alcune delle più usate rappresentazioni grafiche per le distribuzioni di frequenza e per le distribuzioni cumulative. Tali grafici sono oggi solitamente ottenuti con l'uso del computer per mezzo di software di tipo statistico; questi consentono, dopo aver immesso i dati, di ottenere rapidamente i vari tipi di grafici.
L'osservazione del grafico può far notare irregolarità o comportamenti anomali non direttamente osservabili sui dati; ad esempio ci si può accorgere di errori di misurazione.
Un primo tipo di diagramma è il **diagramma circolare**; in questo diagramma le frequenze percentuali sono rappresentate da settori circolari aventi ampiezze proporzionali alle frequenze stesse; indicando con f la frequenza percentuale e con g l'ampiezza in gradi, si ha

$$f : 100 = g : 360°$$

Il diagramma circolare è il più adatto per le frequenze percentuali e per le variabili non numeriche.

Esempio 10
La seguente tabella rappresenta il numero di studenti iscritti ai vari anni di corso di un istituto superiore (frequenze assolute) e le corrispondenti frequenze percentuali; la figura 1 rappresenta il diagramma circolare delle frequenze percentuali.

Studenti iscritti ai diversi anni di corso		
	freq. assoluta	*freq. percentuale*
classi prime	187	19.00%
classi seconde	214	21.75%
classi terze	225	22.87%
classi quarte	176	17.89%
classi quinte	182	18.50%
Totale	984	100.01%

Tabella 17

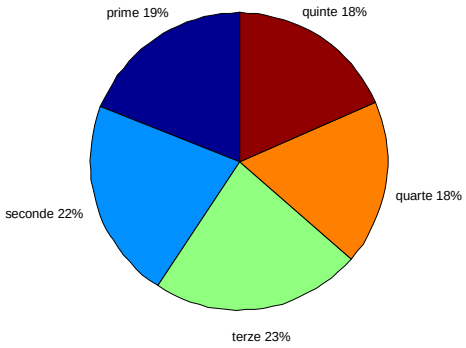


Figura 1

Un altro tipo di grafico molto usato per rappresentare dati raggruppati è il **diagramma a barre**. Per costruire un diagramma a barre si raggruppano i dati in classi, come già descritto; per ciascuna classe si disegna un rettangolo avente base di ampiezza costante e altezza uguale alla frequenza di classe; i rettangoli di solito non sono adiacenti e sono equidistanti fra loro. Questo tipo di diagramma è particolarmente indicato per variabili non numeriche e per variabili discrete.
Il diagramma a barre della distribuzione di frequenza assoluta della tabella 17 è il seguente

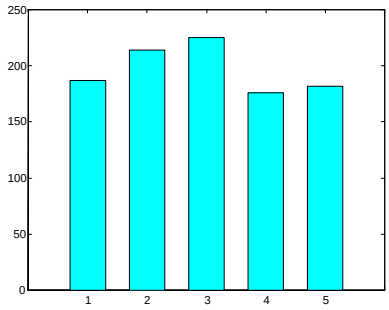


Figura 2

Nel caso della variabile discreta dell'esempio 4, in base alla tabella 4 della distribuzione di frequenza, si può tracciare il diagramma a barre riportato nella figura 3, ottenuto disegnando i rettangoli con le basi centrate nel valore che definisce la classe e riportando in ordinata la frequenza assoluta.

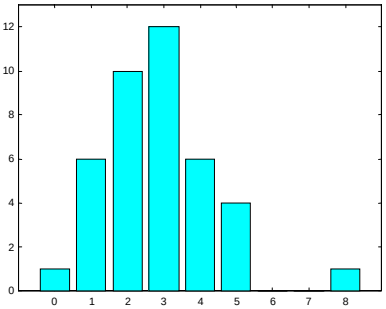


Figura 3

Gli **istogrammi** sono un altro modo molto usato per rappresentare graficamente le informazioni contenute in una tabella di distribuzione di frequenza.
Un istogramma consiste in un insieme di rettangoli adiacenti, aventi base sull'asse orizzontale; le basi sono gli intervalli che definiscono le classi (i punti medi delle basi sono i valori centrali delle classi).
Se le classi hanno tutte la stessa ampiezza le altezze dei rettangoli sono uguali, o proporzionali, alle corrispondenti frequenze assolute (oppure relative o percentuali).
Se invece le classi sono di ampiezza diversa, i rettangoli hanno ancora base uguale alla corrispondente ampiezza della classe, e area (non più altezza!) corrispondente alla frequenza: l'altezza del rettangolo sarà uguale, o proporzionale, al rapporto fra la frequenza e l'ampiezza di classe. Tale rapporto si chiama **densità di frequenza** (vedere figura 4b).
In entrambi i casi quindi l'area di ogni rettangolo è uguale, o proporzionale, alla frequenza della classe.
L'istogramma corrispondente alla distribuzione di frequenza studiata nell'esempio 5 (tabella 5b) è quello della figura 4. Le classi hanno tutte la stessa ampiezza e in ordinata è riportata la frequenza assoluta; le basi dei rettangoli hanno i punti medi nei valori centrali delle classi.
L'istogramma corrispondente alla distribuzione di frequenza dell'esempio 9 è quello della figura 4b; in questo caso le classi non hanno tutte la stessa ampiezza e in ordinata si pone la densità di frequenza (ossia il rapporto fra la frequenza assoluta e l'ampiezza della corrispondente classe).

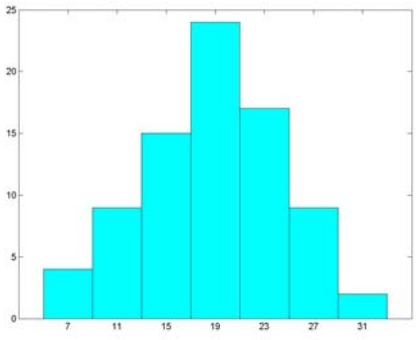


Figura 4

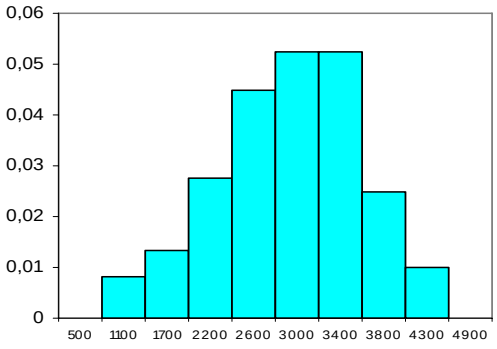


Figura 4b

Una distribuzione di frequenza può essere rappresentata graficamente anche con un altro tipo di grafico: il **poligono di frequenza**. Tale poligono si ottiene unendo fra loro i punti aventi come ascissa il valore centrale di ogni classe e come ordinata il corrispondente valore della frequenza.
Nella figura 5 rappresentiamo il poligono di frequenza per i dati della tabella 5b. La figura 5b riporta il poligono di frequenza sovrapposto all'istogramma della figura 4; questo grafico consente di vedere, per lo stesso insieme di dati, la relazione fra i due tipi di grafico.

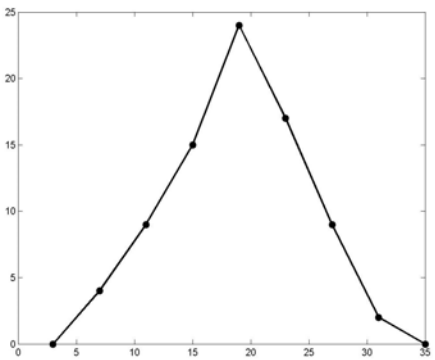


Figura 5

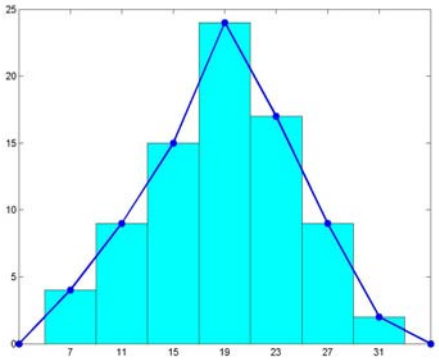


Figura 5b

Esempio 11
Nella tabella 18 sono riportate le lunghezze in mm di 40 sbarrette metalliche; costruire una distribuzione di frequenza assoluta, scegliendo un numero opportuno di classi e disegnare il relativo istogramma.

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

Tabella 18

La lunghezza maggiore è di 176 mm, la minore è di 119 mm; il campo di variazione dei dati è
$R = 176 - 119 = 57$ mm.
Si possono scegliere 7 classi di ampiezza 9 e si ottiene la seguente distribuzione di frequenza assoluta e il corrispondente istogramma

Classe	*Frequenza assoluta*
$118 \le x \le 126$	3
$127 \le x \le 135$	5
$136 \le x \le 144$	9
$145 \le x \le 153$	12
$154 \le x \le 162$	5
$163 \le x \le 171$	4
$172 \le x \le 180$	2
Totale	40

Tabella 19

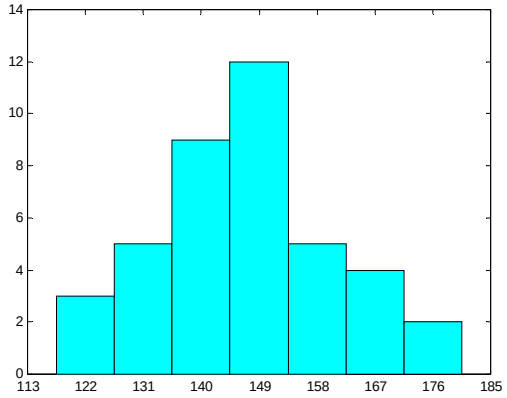


Figura 6

Esempio 12
Si consideri la seguente tabella riepilogativa dei voti finali riportati dagli studenti delle classi terze di un istituto superiore; nella tabella sono riportate due diverse distribuzioni di frequenza assoluta e percentuale relative ai voti finali in italiano e matematica.

voto finale	*studenti che hanno riportato il voto indicato*			
	italiano		***matematica***	
	freq. assoluta	*freq. percentuale*	*freq. assoluta*	*freq. percentuale*
3	10	3.36%	12	4.03%
4	25	8.39%	38	12.75%
5	34	11.41%	35	11.74%
6	136	45.64%	117	39.26%
7	68	22.82%	67	22.48%
8	22	7.38%	26	8.72%
9	3	1.01%	3	1.01%
Totale	298	100.01%	298	99.99%

Tabella 20

Si possono rappresentare le due distribuzioni di frequenza assolute con un unico diagramma a barre, che permette il confronto fra le due distribuzioni ed evidenzia le differenze significative.

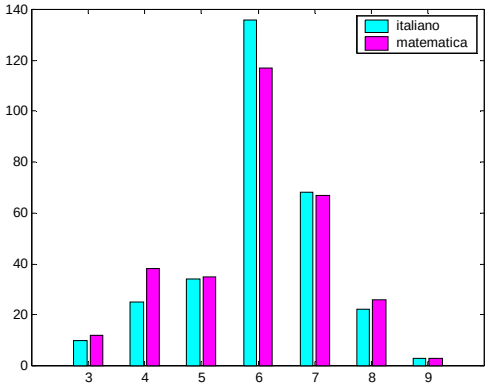


Figura 7

Una distribuzione cumulativa viene rappresentata con un grafico detto **poligono cumulativo** o **ogiva**; il grafico si ottiene riportando sulle ascisse i limiti superiori delle classi e, per ciascuno di essi, in ordinata la frequenza cumulativa della corrispondente classe, e unendo poi tra loro i punti ottenuti.
Per la distribuzione cumulativa di frequenza assoluta dell'esempio 4, tabella 8, si ottiene il grafico della figura 8; per la distribuzione cumulativa di frequenza assoluta dell'esempio 5, tabella 11, si ottiene il grafico della figura 9.

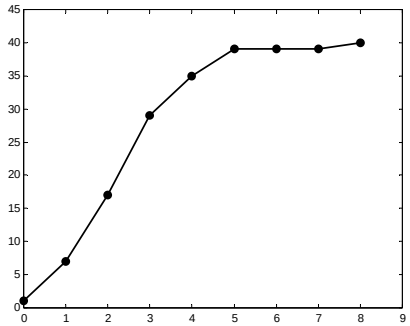


Figura 8

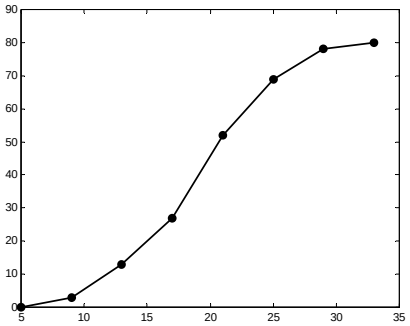


Figura 9

Esempio 13
Riprendendo in esame la tabella relativa all'esempio 12, si costruisce la seguente tabella della distribuzione cumulativa di frequenza assoluta per i voti di italiano

voto finale	*studenti che hanno riportato il voto indicato in italiano*	
	freq. assoluta	*freq. cumulativa assoluta*
$x \leq 3$	10	10
$x \leq 4$	25	35
$x \leq 5$	34	69
$x \leq 6$	136	205
$x \leq 7$	68	273
$x \leq 8$	22	295
$x \leq 9$	3	298

Tabella 21

Dalla tabella 21 si possono ad esempio dedurre i seguenti risultati:

– il numero degli studenti che non hanno la sufficienza in italiano, indicato con $f(x \le 5)$, è uguale alla frequenza cumulata relativa al voto 5, ossia

$$f(x \le 5) = 69 \text{ studenti;}$$

– il numero degli studenti che hanno la sufficienza in italiano, indicato con $f(x \ge 6)$, è uguale al complementare, sul totale, del numero di quelli che non hanno la sufficienza, ossia

$$f(x \ge 6) = 298 - 69 = 229 \text{ studenti.}$$

Il grafico della distribuzione cumulativa di frequenza assoluta è il seguente

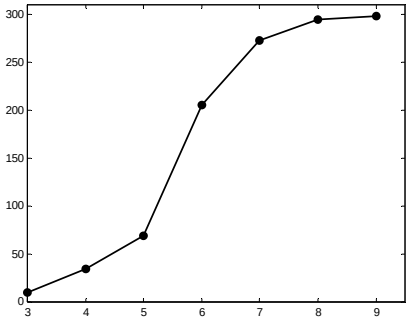


Figura 10

Esempio 14

La tabella 22 riporta la distribuzione dei punteggi ottenuti con 500 lanci di due dadi; il corrispondente istogramma è rappresentato nella figura 11.

punteggio	*freq. assoluta*
2	13
3	35
4	32
5	55
6	74
7	85
8	66
9	56
10	34
11	35
12	15

Tabella 22

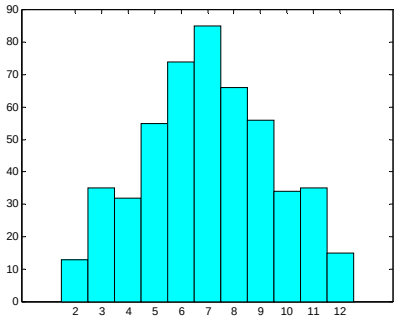


Figura 11

La tabella di distribuzione delle frequenze cumulative assolute e percentuali è la seguente

punteggio	*freq. cumul. assoluta*	*freq. cumul. percentuale*
$x \le 2$	13	2.6%
$x \le 3$	48	9.6%
$x \le 4$	80	16%
$x \le 5$	135	27%
$x \le 6$	209	41.8%
$x \le 7$	294	58.8%
$x \le 8$	360	72%
$x \le 9$	416	83.2%
$x \le 10$	450	90%
$x \le 11$	485	97%
$x \le 12$	500	100%

Tabella 23

Utilizzando la tabella delle frequenze cumulative percentuali si possono calcolare ad esempio le frequenze percentuali dei seguenti risultati

– punteggio minore o uguale a 8: $f(x \le 8) = 72\%$

– punteggio minore di 9: $f(x < 9) = f(x \le 8) = 72\%$

– punteggio compreso fra 4 e 8: $f(4 \le x \le 8) = f(x \le 8) - f(x < 4) = f(x \le 8) - f(x \le 3) =$
$= 72\% - 9.6\% = 62.4\%$

– punteggio maggiore di 7: $f(x > 7) = f(x \le 12) - f(x \le 7) = 100\% - 58.8\% = 41.2\%$

Esempio 15

Sono stati misurati i diametri di 20 sferette prodotte da una linea produttiva; le misure in cm sono date da

2.08	1.72	1.92	1.95	1.89	1.85	1.80	1.84	1.82	1.84
1.93	1.86	2.00	1.80	1.82	2.08	1.90	1.85	2.02	2.00

Tabella 24

Per raggruppare i dati utilizziamo 5 classi, aventi ampiezza uguale a 0.08.
Tabella della distribuzione di frequenza

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$1.69 < x \le 1.77$	1	0.05	5%
$1.77 < x \le 1.85$	8	0.4	40%
$1.85 < x \le 1.93$	5	0.25	25%
$1.93 < x \le 2.01$	3	0.15	15%
$2.01 < x \le 2.09$	3	0.15	15%
Totale	20	1	100%

Tabella 25

Istogramma della distribuzione di frequenza assoluta (figura 12); tabella e grafico della distribuzione cumulativa di frequenza assoluta (tabella 26 e figura 13)

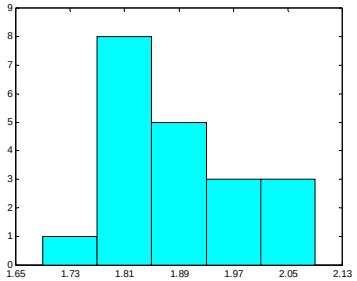


Figura 12

Classe	*Freq. cumul. assoluta*
$x \le 1.69$	0
$x \le 1.77$	1
$x \le 1.85$	9
$x \le 1.93$	14
$x \le 2.01$	17
$x \le 2.09$	20

Tabella 26

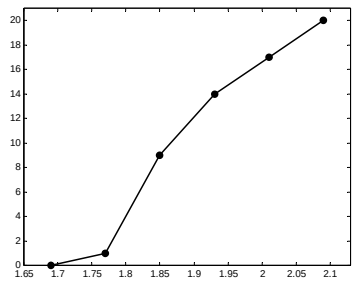


Figura 13

I dati possono anche essere raggruppati scegliendo altre 5 classi, di ampiezza uguale a 0.10; in questo caso si ottengono i seguenti risultati

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$1.64 < x \leq 1.74$	1	0.05	5 %
$1.74 < x \leq 1.84$	6	0.30	30 %
$1.84 < x \leq 1.94$	7	0.35	35 %
$1.94 < x \leq 2.04$	4	0.20	20 %
$2.04 < x \leq 2.14$	2	0.10	10 %
Totale	20	1	100 %

Tabella 27

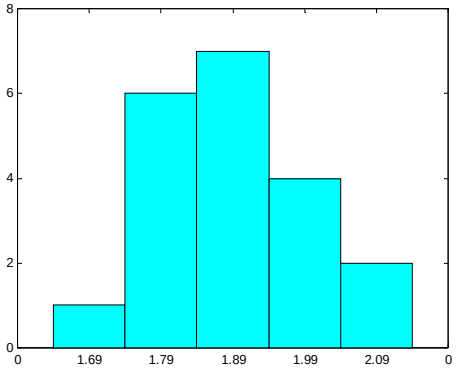


Figura 14

Classe	*Freq. cumul. assoluta*
$x \leq 1.64$	0
$x \leq 1.74$	1
$x \leq 1.84$	7
$x \leq 1.94$	14
$x \leq 2.04$	18
$x \leq 2.14$	20

Tabella 28

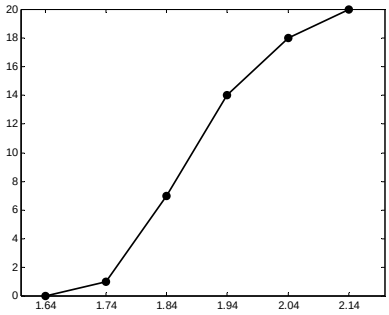


Figura 15

Esempio 16

In uno studio di due settimane sulla qualità della produzione degli operai di un'azienda, si sono ottenuti i dati seguenti, riguardanti il numero totale di pezzi accettabili al controllo qualità, prodotti da 100 operai

65	36	49	84	79	56	28	43	67	36
43	78	37	40	68	72	55	62	22	82
88	50	60	56	57	46	39	57	73	65
59	48	76	74	70	51	40	75	56	45
35	62	52	63	32	80	64	53	74	34
76	60	48	55	51	54	45	44	35	51
21	35	61	45	33	61	77	60	85	68
45	53	34	67	42	69	52	68	52	47
62	65	55	61	73	50	53	59	41	54
41	74	82	58	26	35	47	50	38	70

Tabella 29

Raggruppiamo i dati in una distribuzione di frequenza avente le classi

$20 \le x \le 29$ $30 \le x \le 39$ $40 \le x \le 49$ $50 \le x \le 59$ $60 \le x \le 69$
$70 \le x \le 79$ $80 \le x \le 89$

e disegniamo l'istogramma.
Ricaviamo poi la distribuzione cumulativa di frequenza assoluta e disegniamo l'ogiva.

Tabella della distribuzione di frequenza

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$20 \le x \le 29$	4	0.04	4
$30 \le x \le 39$	13	0.13	13
$40 \le x \le 49$	18	0.18	18
$50 \le x \le 59$	25	0.25	25
$60 \le x \le 69$	20	0.20	20
$70 \le x \le 79$	14	0.14	14
$80 \le x \le 89$	6	0.06	6
Totale	100	1	100

Tabella 30

Istogramma della distribuzione di frequenza assoluta

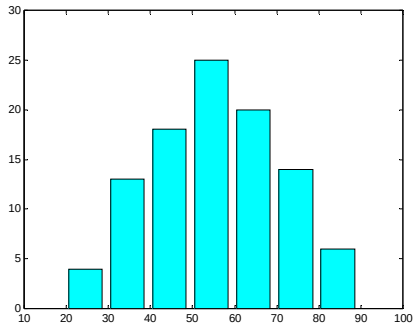


Figura 16

Tabella e grafico della distribuzione cumulativa di frequenza assoluta

Classe	*Freq. cumul. assoluta*
$x \le 19$	0
$x \le 29$	4
$x \le 39$	17
$x \le 49$	35
$x \le 59$	60
$x \le 69$	80
$x \le 79$	94
$x \le 89$	100

Tabella 31

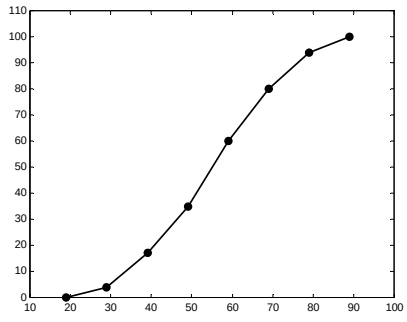


Figura 17

Esempio 17
Cinque monete vengono lanciate 1000 volte contemporaneamente e si osserva ad ogni lancio il numero di teste. Il numero di lanci in cui si sono ottenute 0, 1, 2, 3, 4, 5 teste sono dati dalla tabella seguente

Classe (numero teste)	*Freq. assoluta*
0	38
1	144
2	342
3	287
4	164
5	25
Totale	1000

Tabella 32

Disegniamo l'istogramma della distribuzione di frequenza (figura 18) e costruiamo la tabella (tabella 33) e il grafico della distribuzione cumulativa di frequenza (figura 19).

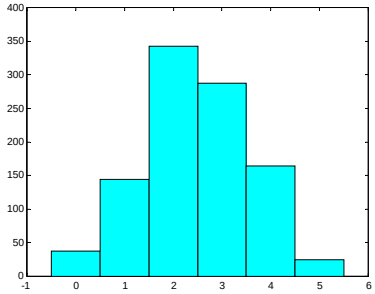


Figura 18

Classe	*Freq. cumul. assoluta*
$x \leq 0$	38
$x \leq 1$	182
$x \leq 2$	524
$x \leq 3$	811
$x \leq 4$	975
$x \leq 5$	1000

Tabella 33

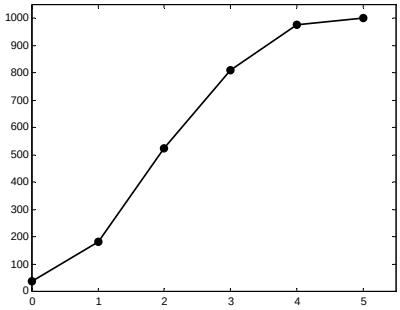


Figura 19

Esempio 18

Nella seguente tabella si riportano i dati riguardanti l'istruzione universitaria in Italia (riferiti all'anno 1996/97).

Disegniamo un diagramma a barre per rappresentare tali dati; rappresentiamo con diagrammi circolari le percentuali di laureati nei vari corsi di laurea calcolate rispetto al numero totale di laureati e le percentuali calcolate rispetto al numero di iscritti in ciascun corso.

corsi di laurea	*studenti in corso*	*studenti fuori corso*	*laureati*
1 – *facoltà scientifiche*	116364	66936	15539
2 – *facoltà di medicina*	50719	21388	7407
3 – *facoltà tecniche*	160106	126158	19099
4 – *facoltà economiche*	278174	179074	35272
5 – *facoltà giuridiche*	193456	125612	18839
6 – *facoltà letterarie*	241824	134622	27128
7 – *diplomi*	62441	16812	9254
Totale	1103084	670602	132538

Tabella 34

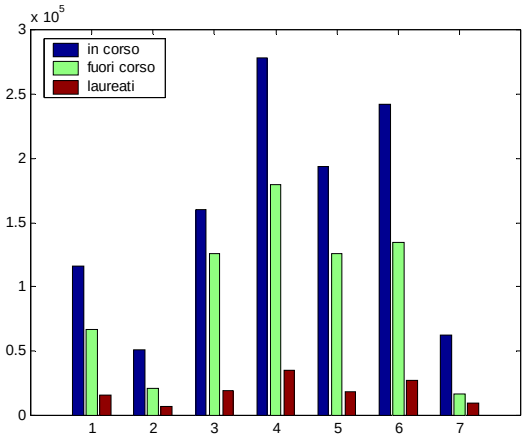


Figura 20

La tabella 35 riporta le percentuali di laureati in ciascun corso di laurea; i dati sono illustrati dal diagramma circolare della figura 21

corsi di laurea	*laureati (freq. assoluta)*	*freq. percentuale*
1 – *facoltà scientifiche*	15539	11.7%
2 – *facoltà di medicina*	7407	5.6%
3 – *facoltà tecniche*	19099	14.4%
4 – *facoltà economiche*	35272	26.6%
5 – *facoltà giuridiche*	18839	14.2%
6 – *facoltà letterarie*	27128	20.5%
7 – *diplomi*	9254	7.0%
Totale	132538	100%

Tabella 35

Nella tabella 36 si riporta per ciascun corso di laurea la percentuale di laureati rispetto al numero di iscritti nel corso stesso; i dati sono illustrati dal diagramma circolare della figura 22

corsi di laurea	*studenti iscritti*	*laureati*	*freq. percentuale*
1 – *facoltà scientifiche*	183300	15539	8.5%
2 – *facoltà di medicina*	72107	7407	10.3%
3 – *facoltà tecniche*	286264	19099	6.7%
4 – *facoltà economiche*	457248	35272	7.7%
5 – *facoltà giuridiche*	319068	18839	6.0%
6 – *facoltà letterarie*	376446	27128	7.2%
7 – *diplomi*	79253	9254	11.7%

Tabella 36

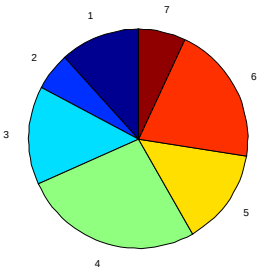


Figura 21

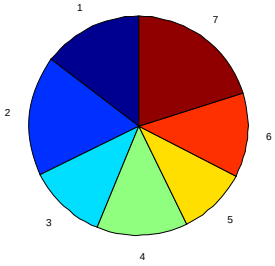


Figura 22

Esempio 19
Nella tabella 37 si riportano le aree dei continenti del mondo, in migliaia di chilometri quadrati; disegniamo il grafico dei dati con un diagramma a barre e con un diagramma circolare.

Continente	*Area (migliaia di Kmq)*
Europa	10368
Asia	45078
Africa	30209
America Sett. e Centr.	24203
America merid.	17855
Oceania	8522
Antartide	14108

Tabella 37

I rettangoli che compongono il diagramma a barre si possono anche disegnare orizzontali, anziché verticali; il diagramma circolare si può anche disegnare in 3 dimensioni.

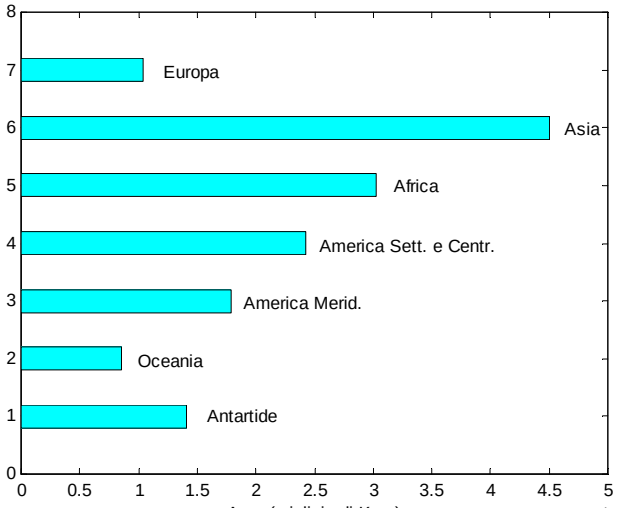


Figura 23

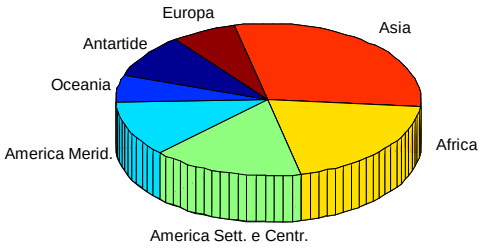


Figura 24

1.3 Indici di posizione e di dispersione

Definiamo alcuni indici numerici, detti anche **statistiche**, utili per descrivere dei dati numerici e la loro distribuzione di frequenza; tali indici prendono il nome di **media**, **mediana**, **moda**, **varianza** e **scarto quadratico medio** o **deviazione standard** e misurano il centro e la dispersione dei dati.
Si osservino i seguenti istogrammi

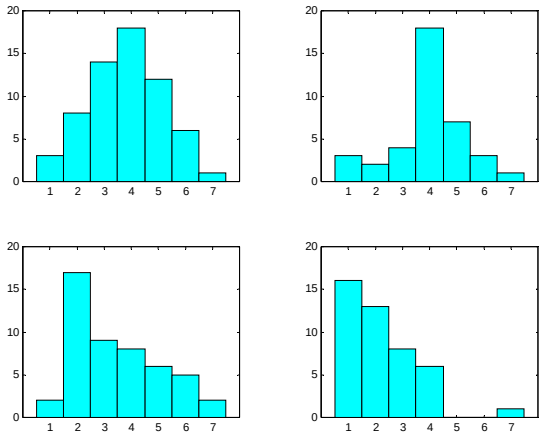


Figura 25

Il primo grafico mostra una distribuzione simmetrica, centrata attorno a 4, valore per cui la frequenza è massima; la seconda distribuzione è ancora centrata attorno a 4, ma per valori lontani da 4 le frequenze sono piccole; la terza distribuzione non è simmetrica, ma ha una coda a destra più lunga che a sinistra; la quarta è decrescente e non simmetrica, con alcuni valori dispersi lontano dagli altri. Gli indici che introdurremo servono per misurare quantitativamente alcune delle caratteristiche osservate qualitativamente su questi grafici esemplificativi.
Si consideri un insieme di n dati $x_1 , x_2 , ... , x_n$.

Definizione 1
Si definisce **media aritmetica** o **media campionaria** di n dati $x_1 , x_2 , ... , x_n$ la quantità

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.1)$$

Per ogni valore x_i della variabile x si definisce lo **scarto dalla media**

$$s_i = x_i - \bar{x}$$

che indica il grado di scostamento del singolo valore x_i dalla media $\bar{x}$.
Si dimostra facilmente che la somma algebrica S degli scarti dalla media è nulla. Infatti

$$S = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} \left(x_i - \bar{x}\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Esempio 20
Media dei dati

$$\bar{x} = \frac{\overset{15}{15} + \overset{14}{14} + \overset{2}{2} + \overset{27}{27} + \overset{13}{13}}{5} = 14.2$$

Definizione 2
La **mediana** M di un insieme di n dati ordinati in ordine di grandezza crescente è il valore centrale dei dati, se il numero di dati è dispari, o la media aritmetica dei due valori centrali, se il numero dei dati è pari.

Questa definizione della mediana assicura che lo stesso numero di dati cade sia a sinistra che a destra della mediana stessa.
L'uso della mediana come indice per descrivere le caratteristiche dei dati ha lo svantaggio di dover prima riordinare i dati in ordine crescente, il che non è richiesto per il calcolo della media.

Esempio 21
a – Mediana dei dati

15 14 2 27 13

Dati ordinati in ordine crescente

2 13 14 15 27

Mediana

$M = 14$

b – Mediana dei dati

11 9 17 19 4 15

Dati ordinati in ordine crescente

4 9 11 15 17 19

Mediana $M = \frac{11+15}{2} = 13$

Un ulteriore indice utile è la moda, denotata con $\tilde{x}$.

Definizione 3
La **moda** $\tilde{x}$ di un insieme di *n* dati è il valore o la classe a cui corrisponde la massima frequenza assoluta.

La moda è per lo più utilizzata quando si trattano dati di tipo qualitativo, per i quali non è possibile calcolare media e mediana.
La moda può non esistere o non essere unica; quando è unica, la distribuzione è detta **unimodale**, quando ci sono più mode diverse è detta **bimodale** o **multimodale**.

Esempio 22
a – Moda dell'insieme di dati

3, 3, 5, 4, 7, 7, 7, 9, 2, 1

L'insieme ha moda $\tilde{x} = 7$.

b – Moda dell'insieme di dati

3, 3, 3, 5, 4, 7, 7, 7, 9, 2, 1

L'insieme ha due mode $\tilde{x} = 3$ e $\tilde{x} = 7$.

c – L'insieme di dati

3, 5, 4, 7, 8, 6, 9, 2, 1

non ha moda, perché ogni dato si presenta una sola volta.

L'ultimo caso mette in rilievo un problema comune con la moda: questo indice non è utile quando i dati sono tanti e per la maggior parte diversi fra loro; in tali casi la moda può non esistere o essere lontana dal centro dell'insieme di dati. Per questa ragione la moda è poco utilizzata.
Media, mediana e moda sono detti **indici di posizione** o **indici di tendenza centrale**, perché descrivono attorno a quale valore è centrato l'insieme di dati.
La mediana è preferibile alla media quando si vogliono eliminare gli effetti di valori estremi molto diversi dagli altri dati: la ragione è che la mediana non utilizza tutti i dati, ma solo il dato centrale o i due dati centrali.
I seguenti esempi mostrano come la mediana in tali casi descriva in modo più adeguato un insieme di dati.
Tuttavia occorre mettere in evidenza che l'utilizzare solo i dati centrali rende la mediana poco sensibile a tutti gli altri valori dei dati e questo può costituire un limite di questo indice.

Esempio 23
Sia dato il seguente insieme di 20 dati, che rappresentano il peso alla nascita (in g) di 20 bambini nati in una settimana in una clinica.

3280	3320	2500	2760
3260	3650	2840	3250
3240	3200	3600	3320
3480	3020	2840	3200
4160	2580	3540	3780

Tabella 38

La media dei dati è

$$\bar{x} = \frac{(3280 + 3320 + ... + 3540 + 3780)}{20} = 3241\,\text{g}$$

Si può osservare che 9 dati sono minori della media e 11 maggiori.
Come già osservato, uno dei limiti della media come misura della tendenza centrale è che essa è molto sensibile ai valori dei dati che cadono agli estremi dell'intervallo di variabilità; in questo senso può non rappresentare bene la collocazione dei dati. Se ad esempio il primo bambino fosse un nato prematuro del peso di 500 g, la media avrebbe il valore

$$\bar{x} = 3102\ \text{g}$$

e in tal caso 7 dati sarebbero minori della media e 13 maggiori.
La mediana in questo caso è

$$M = 3245$$

mentre per l'insieme di dati assegnati inizialmente è

$$M = 3255$$

Esempio 24
In una ditta lavorano 4 giovani ingegneri, che guadagnano € 15.000 all'anno ciascuno, e il proprietario, anch'egli ingegnere, che guadagna € 90.000 all'anno.
Stabilire se la ditta è un buon posto di lavoro per un giovane ingegnere.

Media degli stipendi

$$\bar{x} = \frac{4 \cdot 15.000 + 90.000}{5} = \text{€}\ 30.000$$

Il valore della media sembra indicare che si tratti di un ottimo posto di lavoro.
Mediana degli stipendi

$$M = \text{€}\ 15.000$$

La mediana rappresenta meglio della media quello che guadagna un giovane ingegnere dipendente, quindi il posto di lavoro non è così buono come era stato giudicato con la media.

Esempio 25
I dati seguenti rappresentano i valori dei globuli bianchi (in migliaia) rilevati in 10 pazienti ricoverati in una mattina in un ospedale

7 35 5 9 8 3 10 12 8 7

Dati ordinati in modo crescente

3 5 7 7 8 8 9 10 12 35

La media e la mediana di questi dati valgono rispettivamente

$$\bar{x} = 10.4 \qquad M = 8$$

Se il secondo paziente della tabella avesse un valore di 70.000 globuli bianchi, anziché di 35.000, il valore della mediana resterebbe invariato, mentre la media diventerebbe

$$\bar{x} = 13.9$$

Questi esempi ci ricordano che c'è sempre comunque un rischio a riassumere un insieme di dati con un singolo numero.

Oltre alla mediana, che divide a metà un insieme di dati ordinati, si possono definire altri indici di posizione, detti **quantili** e **percentili**, che dividono l'insieme di dati ordinati in un dato numero di parti uguali. Questi **indici di posizione non centrale** sono usati soprattutto per ampi insiemi di dati. I **quartili** sono un caso particolare dei quantili, e si ottengono dividendo l'insieme di dati ordinati in quattro parti uguali.

Definizione 4
Il **primo quartile Q_1** è un valore tale che il 25 % dei dati ordinati è minore o uguale a Q_1. Il primo quartile Q_1 è detto anche **25-esimo percentile** e indicato con $P_{0.25}$.
Il **terzo quartile Q_3** è un valore tale che il 75 % dei dati ordinati è minore o uguale a Q_3 ed è detto anche **75-esimo percentile** e indicato con $P_{0.75}$.

Il secondo quartile Q_2 (50-esimo percentile) coincide con la mediana.
Per calcolare i quartili si segue una regola simile a quella usata per il calcolo della mediana.

Regola per il calcolo dei quartili
1 – Si ordinano gli n dati assegnati in ordine crescente;
2 – si calcola il prodotto $k = np$, dove $p = 0.25$ per il primo quartile e $p = 0.75$ per il terzo quartile;
3 – se k è un intero, il quartile si ottiene facendo la media del k-esimo e del (k+1)-esimo valore dei dati ordinati;
4 – se k non è intero, si arrotonda k per eccesso al primo intero successivo e si sceglie come quartile il corrispondente valore dei dati ordinati.

La regola può essere generalizzata in modo semplice per trovare un qualsiasi altro percentile.
Ad esempio per trovare il 95-esimo percentile, ossia quel valore tale che il 95 % dei dati ordinati è minore o uguale ad esso, si usa la stessa regola, con $p = 0.95$.[2]

Esempio 26
Calcolare il primo e il terzo quartile dell'insieme di dati

32.2 32.0 30.4 31.0 31.2 31.3 30.3 29.6 30.5 30.7

Dati ordinati

29.6 30.3 30.4 30.5 30.7 31.0 31.2 31.3 32.0 32.2

Primo quartile

$n = 10 \qquad p = 0.25 \qquad k = np = 2.5$

k non è intero, perciò si arrotonda per eccesso $k = 3$: il primo quartile è il terzo dei dati ordinati

$Q_1 = 30.4.$

Terzo quartile

$n = 10 \qquad p = 0.75 \qquad k = np = 7.5$

k non è intero, perciò si arrotonda per eccesso $k = 8$: il terzo quartile è l'ottavo dei dati ordinati

$Q_3 = 31.3.$

Secondo quartile (mediana)

$n = 10 \qquad p = 0.5 \qquad k = np = 5$

k è intero, perciò si fa la media tra il quinto e il sesto dato e si ottiene

$$Q_2 = \frac{30.7 + 31.0}{2} = 30.85$$

(Questo valore coincide con quello che si trova con la regola della mediana).

[2] Molti software calcolano i percentili con una regola un po' più complessa, basata sull'interpolazione lineare fra dati adiacenti, perciò i valori trovati possono differire leggermente da quelli ricavati con la regola più semplice qui indicata.

Esempio 27
Calcolare il primo e il terzo quartile e il 95-esimo percentile per i dati della tabella 2, pag. 3.
Dati ordinati

6.2	7.7	8.3	9.0	9.4	9.8	10.5	10.7	11.0	11.2
11.8	12.3	12.8	13.2	13.3	13.5	13.9	14.4	14.5	14.7
15.2	15.5	15.8	15.9	16.2	16.7	16.9	17.0	17.3	17.5
17.6	17.9	18.0	18.0	18.1	18.1	18.4	18.5	18.7	19.0
19.1	19.2	19.3	19.4	19.4	20.0	20.1	20.1	20.4	20.5
20.8	20.9	21.4	21.6	21.9	22.3	22.5	22.7	22.7	22.9
23.0	23.5	23.7	23.9	24.1	24.3	24.6	24.6	24.8	25.7
25.9	26.1	26.4	26.6	26.8	27.5	28.5	28.6	29.6	31.8

Tabella 39

Primo quartile

$$n = 80 \qquad p = 0.25 \qquad k = np = 20$$

k è intero, perciò si fa la media tra il 20-esimo e il 21-esimo dato e si ottiene

$$Q_1 = \frac{14.7 + 15.2}{2} = 14.95$$

Terzo quartile

$$n = 80 \qquad p = 0.75 \qquad k = np = 60$$

k è intero, perciò si fa la media tra il 60-esimo e il 61-esimo dato e si ottiene

$$Q_3 = \frac{22.9 + 23.0}{2} = 22.95$$

95-esimo percentile

$$n = 80 \qquad p = 0.95 \qquad k = np = 76$$

k è intero, perciò si fa la media tra il 76-esimo e il 77-esimo dato e si ottiene

$$P_{0.95} = \frac{27.5 + 28.5}{2} = 28.0$$

Il 95-esimo percentile fornisce un'importante informazione: soltanto il 5% dei dati sono maggiori di 28.0, ossia, con riferimento al tipo di dati descritti nell'esempio 2, soltanto nel 5% dei giorni l'emissione di gas inquinanti supera la soglia di 28.0 unità.

Gli indici di posizione non tengono conto della variabilità esistente fra i dati; vi sono distribuzioni che, pur avendo la stessa media, sono molto diverse fra loro.
I dati dei seguenti insiemi ad esempio hanno la stessa media ($\bar{x} = 60$)

A = {60 60 60 60 60}
B = {10 20 60 100 110}
C = {50 55 60 65 70}

ma gli insiemi sono molto diversi; il primo è composto da dati tutti uguali, mentre il secondo presenta la maggior differenza tra il valore minimo e il massimo.
Indici significativi per la misura della variabilità di una distribuzione di frequenza sono la **varianza** e lo **scarto quadratico medio**, detto anche **deviazione standard**.

Definizione 5
Si definisce **varianza**, o anche **varianza campionaria**, la quantità

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{1.2}$$

dove $\bar{x}$ indica la media dei dati.

Definizione 6
Si definisce **scarto quadratico medio** o **deviazione standard** la radice quadrata della varianza

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{1.3}$$

Nella formula per la varianza si divide per $n - 1$ anziché per n, perché la varianza s^2 definita in questo modo gode di alcune proprietà che la rendono una misura più adeguata nell'inferenza statistica (Capitolo 7).
Si può facilmente dimostrare che per il calcolo della varianza si possono usare le seguenti formule alternative alla (1.2), che richiedono una minor quantità di calcoli e sono più efficienti dal punto di vista dell'accuratezza computazionale (vedere anche esempi 45, 46, 47)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \tag{1.4}$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \tag{1.5}$$

Varianza e scarto quadratico medio sono detti **indici di dispersione** o **indici di variabilità**, perché misurano la dispersione dei dati attorno alla media.
Dalla definizione 5 risulta che la varianza è tanto più grande quanto più i dati si discostano dalla media.
I valori di s e s^2 , poiché misurano l'effettiva variazione assoluta presente in un insieme di dati, dipendono dall'unità di misura dei dati. In particolare lo scarto quadratico medio s misura la dispersione dei dati con la stessa unità di misura della media dei dati, cosa che non accade per la varianza; questa è la ragione principale per cui lo scarto quadratico medio è più usato della varianza.
La media e lo scarto quadratico medio sono i due indici di posizione e di dispersione più usati; uno dei motivi principali è che la distribuzione normale, che viene largamente utilizzata in molti campi diversi, è definita in termini di questi due parametri. La distribuzione normale verrà trattata nel capitolo 5.

Esempio 28
I seguenti dati sono i tempi di esecuzione di una certa operazione misurati in minuti

0.6 1.2 0.9 1.0 0.6 0.8

Calcoliamo la varianza e la deviazione standard.

$$\bar{x} = \frac{0.6 + 1.2 + 0.9 + 1.0 + 0.6 + 0.8}{6} = 0.85 \text{ minuti}$$

Per la varianza, usando la formula (1.2) si dispongono i calcoli nella tabella seguente

x_i	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$
0.6	–0.25	0.0625
1.2	0.35	0.1225
0.9	0.05	0.0025
1.0	0.15	0.0225
0.6	–0.25	0.0625
0.8	–0.05	0.0025
totale		0.2750

Tabella 40

$$s^2 = \frac{0.2750}{5} = 0.055 \text{ minuti}^2$$
$$s = \sqrt{0.055} \cong 0.23 \text{ minuti}$$

Se per la varianza si usa la formula (1.4), che è più efficiente, i calcoli si dispongono invece nella tabella 41 (non si fa uso in modo esplicito del valor medio)

x_i	x_i^2
0.6	0.36
1.2	1.44
0.9	0.81
1.0	1
0.6	0.36
0.8	0.64
5.10	4.61

Tabella 41

$$s^2 = \frac{1}{5}\left(4.61 - \frac{5.10^2}{6}\right) = 0.055 \text{ minuti}^2$$

Esempio 29
Calcoliamo varianza e deviazione standard dei dati della tabella 38.

Per la varianza, usando la formula (1.4) e disponendo i calcoli in una tabella analoga alla tabella 41, si ottiene

$$\sum_{i=1}^{n} x_i = 64820 \qquad \sum_{i=1}^{n} x_i^2 = 213265000$$
$$s^2 = \frac{1}{19}\left(213265000 - \frac{64820^2}{20}\right) = 167546.3 \text{ g}^2$$
$$s = \sqrt{167546.3} = 409.3 \text{ g}$$

Esempio 30
Per la partecipazione a una gara di matematica una scuola deve formare una squadra di 6 studenti; con una selezione preliminare, attraverso un test con un punteggio massimo di 100 punti, sulla base della media dei migliori 6 punteggi risultano tre squadre a pari merito.
Con quale criterio può essere scelta la squadra da mandare alla gara?

squadra	*punteggi degli studenti*					
A	73	76	77	85	88	90
B	74	74	78	84	88	91
C	72	77	79	82	84	95

Tabella 42

La somma dei punteggi ottenuti da ciascuna squadra è 489; la media aritmetica per le tre squadre vale $\bar{x} = 81.5$ e non è quindi un criterio utilizzabile per la scelta; calcoliamo la varianza e lo scarto quadratico medio

squadra A		*squadra* B		*squadra* C	
x_i	x_i^2	x_i	x_i^2	x_i	x_i^2
73	5329	74	5476	72	5184
76	5776	74	5476	77	5929
77	5929	78	6084	79	6241
85	7225	84	7056	82	6724
88	7744	88	7744	84	7056
90	8100	91	8281	95	9025
489	40103	489	40117	489	40159

Tabella 43

squadra A $\quad s^2 = \frac{1}{5}\left(40117 - \frac{1}{6} 489^2 \right) = 49.9$

squadra B $\quad s^2 = \frac{1}{5}\left(40103 - \frac{1}{6} 489^2 \right) = 52.7$

squadra C $\quad s^2 = \frac{1}{5}\left(40159 - \frac{1}{6} 489^2 \right) = 61.1$

squadra	*varianza*	*scarto quadratico medio*
A	49.9	7.06
B	52.7	7.26
C	61.1	7.82

Tabella 44

Utilizzando il criterio dello scarto quadratico medio, la squadra da inviare alla gara è la squadra A, che ha il minor scarto quadratico medio.

Esempio 31
I voti in trentesimi riportati da 25 studenti in un esame sono riportati nella seguente tabella. Individuare quali studenti si discostano dal voto medio per più di una volta oppure due volte lo scarto quadratico medio.

numero studente	1	2	3	4	5	6	7	8	9	10	11	12	13
voto	15	17	27	25	29	14	16	25	27	18	10	15	27

numero studente	14	15	16	17	18	19	20	21	22	23	24	25
voto	28	19	14	30	21	17	24	29	20	13	30	25

Tabella 45

Elaborando i dati si ottengono i seguenti risultati

$\bar{x} = 21.40$
$s = 6.21$
$\bar{x} - s = 15.19 \qquad \bar{x} + s = 27.61$
$\bar{x} - 2s = 8.98 \qquad \bar{x} + 2s = 33.82$

Tutti i voti appartengono all'intervallo $[\bar{x} - 2s , \bar{x} + 2s]$, cioè non vi è nessun voto che si discosta dalla media per più di due volte lo scarto quadratico medio; ci sono invece 11 voti che non appartengono all'intervallo $[\bar{x} - s , \bar{x} + s]$, ossia si discostano dalla media per più di una volta lo scarto quadratico medio.

Per rappresentare la situazione può essere utile un diagramma nel piano cartesiano (figura 26), con il quale si individuano più facilmente gli studenti che rientrano nella fascia delimitata dai valori $\bar{x}-s$, $\bar{x}+s$.

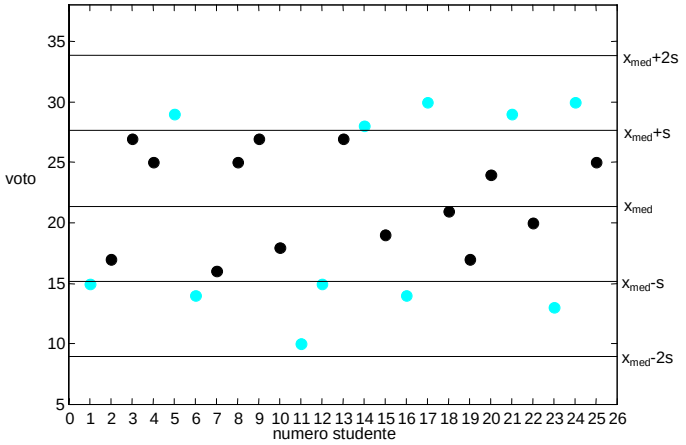


Figura 26

Per confrontare la variazione di molti campioni diversi di dati, ciascuno con media diversa, o misurati in unità di misura diverse, può essere utile usare una misura di variazione relativa, anziché una misura assoluta come lo scarto quadratico medio.

Definizione 7
Il **coefficiente di variazione** CV è definito da

$$CV = \frac{s}{|\bar{x}|} \cdot 100\% \qquad (1.6)$$

Il coefficiente di variazione esprime lo scarto quadratico medio come percentuale della media ed è indipendente dall'unità di misura usata, poiché la media e lo scarto quadratico medio sono espressi nella stessa unità di misura.

Esempio 32
Sia dato un campione di 200 pacchi di cui sono noti il peso e il volume. Calcolando la media e lo scarto quadratico medio delle due misure si ottengono i seguenti valori

Peso medio:	$\bar{x}_P = 9Kg$
Scarto quadratico medio del peso:	$s_P = 1.5Kg$
Volume medio:	$\bar{x}_V = 2.7m^3$
Scarto quadratico medio del volume:	$s_V = 0.6m^3$

Confrontiamo la variabilità del peso e del volume.
Siccome il peso e il volume sono espressi in unità di misura diverse, occorre prendere in considerazione la variabilità relativa delle osservazioni, calcolando il coefficiente di variazione.
Per il peso il coefficiente di variazione è

$$CV = \frac{1.5}{9} \cdot 100\% = 16.67\% .$$

Per il volume il coefficiente di variazione è

$$CV = \frac{0.6}{2.7} \cdot 100\% = 22.22\% .$$

Pertanto, rispetto alla media, il volume dei pacchi è più variabile del peso.

Esempio 33
Le misure del diametro di un cuscinetto a sfera effettuate con uno strumento hanno un valor medio $\bar{x}$ = 3.92 mm e uno scarto quadratico medio s = 0.015 mm; le misure della lunghezza di una sbarra rigida effettuate con un altro strumento hanno invece un valor medio $\bar{x}$ = 1.54 m e uno scarto quadratico medio s = 0.008 m. Quale dei due strumenti è relativamente più preciso?

Per il primo strumento il coefficiente di variazione è

$$CV = \frac{0.015}{3.92} \cdot 100 = 0.38\%$$

Per il secondo strumento è invece

$$CV = \frac{0.008}{1.54} \cdot 100 = 0.52\%$$

Il primo strumento è relativamente più preciso del secondo.

1.4 Calcolo di media e varianza per dati raggruppati

Nel caso in cui i dati siano molto numerosi, non disponendo di un computer il calcolo della media e della varianza viene semplificato se si raggruppano i dati prima di utilizzarli; può inoltre succedere di dover calcolare media e varianza di dati che sono noti solo nella forma di dati raggruppati. In questi casi il calcolo esatto non è possibile, ma si può calcolare una buona approssimazione di media e varianza, supponendo che i dati di ogni classe siano approssimati dal valore centrale della classe.
Dopo aver raggruppato gli n dati in k classi, indichiamo con m_i il valore centrale della generica classe e con f_i la corrispondente frequenza assoluta della classe.

Definizioni 8
La **media per dati raggruppati** è definita da

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} m_i f_i \tag{1.7}$$

La **varianza per dati raggruppati** è definita da

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} \left(m_i - \bar{x}\right)^2 f_i \tag{1.8}$$

Per il calcolo della varianza per dati raggruppati si possono usare le seguenti formule alternative alla (1.8) e più accurate dal punto di vista computazionale

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^{k} f_i m_i^2 - \frac{1}{n} \left(\sum_{i=1}^{k} f_i m_i \right)^2 \right] \tag{1.9}$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} f_i m_i^2 - n\bar{x}^2 \right] \tag{1.10}$$

Osserviamo che, se sono disponibili i dati grezzi, con la diffusione dei computer e dei software statistici queste formule per dati raggruppati hanno perso molta della loro importanza.

Esempio 34
Riprendiamo l'esempio 5 nel quale, raggruppando i dati con 7 classi aperte a sinistra, abbiamo ottenuto la tabella seguente (tabella 5b)

Classi	m_i	f_i
$5 < x \le 9$	7	4
$9 < x \le 13$	11	9
$13 < x \le 17$	15	15
$17 < x \le 21$	19	24
$21 < x \le 25$	23	17
$25 < x \le 29$	27	9
$29 < x \le 33$	31	2
Totale		80

Tabella 47

Applicando le formule (1.7) e (1.10) per i dati raggruppati si ottiene per la media

$$\bar{x} = \frac{1}{80}(7 \cdot 4 + 11 \cdot 9 + 15 \cdot 15 + 19 \cdot 24 + 23 \cdot 17 + 27 \cdot 9 + 31 \cdot 2) = 18.8$$

e per la varianza

$$s^2 = \frac{1}{79}[7^2 \cdot 4 + 11^2 \cdot 9 + 15^2 \cdot 15 + 19^2 \cdot 24 + 23^2 \cdot 17 + 27^2 \cdot 9 + 31^2 \cdot 2] = 31.96$$

Se il calcolo viene fatto sui dati non raggruppati (tabella 2) si ottiene invece

$$\bar{x} = 18.89$$
$$s^2 = 32.00$$

Come si nota, i valori ottenuti dai dati raggruppati sono un'approssimazione dei valori più precisi calcolati su tutti i dati.

Esempio 35
Quattro gruppi di 18, 20, 10 e 15 scolari hanno un'altezza media rispettivamente di 140 cm, 148 cm, 153 cm e 162 cm.
Determinare l'altezza media di tutti gli scolari e la varianza, con le formule dei dati raggruppati.

Tabella della distribuzione di frequenza

m_i	f_i	$m_i - \bar{x}$	$(m_i - \bar{x})^2$
140	18	–10	100
148	20	–2	4
153	10	3	9
162	15	12	144

Tabella 49

Numero totale degli scolari $n = 63$
Media

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{4} m_i f_i = \frac{15 \cdot 162 + 20 \cdot 148 + 10 \cdot 153 + 18 \cdot 140}{63} = 150 \text{ cm}$$

Varianza, con la formula (1.8)

$$s^2 = \frac{1}{62}(100 \cdot 18 + 4 \cdot 20 + 9 \cdot 10 + 144 \cdot 15) = \frac{4130}{62} \cong 66.6 \text{cm}^2$$

La mediana è uguale al 32° dato

$$M = 148$$

La moda è uguale al dato che si presenta con maggior frequenza

$$\tilde{x} = 148$$

Esempio 36

La tabella 48 riassume i voti finali in matematica degli studenti di una classe; calcolare il voto medio della classe.

voto	3	4	5	6	7	8	9	10
numero studenti	3	5	2	8	5	1	1	0

Tabella 48

I voti finali in una materia sono una distribuzione di frequenza, in cui alcuni voti sono generalmente attribuiti a più studenti; il numero complessivo degli studenti è

$$3+5+2+8+5+1+1=25$$

Il voto medio è

$$\bar{x} = \frac{3\cdot 3+4\cdot 5+5\cdot 2+6\cdot 8+7\cdot 5+8\cdot 1+9\cdot 1+10\cdot 0}{25} = 5.56$$

Esempio 37

In un insieme di numeri compaiono dieci volte il 6, cinque volte il 7, nove volte l'8, dodici volte il 9 e quattro volte il 10. Trovare la media aritmetica di questi numeri.

Si tratta di 40 dati raggruppati, la cui media vale

$$\bar{x} = \frac{10\cdot 6+5\cdot 7+9\cdot 8+12\cdot 9+4\cdot 10}{40} = 7.875$$

Esempio 38

Nella tabella seguente si riportano i punteggi ottenuti in 40 lanci successivi di un dado

classe (punteggio)	f_i
1	9
2	8
3	5
4	5
5	6
6	7

Tabella 49

Calcolare la media, la mediana, la moda e la varianza.

Media

$$\bar{x} = \frac{1}{40}(9+16+15+20+30+42) = 3.3$$

Mediana: è la semisomma del 20-esimo e del 21-esimo valore (i dati devono essere prima disposti in ordine crescente)

$$M = \frac{3+3}{2}$$

Moda: è il punteggio a cui corrisponde la maggior frequenza

$$\tilde{x} = 1$$

Varianza

$$s^2 = \frac{1}{39}[(1-3.3)^2 \cdot 9 + (2-3.3)^2 \cdot 8 + (3-3.3)^2 \cdot 5 +$$
$$(4-3.3)^2 \cdot 5 + (5-3.3)^2 \cdot 6 + (6-3.3)^2 \cdot 7] = 3.395$$

1.5 Forma di una distribuzione

Un'altra caratteristica dei dati che prendiamo in considerazione è la **forma** della loro distribuzione. Le distribuzioni di frequenza possono assumere più forme diverse, e fra queste le più importanti sono quelle che assumono una **forma a campana**. In questo caso la distribuzione dei dati è **simmetrica** rispetto a una linea verticale (linea tratteggiata - figura 27); i **dati** di questo tipo si dicono **normali**.
Se la distribuzione dei dati non è perfettamente simmetrica, i **dati** si dicono **approssimativamente normali** (figura 28).

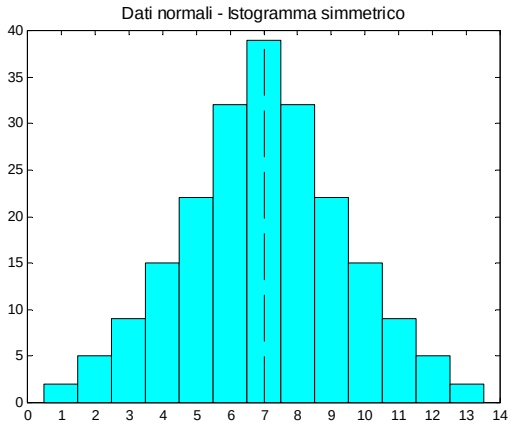


Figura 27

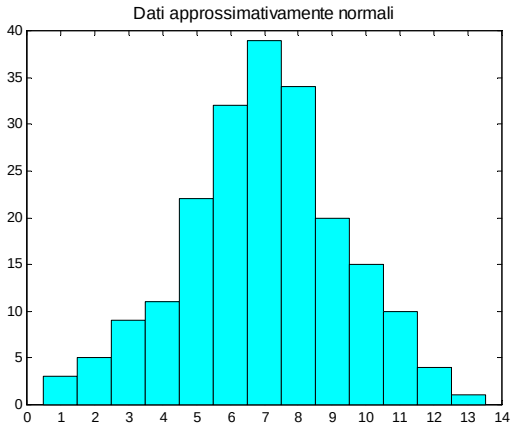


Figura 28

Una **distribuzione asimmetrica,** detta anche **obliqua**, può avere una "coda" a destra e viene detta **distribuzione obliqua a destra** o con **asimmetria positiva** (figura 29); se invece la coda è a sinistra, si dice che la **distribuzione** è **obliqua a sinistra** o con **asimmetria negativa** (figura 30).

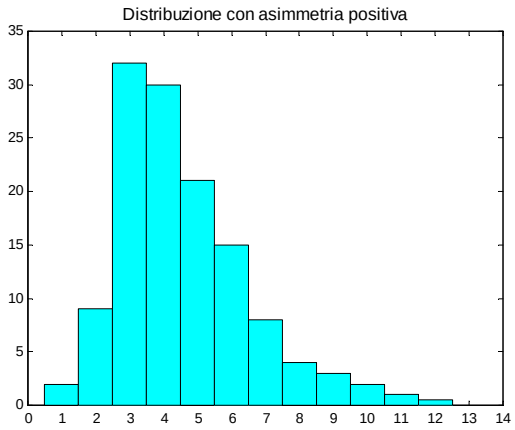


Figura 29

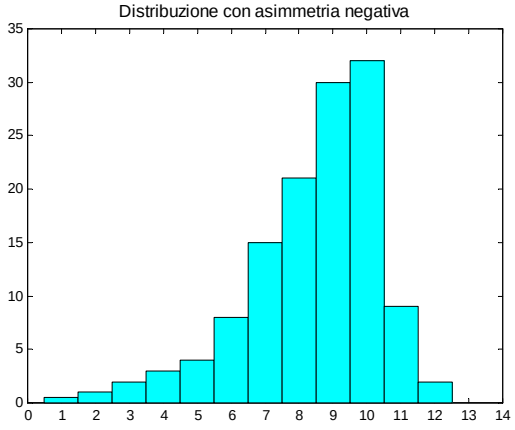


Figura 30

Per descrivere la forma della distribuzione è sufficiente confrontare la media con la mediana: se queste due misure sono uguali la distribuzione è simmetrica; se la media è maggiore della mediana, la distribuzione ha asimmetria positiva (obliqua a destra, figura 31); se invece la media è minore della mediana, la distribuzione ha asimmetria negativa (obliqua a sinistra).

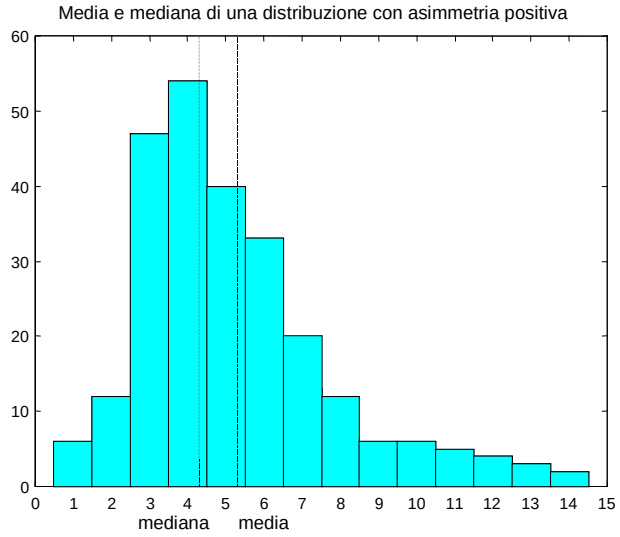


Figura 31

Questa relazione fra la media e la mediana può essere usata per definire una misura di asimmetria, detta **coefficiente di asimmetria di Pearson.**

Definizione 9
Siano $\bar{x}$, M e s rispettivamente la media, la mediana e lo scarto quadratico medio di un insieme di dati; il **coefficiente di asimmetria di Pearson** è definito da

$$SK = \frac{3(\bar{x} - M)}{s} \qquad (1.11)$$

Il coefficiente SK è indipendente dall'unità di misura dei dati. Per una distribuzione perfettamente simmetrica SK vale 0; per una distribuzione asimmetrica positivamente il valore di SK è positivo, mentre è negativo per una distribuzione asimmetrica negativamente.
In generale i valori di SK cadono fra -3 e 3. La divisione per lo scarto quadratico medio rende il valore di SK indipendente dall'unità di misura dei dati.

Sebbene la media e lo scarto quadratico medio siano solo misure descrittive di un insieme di dati, esse forniscono importanti informazioni sulla distribuzione dei dati. Se la distribuzione dei dati è approssimativamente normale vale infatti la seguente regola.

Regola empirica
Se un insieme di dati è approssimativamente normale, con media $\bar{x}$ e scarto quadratico medio s, allora:
1 – circa il 68% dei dati è compreso fra $\bar{x} - s$ e $\bar{x} + s$;
2 – circa il 95% dei dati è compreso fra $\bar{x} - 2s$ e $\bar{x} + 2s$;
3 – circa il 99.7% dei dati è compreso fra $\bar{x} - 3s$ e $\bar{x} + 3s$;

Questo risultato, noto come regola empirica presumibilmente perché le percentuali indicate sono osservate nella pratica, è in realtà un risultato teorico basato sulle proprietà della distribuzione normale, che sarà studiata nel capitolo 5.

Esempio 39
Per i dati dell'esempio 2 (tabella 2) si possono calcolare i seguenti valori per la media e lo scarto quadratico (si veda anche l'esempio 34)

$$\bar{x} = 18.89 \qquad s^2 = 32.00 \qquad s = 5.66$$

La regola empirica in questo caso afferma che circa il 68% dei dati cade fra i valori

$$\bar{x} - s = 18.89 - 5.66 = 13.23 \qquad \text{e} \qquad \bar{x} + s = 18.89 + 5.66 = 24.55$$

Usando la tabella 39, dove compaiono gli stessi dati in ordine crescente, si può facilmente verificare che 14 dati cadono prima di 13.23 e 14 dati cadono dopo 24.55, quindi 80 – 28 = 52 dati cadono nell'intervallo (13.23 , 24.55), ossia il $\frac{52}{80} \cdot 100\% = 65\%$ dei dati.
Con lo stesso metodo si osserva sulla tabella che il 97.5% dei dati cade fra

$$\bar{x} - 2s = 18.89 - 2 \cdot 5.66 = 7.57 \qquad \text{e} \qquad \bar{x} + 2s = 18.89 + 2 \cdot 5.66 = 30.21$$

e la regola empirica prevede il 95%.

Esempio 40
Riprendiamo l'esempio 17; l'istogramma rappresentato nella figura 18 evidenzia una distribuzione dei dati approssimativamente normale.
Calcoliamo media e varianza con le formule (1.7) e (1.8) per dati raggruppati.

$$\bar{x} = \frac{1}{1000}\left(0 \cdot 38 + 1 \cdot 144 + 2 \cdot 342 + 3 \cdot 287 + 4 \cdot 164 + 5 \cdot 25\right) = 2.47$$

$$s^2 = \frac{1}{999}\left[0^2 \cdot 38 + 1^2 \cdot 144 + 2^2 \cdot 342 + 3^2 \cdot 287 + \right.$$
$$\left. + 4^2 \cdot 164 + 5^2 \cdot 25 - 1000 \cdot 2.47^2\right] = 1.244$$

$$s = 1.12$$

Se si immagina di disporre in ordine crescente i 1000 dati (numero di teste ottenute ad ogni lancio), si può osservare che i valori che occupano le due posizioni centrali sono uguali a 2, perciò la mediana è

$$M = 2$$

Il valore che si presenta con la maggior frequenza (342 volte) è 2, ossia la moda è

$$\tilde{x} = 2\,.$$

Si ha

$$\bar{x} - s = 2.47 - 1.12 = 1.35 \qquad \bar{x} + s = 2.47 + 1.12 = 3.59$$

$$\bar{x} - 2s = 2.47 - 2 \cdot 1.12 = 0.23 \qquad \bar{x} + 2s = 2.47 + 2 \cdot 1.12 = 4.75$$

Il numero di dati compresi fra $\bar{x} - s$ e $\bar{x} + s$ è dato dal numero di dati uguali a 2 e a 3, ossia 629, ed è il 63% circa dei dati; il numero di dati compresi fra $\bar{x} - 2s$ e $\bar{x} + 2s$ è dato dal numero di dati uguali a 1, 2, 3 e 4, ossia 917 ed è il 92% circa dei dati.

1.6 Correlazione fra variabili

Spesso nell'indagine statistica si eseguono analisi di tipo comparativo, ad esempio si osservano più variabili su un medesimo gruppo di individui.
Un problema tipico consiste nel chiedersi se esiste una **correlazione** fra le variabili osservate.
Il primo passo utile per indagare qualitativamente l'eventuale dipendenza fra due variabili *x* e *y* consiste nel disegnare un grafico, detto **diagramma di dispersione** o **scatterplot**.
Si pongono in ascissa i dati relativi a una delle due variabili, in ordinata quelli relativi all'altra variabile e si rappresentano con punti o cerchietti le singole osservazioni. Se esiste una relazione semplice fra le due variabili, il diagramma dovrebbe evidenziarla.

Si osservino ad esempio i due diagrammi seguenti

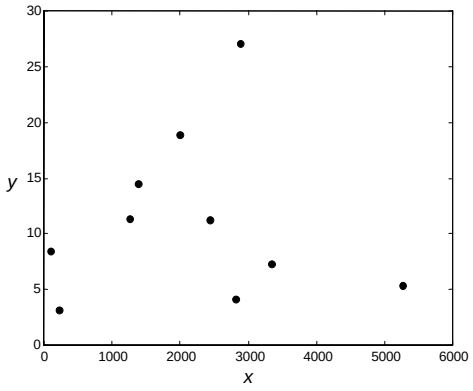


Figura 32

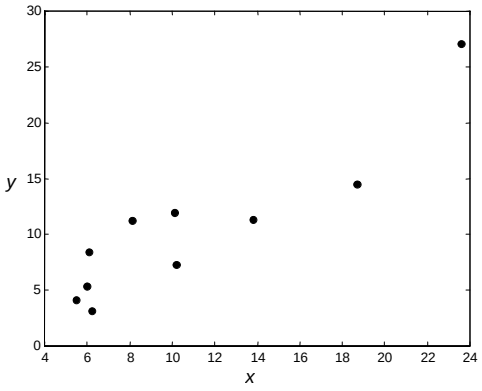


Figura 33

Il primo diagramma non suggerisce che vi sia una correlazione fra le due variabili: i punti sono sparsi senza apparenti regolarità. Il secondo diagramma evidenzia invece una certa regolarità: punti con ascissa piccola hanno ordinata piccola e punti con ascissa grande hanno ordinata grande; in questo caso si dice che esiste una correlazione diretta fra le due variabili. Analogamente si parla di correlazione inversa fra le due variabili se al crescere di una di esse l'altra decresce.
Nella figura 33 si può ipotizzare una correlazione tra le due variabili di tipo lineare; in tal caso si può tracciare la retta di regressione, cioè la retta che "più si avvicina" a tutti i punti.
Esaminiamo dapprima il concetto di correlazione fra variabili.

Definizione 10
Date *n* osservazioni congiunte di due variabili *x* e *y*

$$(x_1, y_1), \quad (x_2, y_2), \quad ....., (x_n, y_n)$$

si dice **covarianza** delle due variabili *x, y* il numero

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \tag{1.12}$$

Definizione 11
Si dice **coefficiente di correlazione** delle due variabili *x, y* il numero

$$r = \frac{S_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} \tag{1.13}$$

dove s_x^2 e s_y^2 sono le varianze delle variabili *x* e *y*.

Per il calcolo della covarianza si può anche usare la formula seguente

$$S_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}\right) \tag{1.14}$$

La covarianza può avere segno positivo o negativo, e il coefficiente di correlazione ha lo stesso segno della covarianza.
Per il calcolo del coefficiente di correlazione si può anche usare la seguente formula

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{1.15}$$

Definizione 12
Si dice che fra le variabili *x, y* c'è una **correlazione diretta o positiva** se $S_{xy} > 0$; si dice che c'è una **correlazione inversa o negativa** se $S_{xy} < 0$; si dice infine che le variabili sono **non correlate** se $S_{xy} = 0$

Si può dimostrare che il coefficiente di correlazione r varia tra -1 e 1; in particolare $r = \pm 1$ se e solo se i punti sono tutti perfettamente allineati sulla stessa retta, ossia esistono due numeri A e B tali che

$$y_i = Ax_i + B \qquad i = 1,2,\ldots,n$$

dati correlati positivamente
(covarianza positiva)

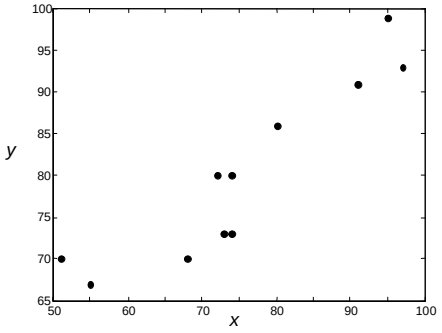


Figura 34

dati correlati negativamente
(covarianza negativa)

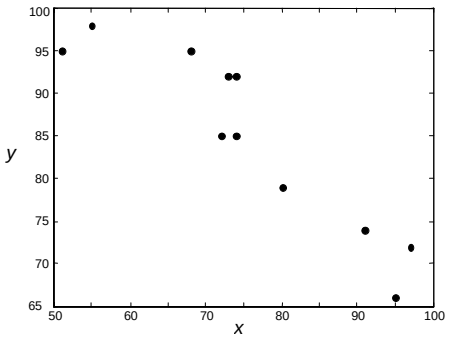


Figura 35

dati non correlati

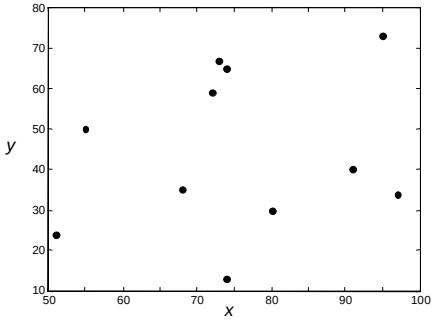


Figura 36

Esempio 41
I seguenti dati sono i punteggi che 10 studenti hanno conseguito in due esami di Analisi Matematica (punteggio massimo = 100). Calcolare la covarianza e il coefficiente di correlazione.

Analisi I	*Analisi* II
51	74
68	70
97	93
55	67
95	99
74	73
20	33
91	91
74	80
80	86

Tabella 50

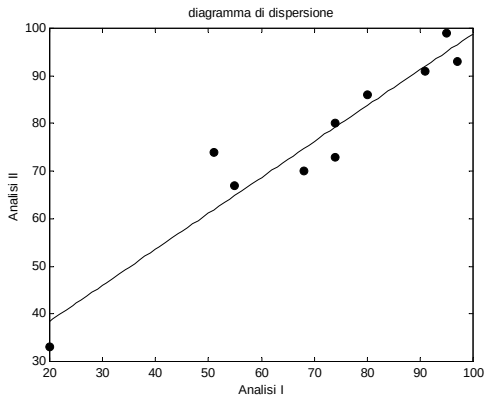


Figura 37

Per calcolare covarianza e coefficiente di correlazione, se non si dispone di un computer, si dispongono i calcoli in una tabella (nell'ultima riga sono indicate le somme delle colonne)

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})\cdot(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
51	74	–19.5	–2.6	50.7	380.2	6.8
68	70	–2.5	–6.6	16.5	6.3	43.6
97	93	26.5	16.4	434.6	702.3	269.0
55	67	–15.5	–9.6	148.8	240.2	92.2
95	99	24.5	22.4	548.8	600.2	501.8
74	73	3.5	–3.6	–12.6	12.3	13.0
20	33	–50.5	–43.6	2201.8	2550.3	1901.0
91	91	20.5	14.4	295.2	420.3	207.4
74	80	3.5	3.4	11.9	12.3	11.6
80	86	9.5	9.4	89.3	90.2	88.4
705	766			3785.0	5014.5	3134.4

Tabella 51

Si ottengono i seguenti risultati

$$\bar{x} = 70.5 \qquad \bar{y} = 76.6$$

$$S_{xy} = 420.55 \qquad s_x^2 = 557.17 \qquad s_y^2 = 348.27$$

$$r = 0.955$$

I dati sono positivamente correlati; il diagramma di dispersione e il valore di *r* prossimo al valore 1 indicano una relazione lineare fra i dati.

1.7 Metodo dei minimi quadrati. Regressione lineare

In base a quanto detto nel § 1.6, se il coefficiente di correlazione *r* non vale ± 1, certamente i dati y_i non sono esattamente una funzione lineare dei dati x_i. Tuttavia, se il diagramma di dispersione suggerisce una relazione di tipo lineare e il valore di *r* è prossimo a +1 o a –1, ha senso determinare l'equazione di una retta che approssimi "nel modo migliore" i dati assegnati.

Sia dato un insieme di *n* punti $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ e sia

$$y = Ax + B$$

l'equazione della retta che si vuole determinare.

Una strategia per determinare tale retta può consistere nel trovare i valori *A* e *B* per i quali è minima la somma

$$\sum_{i=1}^{n} (Ax_i + B - y_i) \tag{1.16}$$

Questo criterio risulta però inadeguato, come mostra la figura 38, che rappresenta l'approssimazione di due soli punti con una retta. Ovviamente la retta migliore è quella che congiunge i due punti, ma qualsiasi retta passante per il punto medio del segmento che congiunge i due punti rende minima la quantità (1.16) (la somma vale zero perché si sommano due valori uguali e di segno opposto).

Si potrebbe allora pensare di minimizzare la somma dei valori assoluti

$$\sum_{i=1}^{n} |Ax_i + B - y_i| \tag{1.17}$$

ma anche questo criterio non è adeguato, come mostra la figura 39; nel caso dei quattro punti rappresentati nella figura 39, qualunque retta compresa tra le due rette r e s che uniscono i punti a due a due soddisfa il criterio (1.17).

Entrambi i criteri (1.16) e (1.17) sono insoddisfacenti perché non conducono ad una soluzione unica.

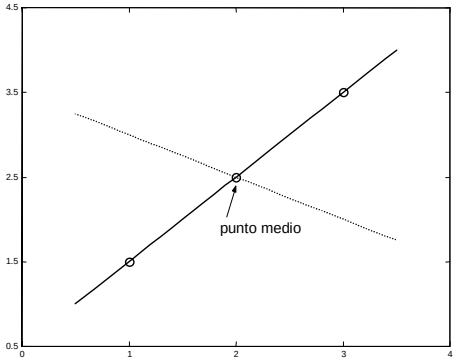


Figura 38

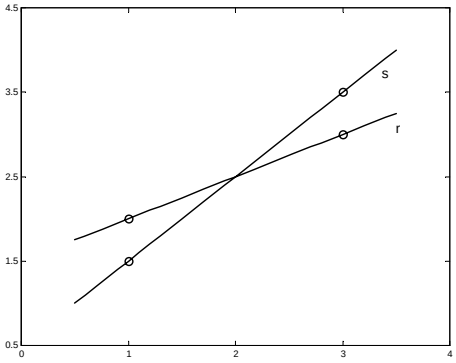


Figura 39

Il criterio che viene usato per definire "il modo migliore" di approssimare i dati, e permette di trovare l'equazione della retta che li approssima, consiste nel minimizzare la quantità

$$E = \sum_{i=1}^{n} \left(Ax_i + B - y_i\right)^2$$

Questo criterio è detto **metodo dei minimi quadrati**. La caratteristica più importante di questo criterio è che consente di determinare un'unica retta di regressione per ogni insieme di dati.
Il grafico che segue illustra il criterio adottato: si richiede che sia minima la somma dei quadrati delle lunghezze dei segmenti che costituiscono le distanze verticali dei punti dalla retta.

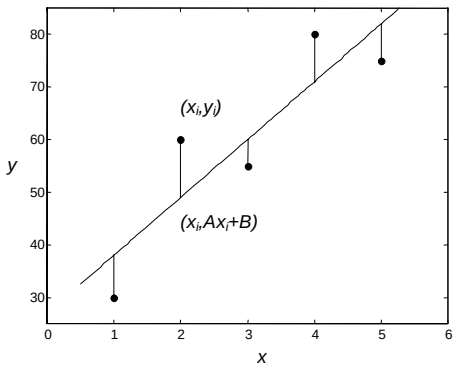


Figura 40

Definizione 13
La **retta dei minimi quadrati** o **retta di regressione** è la retta di equazione

$$y = Ax + B$$

per la quale è minima la quantità

$$E = \sum_{i=1}^{n} \left(Ax_i + B - y_i\right)^2 \qquad (1.18)$$

Si può dimostrare[3] che i coefficienti A e B dell'equazione della retta di regressione sono le soluzioni del seguente sistema lineare di due equazioni nelle incognite A e B, detto **sistema delle equazioni normali**

$$\begin{cases} A\sum_{i=1}^{n} x_i^2 + B\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ A\sum_{i=1}^{n} x_i + nB = \sum_{i=1}^{n} y_i \end{cases} \tag{1.19}$$

Si dimostra che la soluzione del sistema esiste ed è unica, purché i punti non siano tutti allineati verticalmente. La soluzione di questo sistema può essere trovata ad esempio con il metodo di Cramer.

Esempio 42
Determinare la retta di regressione lineare per i dati riportati nelle prime due colonne della tabella seguente

Per scrivere il sistema lineare (1.19) conviene disporre i calcoli in una tabella (nell'ultima riga si riportano le somme delle colonne).

x_i	y_i	$x_i\, y_i$	x_i^2
–1	10	–10	1
0	9	0	0
1	7	7	1
2	5	10	4
3	4	12	9
4	3	12	16
5	0	0	25
6	–1	–6	36
20	37	25	92

Tabella 52

Il sistema delle equazioni normali è il seguente

[3] Per trovare la retta di regressione basta minimizzare la funzione $E(A,B)$ data da (1.18), dove A e B sono le variabili e i punti (x_i, y_i) sono noti. In un punto di minimo della funzione $E(A,B)$ le derivate parziali $\frac{\partial E}{\partial A}$ e $\frac{\partial E}{\partial B}$ si annullano.
Calcolando le derivate parziali e imponendo che siano nulle si trova il sistema delle equazioni normali (1.19)

$$\begin{cases} \frac{\partial E}{\partial A} = \sum_{i=1}^{n} 2x_i\left(Ax_i + B_i - y_i\right) = 2\sum_{i=1}^{n}\left(Ax_i^2 + Bx_i - x_i y_i\right) \\ \frac{\partial E}{\partial B} = \sum_{i=1}^{n} 2\left(Ax_i + B - y_i\right) = 2\sum_{i=1}^{n}\left(Ax_i + B - y_i\right) \end{cases}$$

$$\begin{cases} 2\sum_{i=1}^{n}\left(Ax_i^2 + Bx_i - x_i y_i\right) = 0 \\ 2\sum_{i=1}^{n}\left(Ax_i + B - y_i\right) = 0 \end{cases} \qquad \begin{cases} A\sum_{i=1}^{n} x_i^2 + B\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ A\sum_{i=1}^{n} x_i + nB = \sum_{i=1}^{n} y_i \end{cases}$$

$$\begin{cases} 92A + 20B = 25 \\ 20A + 8B = 37 \end{cases}$$

Troviamo la soluzione con il metodo di Cramer

$$D = \begin{vmatrix} 92 & 20 \\ 20 & 8 \end{vmatrix} = 336 \qquad D_A = \begin{vmatrix} 25 & 20 \\ 37 & 8 \end{vmatrix} = -540 \qquad D_B = \begin{vmatrix} 92 & 25 \\ 20 & 37 \end{vmatrix} = 2904$$

$$A = \frac{D_A}{D} = -\frac{540}{336} \cong -1.61 \qquad B = \frac{D_B}{D} = \frac{2904}{336} \cong 8.64$$

La retta di regressione ha equazione

$$y = -1.61x + 8.64$$

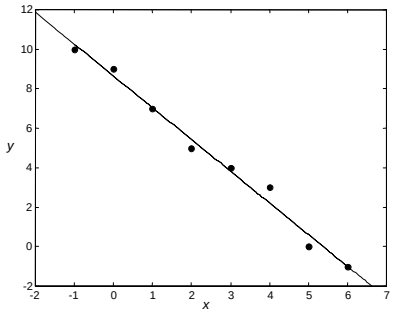


Figura 41

Esempio 43

Nella seguente tabella si riportano le misure dell'ossigeno consumato da una persona che cammina, in corrispondenza a varie velocità della persona.

velocità (Km/h)	*ossigeno (litri/h)*
0	19
1	20
2	20.5
3	21.5
4	22
5	23
6	23
7	23.5
8	24

Tabella 53

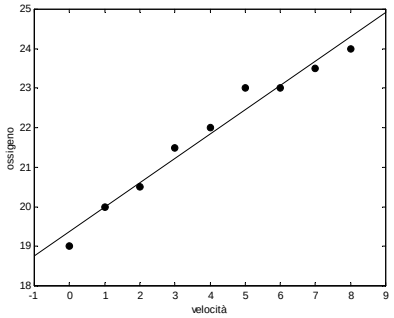


Figura 42

Il diagramma di dispersione mostra chiaramente che il volume dell'ossigeno consumato è all'incirca una funzione lineare della velocità dell'individuo.

x_i	y_i	$x_i\, y_i$	x_i^2
0	19	0	0
1	20	20	1
2	20.5	41	4
3	21.5	64.5	9
4	22	88	16
5	23	115	25
6	23	138	36
7	23.5	164.5	49
8	24	192	64
36	196.5	823	204

Tabella 54

Il sistema delle equazioni normali è

$$\begin{cases} 204A + 36B = 823 \\ 36A + 9B = 196.5 \end{cases}$$

La sua soluzione con il metodo di Cramer è

$$D = \begin{vmatrix} 204 & 36 \\ 36 & 9 \end{vmatrix} = 540 \qquad D_A = \begin{vmatrix} 823 & 36 \\ 196.5 & 9 \end{vmatrix} = 333 \qquad D_B = \begin{vmatrix} 204 & 823 \\ 36 & 196.5 \end{vmatrix} = 10458$$

$$A = \frac{333}{540} \cong 0.62 \qquad B = \frac{10458}{540} \cong 19.37$$

La retta di regressione ha equazione

$$y = 0.62x + 19.37$$

Esempio 44

Si vuole studiare la relazione che intercorre tra il numero di anni di studio di una lingua straniera e il punteggio ottenuto in un test di conoscenza della lingua.

x_i *n° anni studio*	y_i *punteggio*
3	57
4	78
4	72
2	58
5	89
3	63
4	73
5	84
3	75
2	48

Tabella 55

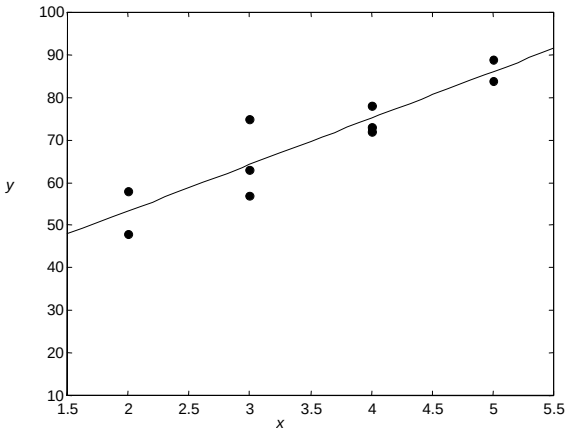


Figura 43

Il diagramma di dispersione evidenzia la relazione lineare fra i dati. Determiniamo l'equazione della retta di regressione lineare.

x_i	y_i	$x_i\, y_i$	x_i^2
3	57	171	9
4	78	312	16
4	72	288	16
2	58	116	4
5	89	445	25
3	63	189	9
4	73	292	16
5	84	420	25
3	75	225	9
2	48	96	4
35	697	2554	133

Tabella 56

Il sistema delle equazioni normali é

$$\begin{cases} 133A + 35B = 2554 \\ 35A + 10B = 697 \end{cases}$$

La soluzione con il metodo di Cramer é

$$D = \begin{vmatrix} 133 & 35 \\ 35 & 10 \end{vmatrix} = 105 \qquad D_A = \begin{vmatrix} 2554 & 35 \\ 697 & 10 \end{vmatrix} = 1145 \qquad D_B = \begin{vmatrix} 133 & 2554 \\ 35 & 697 \end{vmatrix} = 3311$$

$$A = \frac{1145}{105} \cong 10.90 \qquad B = \frac{3311}{105} \cong 31.53$$

La retta di regressione lineare ha equazione

$$y = 10.90x + 31.53$$

Si osservi (figura 43) che in questo esempio alcuni dei punti (non tutti!) sono allineati verticalmente.

In alternativa al metodo basato sulla soluzione del sistema lineare delle equazioni normali, si dimostra che l'equazione della retta di regressione può anche essere ricavata con le seguenti formule, che utilizzano la covarianza delle due distribuzioni di *x* e *y* e la varianza di *x*. Queste formule sono più efficienti dal punto di vista computazionale.

$$y = Ax + B$$

$$A = \frac{S_{xy}}{s_x^2} \qquad B = \bar{y} - A\bar{x} \tag{1.20}$$

S_{xy} = covarianza di *x* e *y* $\qquad s_x^2$ = varianza di *x*

Per il calcolo della covarianza S_{xy} e della varianza s_x^2 si usano preferibilmente le formule (1.14) e (1.5).

Si osservi che il coefficiente angolare della retta ha il segno della covarianza, coerentemente con la definizione data di correlazione diretta e inversa: se tra *x* e *y* c'è una correlazione diretta (inversa), la retta di regressione sarà una retta crescente (decrescente).

La retta di regressione può essere usata per fare delle previsioni sul valore $\hat{y}$ della variabile *y* in corrispondenza a un valore $\hat{x}$ della variabile *x*, diverso dai valori x_i osservati; la previsione sarà tanto più affidabile, quanto più il valore di $\hat{x}$ è vicino ai valori x_i già osservati.

Esempio 45

Si considerino i valori della tabella 58, ottenuti osservando il tempo che impiega un computer a processare dei dati; *x* è il numero di dati processati, *y* il tempo impiegato in secondi.

Tracciamo un diagramma di dispersione, che evidenzia un andamento di tipo lineare.

x_i *n° dati*	y_i *tempo*
105	44
511	214
401	193
622	299
330	143

Tabella 57

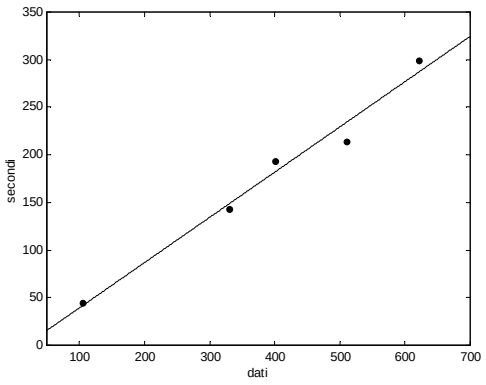


Figura 44

Per il calcolo della covarianza e del coefficiente di correlazione disponiamo i calcoli nella tabella 58.

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
105	44	4620	11025	1936
511	214	109354	261121	45796
401	193	77393	168801	37249
622	299	185978	386884	89401
330	143	47190	108900	20449
1969	893	424535	928731	194831

Tabella 58

$$\bar{x} = \frac{1969}{5} = 393.8 \qquad \bar{y} = \frac{893}{5} = 178.6$$

$$S_{xy} = \frac{1}{4}\left(424535 - 5 \cdot 393.8 \cdot 178.6\right) = 18218$$

$$s_x^2 = \frac{1}{4}\left(928731 - 5 \cdot 393.8^2\right) = 38335$$

$$s_y^2 = \frac{1}{4}\left(194831 - 5 \cdot 178.6^2\right) = 8835$$

$$r = \frac{18218}{\sqrt{38335 \cdot 8835}} = 0.99$$

Il valore del coefficiente di correlazione mostra che esiste una forte correlazione positiva fra le variabili: infatti il coefficiente *r* è molto vicino a 1. E' perciò significativo determinare la retta di regressione, la cui equazione è

$$y = Ax + B$$

$$A = \frac{S_{xy}}{s_x^2} = \frac{18218}{38335} \cong 0.475$$

$$B = \bar{y} - A\bar{x} = 178.6 - 0.475 \cdot 393.8 \cong -8.455$$

$$y = 0.475x - 8.455$$

La retta di regressione può essere usata per fare ad esempio le seguenti previsioni

dati da processare	*tempo previsto*
200	0.475·200–8.455 = 86.55
300	0.475·300–8.455 = 134.05
400	0.475·400–8.455 = 181.55
500	0.475·500–8.455 = 229.05

Tabella 59

Esempio 46

Nella seguente tabella sono riportate le misure del volume di una quantità di un gas a differenti temperature

temperatura	10	20	30	40	50	60
volume	10.4	11.1	11.2	11.9	11.8	12.3

Tabella 60

Verifichiamo che esiste dipendenza lineare del volume dalla temperatura e determiniamo l'equazione della retta di regressione.

Per il calcolo della covarianza e del coefficiente di correlazione disponiamo i calcoli nella tabella 61.

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
10	10.4	104	100	108.16
20	11.1	222	400	123.21
30	11.2	336	900	125.44
40	11.9	476	1600	141.61
50	11.8	590	2500	139.24
60	12.3	738	3600	151.29
210	68.7	2466	9100	788.95

Tabella 61

$$\bar{x} = 35 \qquad \bar{y} = 11.45$$

$$S_{xy} = \frac{1}{5}\left(2466 - 6 \cdot 35 \cdot 11.45\right) = 12.3$$

$$s_x^2 = \frac{1}{5}\left(9100 - 6 \cdot 35^2\right) = 350$$

$$s_y^2 = \frac{1}{5}\left(788.95 - 6 \cdot 11.45^2\right) = 0.467$$

$$r = \frac{12.3}{\sqrt{350 \cdot 0.467}} \cong 0.96$$

Il valore del coefficiente di correlazione mostra che esiste una forte correlazione positiva fra le variabili: infatti il coefficiente r è molto vicino a 1. E' perciò significativo determinare la retta di regressione, la cui equazione è

$$y = Ax + B$$

$$A = \frac{S_{xy}}{s_x^2} = \frac{12.3}{350} \cong 0.035 \qquad B = \bar{y} - A\bar{x} = 11.45 - 0.035 \cdot 35 \cong 10.22$$

$$y = 0.035x + 10.22$$

Il calcolo del coefficiente di correlazione può anche essere fatto con la formula (1.15). In tal caso, usando ancora la tabella 61, si ha

$$r = \frac{6 \cdot 2466 - 210 \cdot 68.7}{\sqrt{6 \cdot 9100 - 210^2}\sqrt{6 \cdot 788.95 - 68.7^2}} \cong 0.96$$

Esempio 47
Trovare la retta di regressione lineare per i dati riportati nelle prime due colonne della tabella seguente.

Per scrivere il sistema delle equazioni normali ci serviamo della tabella 62.

x_i	y_i	$x_i\, y_i$	x_i^2	y_i^2
0	1.0	0	0	1.0
1	1.4	1.4	1	1.96
3	2.1	6.3	9	4.41
4	2.3	9.2	16	5.29
6	3.0	18.0	36	9.00
14	9.8	34.9	62	21.66

Tabella 62

Il sistema delle equazioni normali è

$$\begin{cases} 62A + 14B = 34.9 \\ 14A + 5B = 9.8 \end{cases}$$

La soluzione con il metodo di Cramer è

$$D = \begin{vmatrix} 62 & 14 \\ 14 & 5 \end{vmatrix} = 114 \qquad D_A = \begin{vmatrix} 34.9 & 14 \\ 9.8 & 5 \end{vmatrix} = 37.3 \qquad D_B = \begin{vmatrix} 62 & 34.9 \\ 14 & 9.8 \end{vmatrix} = 119$$

$$A = \frac{37.3}{114} \cong 0.33 \qquad B = \frac{119}{114} \cong 1.04$$

La retta di regressione ha equazione

$$y = 0.33x + 1.04$$

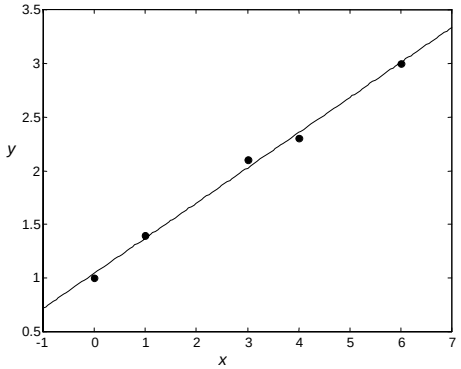


Figura 45

Per calcolare la covarianza e il coefficiente di correlazione per questo insieme di dati si possono usare le formule (1.5) e (1.14), ottenendo (con la tabella precedente)

$$s_x^2 = \frac{1}{4}\left(62 - 5 \cdot 2.8^2\right) = 5.7 \qquad s_y^2 = \frac{1}{4}\left(21.66 - 5 \cdot 1.96^2\right) = 0.613$$

$$S_{xy} = \frac{1}{4}\left(34.9 - 5 \cdot 2.8 \cdot 1.96\right) = 1.865 \qquad r = \frac{1.865}{\sqrt{5.7 \cdot 0.613}} = 0.998$$

Il valore del coefficiente di correlazione indica una relazione di tipo lineare fra i dati.
I coefficienti *A* e *B* della retta di regressione possono in questo caso essere calcolati con le (1.20)

$$A = \frac{S_{xy}}{s_x^2} = \frac{1.865}{5.7} \cong 0.33 \qquad B = \bar{y} - A\bar{x} = 1.96 - 0.33 \cdot 2.8 \cong 1.04$$

1.8 Regressione polinomiale

In molti casi, dopo aver disegnato su un diagramma di dispersione i dati sperimentali $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, si può rilevare una correlazione fra le due variabili osservate, ma non di tipo lineare, ossia i punti appaiono disposti su una curva e non su una retta. Un modo più generale per risolvere il problema di trovare una funzione che approssimi i dati consiste nell'usare come funzione approssimante un polinomio di grado più elevato. Nel caso più semplice del polinomio di secondo grado si trova la **parabola dei minimi quadrati**.
Siano dati i punti $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$; cerchiamo la parabola

$$y = Ax^2 + Bx + C$$

per cui è minima la quantità

$$E = \sum_{i=1}^{n} \left(Ax_i^2 + Bx_i + C - y_i \right)^2 .$$

Si può dimostrare che i coefficienti *A*, *B*, *C* della parabola si trovano risolvendo il **sistema delle equazioni normali**

$$\begin{cases} A\left(\sum_{i=1}^{n} x_i^4\right) + B\left(\sum_{i=1}^{n} x_i^3\right) + C\left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i^2 y_i \\ A\left(\sum_{i=1}^{n} x_i^3\right) + B\left(\sum_{i=1}^{n} x_i^2\right) + C\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} x_i\, y_i \\ A\left(\sum_{i=1}^{n} x_i^2\right) + B\left(\sum_{i=1}^{n} x_i\right) + nC = \sum_{i=1}^{n} y_i \end{cases} \qquad (1.21)$$

Si dimostra che questo sistema possiede soluzione unica, purché i punti non siano tutti allineati verticalmente.

Esempio 48
Troviamo la parabola dei minimi quadrati per i punti (–3,3), (–2,2), (0,1), (2,1), (4,3).

Per ricavare il sistema delle equazioni normali disponiamo i calcoli nella tabella 63

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i\, y_i$	$x_i^2 y_i$
–3	3	9	–27	81	–9	27
–2	2	4	–8	16	–4	8
0	1	0	0	0	0	0
2	1	4	8	16	2	4
4	3	16	64	256	12	48
1	10	33	37	369	1	87

Tabella 63

Il sistema delle equazioni normali è

$$\begin{cases} 369A + 37B + 33C = 87 \\ 37A + 33B + C = 1 \\ 33A + B + 5C = 10 \end{cases}$$

La soluzione è

$$A = \frac{3596}{20176} \cong 0.178 \qquad B = -\frac{3948}{20176} \cong -0.196 \qquad C = \frac{17408}{20176} \cong 0.863$$

La parabola dei minimi quadrati ha equazione (figura 46)

$$y = 0.178x^2 - 0.196x + 0.863$$

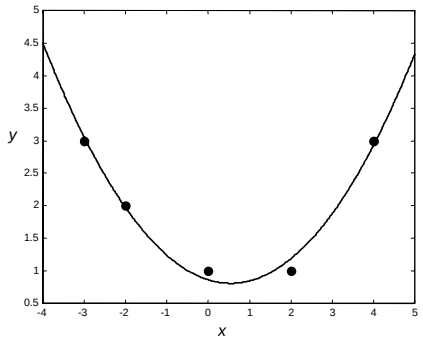


Figura 46

Polinomi di grado più elevato del secondo vengono usati raramente, a meno che sia noto a priori che i dati hanno un andamento di tipo polinomiale; infatti un polinomio di grado *m* ha *m*−1 punti di massimo o minimo, quindi può oscillare molto, specialmente se *m* è elevato.

1.9 Metodi di linearizzazione

Talvolta, nei casi in cui i dati sperimentali non evidenziano una correlazione di tipo lineare, anziché cercare un polinomio ai minimi quadrati, è possibile con un semplice cambiamento di variabile ricondursi alla ricerca della retta di regressione. Tale procedimento è detto **linearizzazione dei dati**. Esaminiamo questo metodo in alcuni casi semplici. Supponiamo che siano assegnati i dati $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ e che la relazione che intercorre tra i dati sia del tipo

$$y = C \cdot x^A \qquad C > 0$$

ossia *y* cresce proporzionalmente a una potenza di *x*.
Prendendo i logaritmi naturali di entrambi i membri si ottiene

$$\ln y = \ln C + A \cdot \ln x$$

e con le sostituzioni

$$X = \ln x \qquad Y = \ln y \qquad B = \ln C$$

si ha

$$Y = AX + B$$

Questa equazione esprime un legame lineare tra le variabili *X* e *Y*.
Si determina perciò la retta di regressione relativa ai dati

$$(X_1, Y_1) = (\ln x_1, \ln y_1), ..., (X_n, Y_n) = (\ln x_n, \ln y_n)$$

e si ricava l'equazione della curva approssimante

$$y = C \cdot x^A \qquad \text{con} \qquad C = e^B$$

dove *A* e *B* sono i coefficienti della retta di regressione $Y = AX + B$.

Esempio 49

Trovare la curva del tipo $y = C \cdot x^A$ che approssima i seguenti dati

x_i	1	1.5	2	2.5	3
y_i	0.9	3.5	7.5	17	30.5

Tabella 64

Linearizziamo i dati con il cambiamento di variabile

$$X = \ln x \qquad Y = \ln y$$

e determiniamo la retta di regressione

$$Y = AX + B\,.$$

x_i	y_i	$X_i = \ln x_i$	$Y_i = \ln y_i$	$X_i Y_i$	X_i^2
1	0.9	0	–0.1054	0	0
1.5	3.5	0.4055	1.2528	0.5080	0.1644
2	7.5	0.6931	2.0149	1.3966	0.4805
2.5	17	0.9163	2.8332	2.5960	0.8396
3	30.5	1.0986	3.4177	3.7548	1.2069
		3.1135	9.4132	8.2554	2.6914

Tabella 65

Il sistema delle equazioni normali è

$$\begin{cases} 2.6914A + 3.1135B = 8.2554 \\ 3.1135A + 5B = 9.4132 \end{cases}$$

ed ha la soluzione

$$A = 3.181 \qquad B = -0.0979$$

$$B = \ln C \quad \Rightarrow \quad C = e^B = e^{-0.0979} = 0.907$$

La funzione che approssima i dati della tabella 64 è

$$y = 0.907 \cdot x^{3.181}$$

Nella figura 47 sono rappresentati i dati e la funzione approssimante.

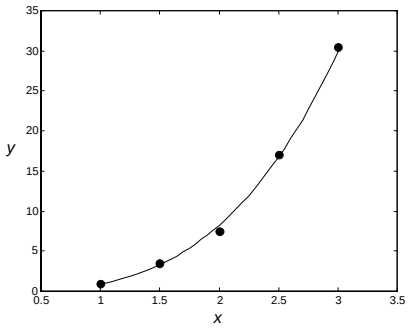


Figura 47

Supponiamo ora che la relazione che intercorre tra i dati

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

sia del tipo

$$y = C \cdot e^{Ax} \qquad C > 0$$

ossia y cresce in modo proporzionale ad una funzione esponenziale.
Prendendo i logaritmi di entrambi i membri si ottiene

$$\ln y = \ln C + A \cdot x$$

e con le sostituzioni

$$X = x \qquad Y = \ln y \qquad B = \ln C$$

si ha

$$Y = AX + B$$

Questa equazione esprime un legame lineare tra le variabili X e Y.
Si determina perciò la retta di regressione relativa ai dati

$$(X_1, Y_1) = (x_1, ln\, y_1), \ldots, (X_n, Y_n) = (x_n, ln\, y_n)$$

e si ricava l'equazione della curva approssimante

$$y = C \cdot e^{Ax} \qquad \text{con} \qquad C = e^B$$

dove A e B sono i coefficienti della retta di regressione $Y = AX + B$.

Esempio 50

Trovare la curva del tipo $y = C \cdot e^{Ax}$ che approssima i seguenti dati

x_i	–1	0	1	2	3
y_i	6.7	4.1	2.1	1.3	0.9

Tabella 66

Linearizziamo i dati con il cambiamento di variabile

$$X = x \qquad Y = ln\, y$$

e determiniamo la retta di regressione

$$Y = AX + B$$

x_i	y_i	$X_i = x_i$	$Y_i = ln\, y_i$	$X_i Y_i$	X_i^2
–1	6.7	–1	1.9021	–1.9021	1
0	4.1	0	1.4110	0	0
1	2.1	1	0.7419	0.7419	1
2	1.3	2	0.2624	0.5247	4
3	0.9	3	–0.1054	–0.3161	9
		5	4.2120	–0.9515	15

Tabella 67

Il sistema delle equazioni normali è

$$\begin{cases} 15A + 5B = -0.9515 \\ 5A + 5B = 4.2120 \end{cases}$$

ed ha la soluzione

$$A = -0.516 \qquad B = 1.359$$

$$B = ln\, C \quad \Rightarrow \quad C = e^B = e^{1.359} = 3.892$$

La funzione che approssima i dati della tabella 66 è

$$y = 3.892 \cdot e^{-0.516x}$$

Nella figura 48 sono rappresentati i dati e la funzione approssimante.

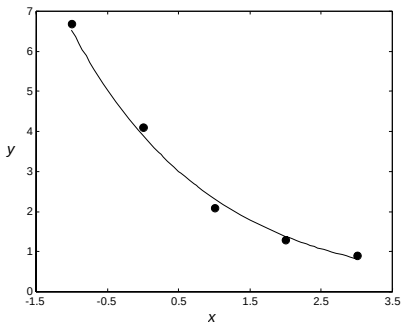


Figura 48

Esempio 51

Trovare la curva del tipo $y = C \cdot e^{Ax}$ che approssima i seguenti dati

x_i	0	1	2	3	4
y_i	1.5	2.5	3.5	5.0	7.5

Tabella 68

Linearizziamo i dati con il cambiamento di variabile

$$X = x \qquad Y = ln\, y$$

e determiniamo la retta di regressione

$$Y = AX + B\,.$$

x_i	y_i	$X_i = x_i$	$Y_i = ln\, y_i$	$X_i Y_i$	X_i^2
0	1.5	0	0.4055	0	0
1	2.5	1	0.9163	0.9163	1
2	3.5	2	1.2528	2.5055	4
3	5.0	3	1.6094	4.8283	9
4	7.5	4	2.0149	8.0596	16
		10	6.1989	16.3097	30

Tabella 69

Il sistema delle equazioni normali è

$$\begin{cases} 30A + 10B = 16.3097 \\ 10A + 5B = 6.1989 \end{cases}$$

ed ha la soluzione

$$A = 0.3912 \qquad B = 0.4574$$

$$B = ln\, C \quad \Rightarrow \quad C = e^B = e^{0.4574} = 1.5800$$

La funzione che approssima i dati della tabella 68 è

$$y = 1.5800 \cdot e^{0.3912x}$$

Nella figura 49 sono rappresentati i dati e la funzione approssimante.

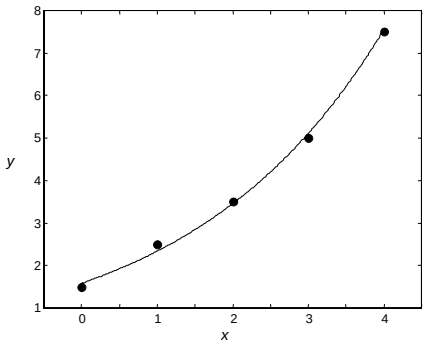


Figura 49

Esempio 52

Siano assegnati i seguenti dati

x_i	1	2	3	4	5
y_i	0.6	1.9	4.3	7.6	12.6

Tabella 70

a – Trovare la curva del tipo $y = C \cdot e^{Ax}$ che approssima i dati;

b – trovare la curva del tipo $y = C \cdot x^A$ che approssima i dati.

c – Usare il criterio dei minimi quadrati per stabilire qual è la curva che approssima meglio i dati.

a – Linearizziamo i dati con il cambiamento di variabile

$$X = x \qquad Y = \ln y$$

e determiniamo la retta di regressione

$$Y = AX + B$$

x_i	y_i	$X_i = x_i$	$Y_i = \ln y_i$	$X_i Y_i$	X_i^2
1	0.6	1	–0.5108	–0.5118	1
2	1.9	2	0.6419	1.2837	4
3	4.3	3	1.4586	4.3758	9
4	7.6	4	2.0281	8.1126	16
5	12.6	5	2.5337	12.6685	25
		15	6.1515	25.9298	55

Tabella 71

Il sistema delle equazioni normali è

$$\begin{cases} 55A + 15B = 25.9298 \\ 15A + 5B = 6.1515 \end{cases}$$

ed ha la soluzione

$$A = 0.747 \qquad B = -1.012$$

$$B = \ln C \quad \Rightarrow \quad C = e^B = e^{-1.012} = 0.363$$

La funzione del tipo $y = C \cdot e^{Ax}$ che approssima i dati della tabella 70 è

$$y = 0.363 \cdot e^{0.747x}$$

b – Linearizziamo i dati con il cambiamento di variabile

$$X = \ln x \qquad Y = \ln y$$

e determiniamo la retta di regressione

$$Y = AX + B$$

x_i	y_i	$X_i = \ln x_i$	$Y_i = \ln y_i$	$X_i Y_i$	X_i^2
1	0.6	0	–0.5108	0	0
2	1.9	0.6931	0.6419	0.4449	0.4805
3	4.3	1.0986	1.4586	1.6025	1.2069
4	7.6	1.3863	2.0281	2.8116	1.9218
5	12.6	1.6094	2.5337	4.0778	2.5903
		4.7875	6.1515	8.9368	6.1995

Tabella 72

Il sistema delle equazioni normali è

$$\begin{cases} 6.1995A + 4.7875B = 8.9368 \\ 4.7875A + 5B = 6.1515 \end{cases}$$

ed ha la soluzione

$$A = 1.886 \qquad B = -0.576$$

$$B = \ln C \quad \Rightarrow \quad C = e^B = e^{-0.576} = 0.562$$

La funzione del tipo $y = C \cdot x^A$ che approssima i dati della tabella 70 è

$$y = 0.562 \cdot x^{1.886}.$$

c – Per stabilire qual è la curva che approssima meglio i dati assegnati ci serviamo del criterio dei minimi quadrati[4] e calcoliamo nei due casi il valore della quantità (errore)

$$E = \sum_{i=1}^{n} \left(AX_i + B - Y_i\right)^2 .$$

La curva che approssima meglio i dati sarà quella per cui il valore di E è più piccolo.
Per la curva trovata al punto a, si ha

$$A = 0.747 \qquad B = -1.012$$

$X_i = x_i$	$Y_i = \ln y_i$	$AX_i + B$	$AX_i + B - Y_i$
1	–0.5108	–0.2650	0.2458
2	0.6419	0.4820	–0.1599
3	1.4586	1.2290	–0.2296
4	2.0281	1.9760	–0.0521
5	2.5337	2.7230	0.1893

Tabella 73

L'errore nel caso a vale

$$E = \sum_{i=1}^{5} \left(AX_i + B - Y_i\right)^2 = 0.2458^2 + 0.1599^2 + 0.2296^2 + 0.0521^2 + 0.1893^2 \cong 0.177$$

Per la curva trovata al punto b, si ha

$$A = 1.886 \qquad B = -0.576$$

$X_i = \ln x_i$	$Y_i = \ln y_i$	$AX_i + B$	$AX_i + B - Y_i$
0	–0.5108	–0.5760	–0.0652
0.6931	0.6419	0.7313	0.0894
1.0986	1.4586	1.4960	0.0374
1.3863	2.0281	2.0386	0.0104
1.6094	2.5337	2.4594	–0.0743

Tabella 74

L'errore nel caso b vale

$$E = \sum_{i=1}^{5} \left(AX_i + B - Y_i\right)^2 = 0.0652^2 + 0.0894^2 + 0.0374^2 + 0.0104^2 + 0.0743^2 \cong 0.019$$

L'approssimazione migliore si ottiene con la curva trovata al punto b.
Nella figura 50 sono rappresentati i dati e le due funzioni approssimanti.

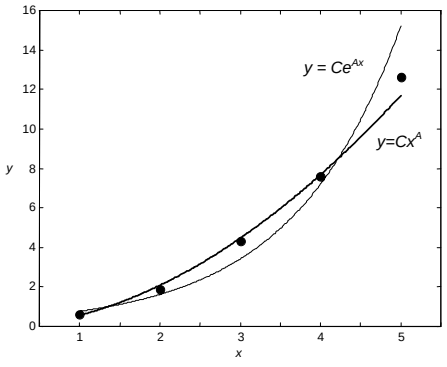


Figura 50

[4] Per scegliere la curva che meglio approssima i dati si può anche calcolare il coefficiente di correlazione R, usando i dati linearizzati $\left(X_i, Y_i\right)$.

La tecnica di linearizzazione può essere usata in molti altri casi; nella tabella 75 sono elencate alcune fra le funzioni approssimanti di uso più comune e i corrispondenti cambiamenti di variabili necessari per linearizzare i dati.

Funzione $y = f(x)$	Forma linearizzata $Y = AX + B$	Cambiamenti di variabili e costanti
$y = C \cdot x^A$	$\ln y = A \ln x + \ln C$	$X = \ln x \quad Y = \ln y$ $C = e^B$
$y = C \cdot e^{Ax}$	$\ln y = A \cdot x + \ln C$	$X = x \quad Y = \ln y$ $C = e^B$
$y = A \ln x + B$	$y = A \ln x + B$	$X = \ln x \quad Y = y$
$y = \frac{A}{x} + B$	$y = A\frac{1}{x} + B$	$X = \frac{1}{x} \quad Y = y$
$y = \frac{1}{Ax + B}$	$\frac{1}{y} = Ax + B$	$X = x \quad Y = \frac{1}{y}$
$y = \frac{x}{A + Bx}$	$\frac{1}{y} = A\frac{1}{x} + B$	$X = \frac{1}{x} \quad Y = \frac{1}{y}$
$y = \frac{D}{x + C}$	$y = -\frac{1}{C}(xy) + \frac{D}{C}$	$X = xy \quad Y = y$ $C = -\frac{1}{A} \quad D = -\frac{B}{A}$
$y = \frac{L}{1 + Ce^{Ax}}$	$\ln\left(\frac{L}{y} - 1\right) = Ax + \ln C$	$X = x \quad Y = \ln\left(\frac{L}{y} - 1\right)$ $C = e^B \quad L =$ costante assegnata
$y = \frac{1}{B + Ae^{-x}}$	$\frac{1}{y} = Ae^{-x} + B$	$X = e^{-x} \quad Y = \frac{1}{y}$

Tabella 75

Esempio 53
Siano assegnati i seguenti dati

x_i	−1	0	1	2	3
y_i	6.62	3.94	2.17	1.35	0.89

Tabella 76

a – Trovare la curva del tipo $y = C \cdot e^{Ax}$ che approssima i dati;

b – trovare la curva del tipo $y = \frac{1}{Ax + B}$ che approssima i dati.

c – Usare il criterio dei minimi quadrati per stabilire qual è la curva che approssima meglio i dati.

a – Linearizziamo i dati con il cambiamento di variabile

$$X = x \quad Y = \ln y$$

e determiniamo la retta di regressione

$$Y = AX + B$$

.

x_i	y_i		$Y_i = ln\, y_i$	$X_i Y_i$	X_i^2
–1	6.62	–1	1.8901	–1.8901	1
0	3.94	0	1.3712	0	0
1	2.17	1	0.7747	0.7747	1
2	1.35	2	0.3001	0.6002	4
3	0.89	3	–0.1165	–0.3496	9
		5	4.2196	–0.8648	15

Tabella 77

Il sistema delle equazioni normali è

$$\begin{cases} 15A + 5B = -0.8648 \\ 5A + 5B = 4.2196 \end{cases}$$

ed ha la soluzione

$$A = -0.508 \qquad B = 1.352$$

$$B = ln\, C \quad \Rightarrow \quad C = e^B = e^{1.352} = 3.865$$

La funzione del tipo $y = C \cdot e^{Ax}$ che approssima i dati della tabella 76 è

$$y = 3.865 \cdot e^{-0.508x}$$

b – Linearizziamo i dati con il cambiamento di variabile (vedere tabella 75)

$$X = x \qquad Y = \frac{1}{y}$$

e determiniamo la retta di regressione

$$Y = AX + B$$

x_i	y_i	$X_i = x_i$	$Y_i = \frac{1}{y_i}$	$X_i Y_i$	X_i^2
–1	6.62	–1	0.1511	–0.1511	1
0	3.94	0	0.2538	0	0
1	2.17	1	0.4608	0.4608	1
2	1.35	2	0.7407	1.4815	4
3	0.89	3	1.1236	3.3708	9
		5	2.7300	5.1620	15

Tabella 78

Il sistema delle equazioni normali è

$$\begin{cases} 15A + 5B = 5.1620 \\ 5A + 5B = 2.7300 \end{cases}$$

ed ha la soluzione

$$A = 0.243 \qquad B = 0.303$$

La funzione del tipo $y = \frac{1}{Ax + B}$ che approssima i dati della tabella 76 è

$$y = \frac{1}{0.243x + 0.303}$$

c – Per stabilire qual è la curva che approssima meglio i dati assegnati ci serviamo del criterio dei minimi quadrati e calcoliamo nei due casi il valore della quantità (errore)

$$E = \sum_{i=1}^{n} \left(AX_i + B - Y_i \right)^2 .$$

La curva che approssima meglio i dati sarà quella per cui il valore di E è più piccolo.
Per la curva trovata al punto a, si ha

$$A = -0.508 \qquad B = 1.352$$

$X_i = x_i$	$Y_i = \ln y_i$	$AX_i + B$	$AX_i + B - Y_i$
–1	1.8901	1.8600	–0.0301
0	1.3712	1.3520	–0.0192
1	0.7747	0.8440	0.0693
2	0.3001	0.3360	0.0359
3	–0.1165	–0.1720	–0.0555

Tabella 79

L'errore nel caso a vale

$$E = \sum_{i=1}^{5} \left(AX_i + B - Y_i \right)^2 = 0.0301^2 + 0.0192^2 + 0.0693^2 + 0.0359^2 + 0.0555^2 \cong 0.0104$$

Per la curva trovata al punto b, si ha

$$A = 0.243 \qquad B = 0.303$$

$X_i = x_i$	$Y_i = \dfrac{1}{y_i}$	$AX_i + B$	$AX_i + B - Y_i$
–1	0.1511	0.0600	–0.0911
0	0.2538	0.3030	0.0492
1	0.4608	0.5460	0.0852
2	0.7407	0.7890	0.0483
3	1.1236	1.0320	–0.0916

Tabella 80

L'errore nel caso b vale

$$E = \sum_{i=1}^{5} \left(AX_i + B - Y_i \right)^2 = 0.0911^2 + 0.0492^2 + 0.0852^2 + 0.0483^2 + 0.0916^2 \cong 0.0287$$

L'approssimazione migliore si ottiene con la curva trovata al punto a.
Nella figura 51 sono rappresentati i dati e le due funzioni approssimanti.

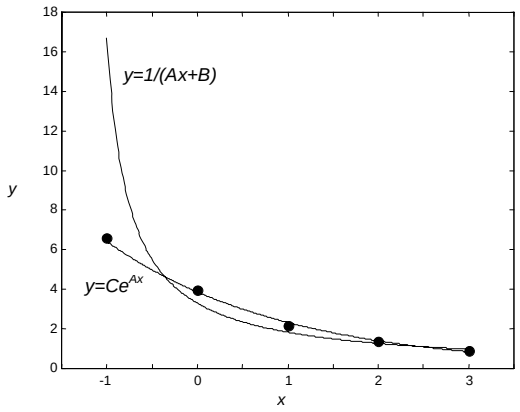


Figura 51

Esempio 54

Quando una popolazione è limitata da un valore limite *L*, la sua crescita è descritta da una funzione avente la forma

$$y = \frac{L}{1 + Ce^{Ax}}$$

La funzione è detta curva logistica.

Trovare *A* e *C* per i dati della tabella 81, con *L* = 1000.

x_i	0	1	2	3	4
y_i	200	400	650	850	950

Tabella 81

b – Linearizziamo i dati con il cambiamento di variabile (vedere tabella 75)

$$X = x \qquad Y = ln\left(\frac{1000}{y} - 1\right)$$

e determiniamo la retta di regressione

$$Y = AX + B$$

x_i	y_i	$X_i = x_i$	$Y_i = ln\left(\frac{1000}{y_i} - 1\right)$	$X_i Y_i$	X_i^2
0	200	0	1.3863	0	0
1	400	1	0.4055	0.4055	1
2	650	2	–0.6190	–1.2381	4
3	850	3	–1.7346	–5.2038	9
4	950	4	–2.9444	–11.7778	16
		10	–3.5063	–17.8142	30

Tabella 82

Il sistema delle equazioni normali è

$$\begin{cases} 30A + 10B = -17.8142 \\ 10A + 5B = -3.5063 \end{cases}$$

ed ha la soluzione

$$A = -1.080 \qquad B = 1.459$$

$$C = e^B = e^{1.459} = 4.302$$

La funzione del tipo $y = \frac{L}{1 + Ce^{Ax}}$ che approssima i dati della tabella 81 è (figura 52)

$$y = \frac{1000}{1 + 4.302\, e^{-1.08x}}$$

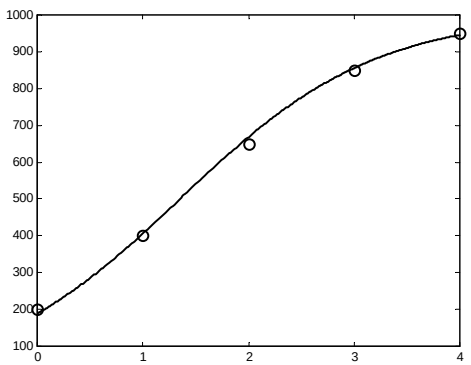


Figura 52

2. Probabilita'

2.1 Esperimenti casuali, spazio dei campioni, eventi

Tutti conoscono l'importanza che hanno gli esperimenti nella scienza e nella tecnologia, ed il fondamentale principio secondo cui, se si esegue ripetutamente un esperimento nelle stesse condizioni, si arriva a risultati che sono essenzialmente uguali.
Ci sono tuttavia esperimenti che, nonostante siano condotti nelle medesime condizioni, possono avere diversi risultati possibili, e il cui risultato non è prevedibile con certezza: **esperimenti** di questo tipo sono detti **casuali**.
Ad esempio nel lancio di una moneta il risultato dell'esperimento può essere T (testa) o C (croce), cioè uno degli elementi dell'insieme {T,C}.
Nel lancio di un dado il risultato può essere uno dei numeri dell'insieme {1,2,3,4,5,6}.
Nell'esperimento consistente in due lanci di una moneta il risultato può essere uno degli elementi dell'insieme {TT,CC,TC,CT}.
Come si osserva dagli esempi, i possibili risultati dell'esperimento si possono esplicitare a priori, ma non si può dire con certezza quale si verificherà.
Un insieme S contenente tutti i possibili risultati di un esperimento casuale è detto **spazio campione**; ciascun risultato è un **elemento** o **punto** di S.
Gli spazi campione vengono classificati in base al numero degli elementi che essi contengono.
Lo spazio campione S corrispondente al lancio di un dado contiene 6 elementi

$$S = \{1,2,3,4,5,6\}$$

e costituisce un esempio di **spazio campione finito**.
Se si considera come evento il numero di volte che un dado deve essere lanciato prima di ottenere un 6, si ha invece uno **spazio campione infinito**: infatti ogni numero intero positivo è un possibile risultato. Il numero degli elementi in questo caso è un'infinità numerabile[1].
Se l'esperimento consiste nel misurare la lunghezza di un segmento, lo spazio S può corrispondere a tutti i punti di un intervallo della retta reale: si ha in questo caso uno **spazio campione continuo**.
Uno **spazio campione** è detto **discreto** se ha un numero finito o un'infinità numerabile di elementi.
Se gli elementi di uno spazio campione costituiscono un insieme continuo, ad esempio i punti di una retta, di una curva, di un piano, lo **spazio campione** è detto **continuo**.
Un **evento** è un sottoinsieme $E \subseteq S$ dello spazio campione S, cioè un insieme di risultati possibili.

Esempio 1
Si effettuano due lanci consecutivi di una moneta; lo spazio campione è l'insieme

$$S = \{TT,CC,TC,CT\}.$$

L'evento che si verifica quando si presenta una sola volta T è il sottoinsieme

$$E_1 = \{TC,CT\}.$$

L'evento che si verifica quando si presenta la prima volta T è

$$E_2 = \{TT,TC\}.$$

Esempio 2
Si estrae una carta a caso da un mazzo di 52 carte; descrivere lo spazio campione quando
a – i semi non sono considerati;
b – i semi sono considerati.

Si indica

1 = asso; 11 = fante; 12 = regina; 13 = re;
C = cuori; Q = quadri; P = picche; F = fiori

[1] Vedere nota pag. 4.

a – S = {1,2,....,9,10,11,12,13}
S contiene 13 elementi.
b – S = {1Q,2Q,....,10Q,11Q,12Q,13Q,1C,...,13C,1P,...,13P,1F,...,13F}
S contiene 52 elementi.

Se il risultato di un esperimento è un elemento di E, si dice che l'evento si è verificato.
Anche l'intero spazio S è un evento: l'**evento** sicuro o **certo**. Ad esempio nel lancio di un dado l'evento certo è che esca uno dei numeri {1,2,3,4,5,6}.
Anche l'insieme vuoto $\varnothing$ è un evento: l'**evento impossibile**.
Dal momento che gli eventi sono insiemi, ogni affermazione concernente gli eventi può essere tradotta nel linguaggio della teoria degli insiemi e viceversa; in particolare avremo un'**algebra degli eventi** corrispondente all'algebra degli insiemi.
Usando le **operazioni insiemistiche** sugli eventi di S si possono ottenere nuovi eventi di S.
Se A e B sono eventi di S, allora
1 – **unione**: $A \cup B$ è l'evento "A oppure B o entrambi";
2 – **intersezione**: $A \cap B$ è l'evento "sia A che B";
3 – **complementare**: $\overline{A}$ è l'evento "non A";

4 – **differenza**: A – B è l'evento "A ma non B".

Definizione 1
Due **eventi** A e B sono **mutuamente esclusivi,** o **incompatibili**, se non possono verificarsi contemporaneamente.

Se gli eventi A e B sono mutuamente esclusivi, essi sono disgiunti, ossia $A \cap B = \varnothing$.
Questi concetti si possono estendere a un numero *k* qualsiasi di eventi. Spesso si illustrano spazi campione ed eventi, in particolare le relazioni fra eventi, con i **diagrammi di Venn**.

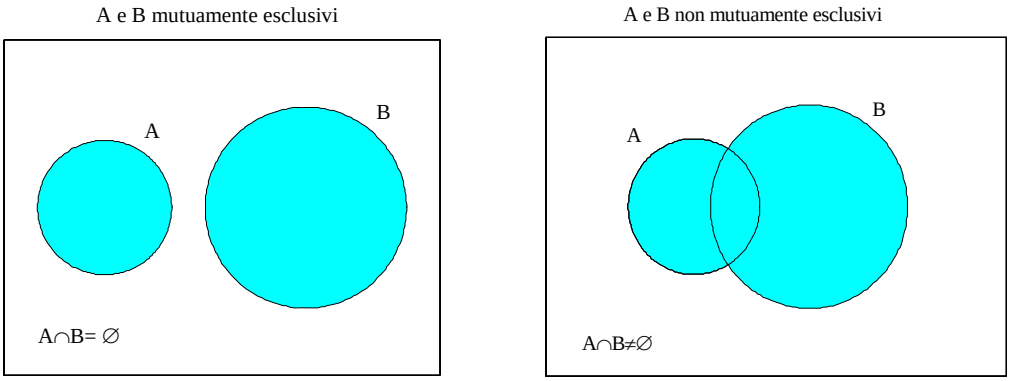


Figura 1

Ricordiamo alcune delle proprietà delle operazioni insiemistiche, valide anche nell'algebra degli eventi.

Proprietà delle operazioni insiemistiche.
Siano A, B, C sottoinsiemi dello spazio S; valgono le proprietà

$1 - A \cup B = B \cup A;\ A \cap B = B \cap A$	proprietà **commutativa** di $\cup$ e $\cap$
$2 - A \cup (B \cup C) = (A \cup B) \cup C;\ A \cap (B \cap C) = (A \cap B) \cap C$	proprietà **associativa** di $\cup$ e $\cap$
$3 - A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	proprietà **distributiva** di $\cup$ rispetto a $\cap$
$4 - A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	proprietà **distributiva** di $\cap$ rispetto a $\cup$
$5 - \overline{A} \cup \overline{B} = \overline{(A \cap B)}$	**legge di De Morgan**
$6 - \overline{A} \cap \overline{B} = \overline{(A \cup B)}$	**legge di De Morgan**

Esempio 3
Si effettuano due lanci di una moneta.
Spazio campione $S = \{TT,CC,TC,CT\}$.
Evento A = "si presenta almeno una T" $A = \{TT,TC,CT\}$
Evento B = "il risultato del secondo lancio è C" $B = \{CC,TC\}$
$A \cup B = \{TT,CC,TC,CT\} = S$ $A \cap B = \{TC\} \neq \varnothing$
$\overline{A} = \{CC\}$ $A - B = \{TT,CT\}$
Gli eventi A e B non sono mutuamente esclusivi.

Esempio 4
Si effettua un lancio di un dado.
Spazio campione $S = \{1,2,3,4,5,6\}$.
Evento A = "uscita di un numero pari" $A = \{2,4,6\}$
Evento B = "uscita di un numero dispari" $B = \{1,3,5\}$
$A \cup B = S \Rightarrow A \cup B$ evento certo
$A \cap B = \varnothing \Rightarrow A \cap B$ evento impossibile
Gli eventi A e B sono mutuamente esclusivi.

Esempio 5
Si estrae una carta a caso da un mazzo di 52 carte; siano dati gli eventi
Evento A = "è uscito un re".
Evento B = "è uscita una carta picche".
Gli eventi sottoelencati si descrivono nel modo seguente:
a – Evento $A \cup B$ = "re o picche o entrambi (cioè re di picche)".
b – Evento $A \cap B$ = "re di picche".
c – Evento $A \cup \overline{B}$ = "re o cuori o quadri o fiori" . Infatti
Evento $\overline{B}$ = "non picche" = evento "cuori o quadri o fiori".
d – Evento $\overline{A} \cup \overline{B}$ = "non re di picche" = "ogni carta diversa dal re di picche". Infatti per la legge di De Morgan (proprietà 5 pag. 60)

$$\overline{A} \cup \overline{B} = \overline{(A \cap B)}$$

e, servendosi del risultato b, $\overline{(A \cap B)}$ = "non re di picche".
e – Evento A – B = "un re, ma non di picche".

2.2 Calcolo Combinatorio

A volte può essere difficile, o almeno noioso, determinare per elencazione diretta gli elementi in uno spazio campione finito. E' preferibile avere dei metodi per contare il numero di tali elementi senza elencarli. Il **calcolo combinatorio** fornisce dei metodi per calcolare il numero di elementi di un insieme. Per illustrare il problema si consideri il seguente esempio.
Esempio 6
Se un uomo ha 3 abiti, 2 camicie e 3 cravatte, quanti modi ha per scegliere una giacca, poi una camicia e infine una cravatta?

Per trattare problemi di questo tipo è utile disegnare un **diagramma ad albero**, dove le alternative per l'abito sono indicate con A_1, A_2, A_3 , per la camicia con C_1, C_2 e per la cravatta con T_1, T_2, T_3

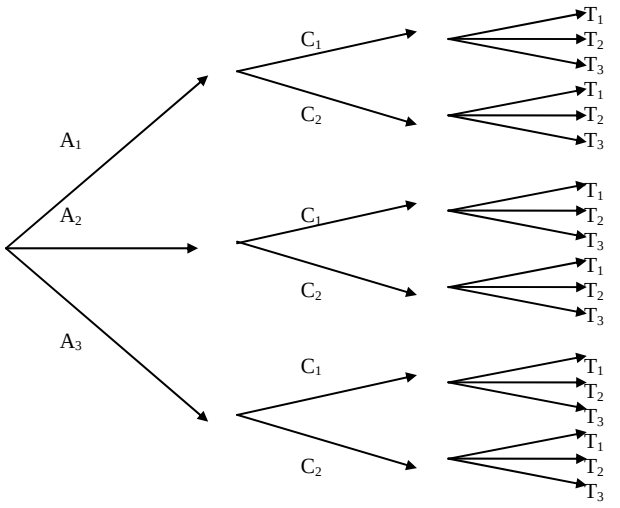


Figura 2

Seguendo un dato cammino da sinistra verso destra lungo i rami dell'albero, si ottiene una particolare scelta, cioè un elemento dello spazio campione, e si può vedere che le possibilità di scelta sono 18. Questo risultato può essere ottenuto osservando che ci sono 3 rami A, che ciascun ramo A si biforca in 2 rami C e che ciascun ramo C si biforca in 3 rami T; ci sono quindi $3 \cdot 2 \cdot 3 = 18$ combinazioni possibili (cammini).
Vale il seguente risultato generale

Teorema 1
Se gli insiemi $A_1, A_2, ..., A_k$ contengono rispettivamente $n_1, n_2, ..., n_k$ oggetti, il numero di modi diversi di scegliere prima un oggetto di A_1, poi un oggetto di A_2,…, infine un oggetto di A_k è

$$N = n_1 \cdot n_2 \cdot ... \cdot n_k \tag{2.1}$$

Esempio 7
In quanti modi diversi una commissione di 25 persone può scegliere un presidente e un vicepresidente?

Il presidente può essere scelto in 25 modi diversi, quindi il vicepresidente in 24 modi diversi; ci sono in tutto

$$N = 25 \cdot 24 = 600$$

modi diversi in cui la scelta richiesta può essere fatta.

Esempio 8
Se un test consiste di 12 domande con risposta Vero-Falso, in quanti modi diversi uno studente può svolgere l'intero test con una risposta per ciascuna domanda?

Poiché a ogni domanda si può rispondere in 2 modi, le possibilità sono in numero di

$$N = \underbrace{2 \cdot 2 \cdot ..... \cdot 2}_{\text{12 fattori}} = 2^{12} = 4096 .$$

Se in particolare $n_1 = n_2 = ... = n_k = n$, si ha $N = n^k$, che rappresenta il numero delle disposizioni con ripetizione di *n* oggetti a gruppi di *k*, ossia dei gruppi che si possono formare scegliendo *k* oggetti, anche ripetibili, fra *n* oggetti disponibili.

Teorema 2
Il numero di **disposizioni con ripetizione di *n* oggetti a gruppi di *k*** è dato da

$$D_{n,k}^{(r)} = n^k \tag{2.2}$$

Esempio 9
Quante parole di 3 lettere (anche senza significato) si possono scrivere con l'alfabeto di 21 lettere?

Le parole sono

aaa, aab, aac, ….. , zzz

Il loro numero è

$$D_{21,3}^{(r)} = 21^3 = 9261$$

Esempio 10
Nella schedina del totocalcio tutti i possibili pronostici sono dati dalle disposizioni con ripetizione dei 3 elementi 1 2 X a gruppi di 13 (i tre simboli si possono ripetere); il loro numero è

$$D_{3,13}^{(r)} = 3^{13} = 1594323$$

Definizione 2
Dati n oggetti distinti, si chiamano **disposizioni semplici (senza ripetizione)** i gruppi che si possono formare scegliendo k ($k \le n$) degli n oggetti; i gruppi devono differire o per qualche oggetto o per l'ordine in cui sono disposti.

Per trovare una formula per il numero delle disposizioni di k oggetti scelti da un insieme di n oggetti distinti, si osservi che la prima scelta è fatta dall'intero insieme di n oggetti, la seconda è fatta fra gli $n - 1$ oggetti rimanenti dopo la prima scelta, in generale la k-esima scelta è fatta fra gli $n - (k - 1) = n - k + 1$ oggetti rimanenti dopo le prime $k - 1$ scelte.
Pertanto, per il teorema 1, il numero delle disposizioni è

$$D_{n,k} = n(n-1)(n-2)....(n-k+1) \tag{2.3}$$

Si può usare la notazione del fattoriale $n! = 1 \cdot 2 \cdot 3 \cdot ... \cdot n$.
Moltiplicando e dividendo nella (2.3) per $(n-k)!$ si ottiene

$$D_{n,k} = \frac{n(n-1)....(n-k+1)(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}$$

Pertanto vale il risultato seguente

Teorema 3
Il numero delle **disposizioni semplici (senza ripetizione)** di k oggetti scelti da un insieme di n oggetti distinti è dato da

$$D_{n,k} = n(n-1)(n-2)....(n-k+1) = \frac{n!}{(n-k)!} \tag{2.4}$$

Esempio 11
Quante parole di 3 lettere diverse si possono formare con l'alfabeto di 21 lettere?

Sono le disposizioni semplici di 21 oggetti diversi a gruppi di 3

$$D_{21,3} = \frac{21!}{18!} = 19 \cdot 20 \cdot 21 = 7980 \ .$$

Esempio 12
In quanti modi 10 persone possono sedersi su una panchina che ha solo 4 posti?

Il numero dei modi è dato dalle disposizioni semplici di 10 elementi a gruppi di 4

$$D_{10,4} = 10 \cdot 9 \cdot 8 \cdot 7 = 5040$$

Esempio 13
In una gara con 40 concorrenti, quante sono le possibili classifiche dei primi tre?

Per il 1° posto possiamo scegliere tra 40 possibilità; per il 2° posto possiamo scegliere fra 39 possibilità e per il 3° posto fra 38 possibilità. In tutto quindi le classifiche possibili per i primi tre sono

$$D_{40,3} = 40 \cdot 39 \cdot 38 = 59280$$

Esempio 14
Trovare quanti numeri di 4 cifre possono essere formati con le 10 cifre 0, 1, 2, ..., 9 se
a – si ammettono delle ripetizioni;
b – non si ammettono ripetizioni;
c – l'ultima cifra deve essere 0 e non si ammettono ripetizioni.

a – la prima cifra può essere una delle 9 cifre 1, 2, ...,9 (lo 0 non è ammesso); le altre tre cifre si scelgono fra le 10 disponibili; si possono allora formare N numeri

$$N = 9 \cdot 10 \cdot 10 \cdot 10 = 9000 .$$

b – la prima cifra può essere una delle 9 cifre 1, 2, ..., 9; per le restanti si devono contare le disposizioni senza ripetizioni

$$D_{9,3} = \frac{9!}{6!} = 7 \cdot 8 \cdot 9 = 504 ;$$

si possono allora formare N numeri

$$N = 9 \cdot 504 = 4536 .$$

c – la prima cifra può essere una delle 9 cifre 1, 2, ..., 9; per la seconda e la terza si devono contare le disposizioni senza ripetizioni

$$D_{8,2} = \frac{8!}{6!} = 7 \cdot 8 = 56$$

(ricordare che la quarta cifra è fissata); si possono quindi formare N numeri

$$N = 9 \cdot 56 = 504 .$$

Nel caso particolare in cui $k = n$ le disposizioni semplici si chiamano permutazioni.

Definizione 3
Le **permutazioni** di n oggetti distinti sono tutti i gruppi formati ciascuno da tutti gli n oggetti dati e che differiscono solo per l'ordine degli oggetti.

Ponendo $k = n$ nella formula delle disposizioni semplici si ottiene il seguente risultato.

Teorema 4
Il numero delle **permutazioni di *n* oggetti distinti** è dato da

$$P_n = n! \qquad (2.5)$$

Esempio 15
Quante parole si possono formare con le 5 vocali?

Il numero delle parole è dato dalle permutazioni di 5 elementi

$$P_5 = 5! = 120.$$

Esempio 16
Si sistemano in uno scaffale 4 libri di matematica, 6 di fisica e 2 di chimica. Contare quante sistemazioni sono possibili se
a – i libri di ogni materia devono stare insieme;
b – solo i libri di matematica devono stare insieme.

a – Numero sistemazioni dei libri di matematica = 4!
Numero sistemazioni dei libri di fisica = 6!
Numero sistemazioni dei libri di chimica = 2!
Numero sistemazioni dei tre gruppi diversi = 3!
Il numero complessivo delle sistemazioni dei libri è quindi

$$N = 4! \cdot 6! \cdot 2! \cdot 3! = 207360$$

b – Si considerano i libri di matematica come un'unica opera.
Restano allora 8 libri (fisica+chimica) + 1 libro (matematica) = 9 libri da sistemare in 9! modi diversi. I libri di matematica hanno 4! sistemazioni diverse, quindi il numero complessivo di sistemazioni diverse è

$$N = 9! \cdot 4! = 8709120$$

Esempio 17
Si fanno sedere 5 uomini e 4 donne in fila: in quanti modi le donne possono occupare i posti pari?

Gli uomini possono essere sistemati in 5! modi diversi (permutazioni), le donne in 4! modi diversi.
Ciascuna sistemazione degli uomini può essere associata ad ogni sistemazione delle donne, quindi il numero complessivo di sistemazioni è

$$N = 5! \cdot 4! = 2880 .$$

Esempio 18
Gli **anagrammi**, cioè le parole che si ottengono da una parola qualunque cambiando solo il posto delle sue lettere, sono permutazioni.
Consideriamo dapprima il caso in cui le parole sono formate da lettere tutte diverse: ad esempio gli anagrammi della parola ROMA sono

$$P_4 = 4 ! = 24$$

Per risolvere il problema degli anagrammi nel caso in cui la parola contenga lettere uguali, occorre disporre di un'altra formula. Supponiamo che un insieme sia formato da *n* oggetti non tutti distinti, dei quali cioè n_1 sono di un tipo (indistinguibili), n_2 di un secondo tipo, ..., n_k del *k*-esimo tipo, con $n_1 + n_2 + ... + n_k = n$.
Si dimostra che

Teorema 5
Il numero delle **permutazioni di *n* oggetti non tutti distinti** è dato da

$$P_{n,n_1,...,n_k} = \frac{n!}{n_1! n_2! ... n_k!} \qquad (2.6)$$

Esempio 19
Contare gli anagrammi della parola MATEMATICA.

Ci sono 10 lettere di cui 2 M, 3 A, 2 T; gli anagrammi sono in numero di

$$N = \frac{10!}{2! \cdot 3! \cdot 2!} = 151200 .$$

Esempio 20
5 palline rosse, 2 bianche e 3 azzurre devono essere sistemate in fila; se tutte le palline dello stesso colore sono indistinguibili, quante sistemazioni sono possibili?

Il numero delle possibili sistemazioni è

$$N = \frac{10!}{5! \cdot 2! \cdot 3!} = 2520 .$$

In una disposizione semplice siamo interessati all'ordine degli oggetti, quindi ad esempio il gruppo "abc" è un gruppo diverso da "bca"; se invece l'ordine di scelta non interessa, cioè "abc" e "bca" sono lo stesso gruppo, si ottengono le **combinazioni**.

Definizione 4
Le **combinazioni** sono tutti i gruppi di *k* oggetti, che si possono formare da un insieme di *n* oggetti distinti, in modo che i gruppi differiscano per almeno un oggetto.

Teorema 6
Il numero delle **combinazioni** di *n* oggetti a gruppi di *k* è dato da

$$\binom{n}{k} = C_{n,k} = \frac{D_{n,k}}{k!} = \frac{n!}{k!(n-k)!} \tag{2.7}$$

I numeri

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)...(n-k+1)}{k!}$$

sono chiamati **coefficienti binomiali**, perché compaiono nello sviluppo della potenza del binomio di Newton $(a+b)^n$.

Esempio 21
Quante squadre di calcio si possono formare con 30 giocatori?

Il numero è dato dalle combinazioni di 11 giocatori scelti nell'insieme di 30

$$C_{30,11} = \binom{30}{11} = \frac{30!}{11!\cdot 19!} = 54627300$$

Esempio 22
In quanti modi 10 oggetti diversi possono essere suddivisi in due gruppi contenenti rispettivamente 4 e 6 oggetti?

Il problema è equivalente a quello di cercare il numero delle scelte di 4 oggetti a partire da 10 (o di 6 a partire da 10), non avendo alcuna importanza l'ordine della scelta; si calcolano perciò le combinazioni

$$C_{10,4} = \binom{10}{4} = \frac{10!}{4!\cdot 6!} = 210$$

Esempio 23
Gioco del poker.
In una mano di poker ogni giocatore riceve 5 delle 52 carte del mazzo. In quanti modi può essere servito?

Il numero dei servizi possibili è dato dalle combinazioni di 5 oggetti scelti fra 52

$$C_{52,5} = \binom{52}{5} = \frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{1\cdot 2\cdot 3\cdot 4\cdot 5} = 2598960$$

Gioco del bridge
In una mano di bridge si ricevono 13 carte su 52. In quanti modi il giocatore può essere servito?

Il numero dei servizi possibili è

$$C_{52,13} = \binom{52}{13} = \frac{52!}{13!\cdot 39!} = 635013559600$$

Esempio 24
Gioco del lotto
Nel gioco del lotto vengono estratti, senza rimetterli ogni volta nell'urna, 5 numeri compresi fra 1 e 90. Le estrazioni avvengono su 10 città o "ruote" diverse, e bisogna precisare su quale ruota si gioca.
a – Trovare il numero di tutte le possibili cinquine relative ad ognuna delle ruote.
b – Quante sono le possibili estrazioni che ci fanno vincere se abbiamo giocato ad esempio l'ambo $\{13\ ,\ 48\}$ su una certa ruota?

a – Il numero di tutte le possibili cinquine è dato dalle combinazioni

$$C_{90,5} = \binom{90}{5} = \frac{90 \cdot 89 \cdot 88 \cdot 87 \cdot 86}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 43949268$$

b – Cerchiamo il numero di cinquine che contengono 13 e 48: gli altri numeri estraibili sono i numeri da 1 a 12, da 14 a 47, da 49 a 90, in tutto 88 numeri; calcoliamo le combinazioni di 88 numeri a gruppi di 3

$$C_{88,3} = \binom{88}{3} = \frac{88 \cdot 87 \cdot 86}{1 \cdot 2 \cdot 3} = 109736\,.$$

Esempio 25
Contare quante sono le diagonali di un poligono convesso.
Un poligono di n lati ha n vertici; ci sono $\binom{n}{2}$ segmenti che uniscono tali vertici; n di questi sono i lati del poligono, perciò il numero delle diagonali è

$$N = \binom{n}{2} - n = \frac{n(n-1)}{2} - n = \frac{n(n-3)}{2}\,.$$

Esempio 26
Quante parole (anche senza significato) di 3 diverse consonanti e 2 diverse vocali si possono formare con l'alfabeto di 21 lettere?
I modi di scegliere le 3 consonanti fra le 16 disponibili sono $\binom{16}{3}$.
I modi di scegliere le 2 vocali fra le 5 disponibili sono $\binom{5}{2}$.
Le 5 lettere risultanti possono essere permutate in 5! modi diversi; allora il numero delle parole possibili è

$$N = \binom{16}{3}\binom{5}{2} \cdot 5! = \frac{16 \cdot 15 \cdot 14}{2 \cdot 3} \cdot \frac{5 \cdot 4}{2} \cdot 5! = 672000$$

2.3 Il concetto di probabilità

Con i metodi del calcolo combinatorio si possono contare gli elementi di un insieme, in altre parole possiamo calcolare quanti sono i casi possibili in una data situazione. In ogni esperimento casuale però non sappiamo se un evento si presenterà o no: bisogna quindi studiare ciò che è probabile o improbabile.
La teoria della probabilità studia concetti e metodi per esprimere quantitativamente il grado di fiducia sul verificarsi di certi eventi. A ciascun evento può essere associata una probabilità, che, dal punto di vista matematico, è una funzione definita sull'insieme degli eventi.

Ci sono più modi mediante i quali è possibile definire la probabilità di un evento: qui definiremo la **probabilità a priori** o **probabilità matematica** e la **probabilità a posteriori** o **probabilità statistica** (o **frequentistica**); è possibile dare un'ulteriore definizione di probabilità, detta **probabilità soggettiva**, che non sarà trattata in queste lezioni.
La definizione classica di **probabilità matematica** *P*, dovuta a Bernoulli e Laplace, è

$$P = \frac{\text{numero casi favorevoli}}{\text{numero casi possibili}}$$

Questa definizione assume che tutti i risultati possibili di un esperimento siano ugualmente probabili e che lo spazio dei campioni sia finito.
La misura della probabilità viene perciò assegnata con il seguente procedimento
1 – si determina il numero di tutti i casi possibili;
2 – si determina il numero dei casi favorevoli, cioè di quei casi che rendono verificato l'evento di cui si vuole calcolare la probabilità;
3 – si calcola il rapporto tra il numero dei casi favorevoli e il numero dei casi possibili.
Secondo questa definizione, ogni probabilità *P* è un numero compreso fra 0 e 1; inoltre la probabilità di un evento che non può accadere (**evento impossibile**) è $P = 0$ e la probabilità di un evento che accade sempre (**evento certo**) è $P = 1$.
Talvolta la probabilità *P* viene moltiplicata per 100 ed espressa in percentuale

$$0\ \% \le P \le 100\ \%.$$

I seguenti esempi illustrano la definizione di probabilità a priori; in alcuni di essi, contrassegnati con un asterisco, si applicano i metodi del calcolo combinatorio.

Esempio 27
Si effettua un lancio di un dado. Calcolare
a – la probabilità di ottenere 2;
b – la probabilità di ottenere un numero dispari.

I casi possibili sono 6 e sono gli elementi dell'insieme {1,2,3,4,5,6}.
a – I casi favorevoli si riducono a 1 (i casi possibili si escludono a vicenda perché può apparire una sola faccia). Pertanto la probabilità cercata è $P = \frac{1}{6}$.

b – I casi favorevoli sono 3. La probabilità cercata è $P = \frac{3}{6} = \frac{1}{2}$.

Esempio 28
Si effettuano due lanci di una moneta. Calcolare la probabilità che si presenti T (testa) almeno una volta.

Casi possibili	TT	TC	CT	CC
Casi favorevoli	TT	TC	CT	

La probabilità cercata è $P = \frac{3}{4}$.

Esempio 29
Si estrae una carta da un mazzo di 52 carte. Calcolare
a – la probabilità di estrarre un asso;
b – la probabilità di estrarre un asso oppure un 10 di cuori oppure un 2 di picche.

a – Nel mazzo ci sono 4 assi, quindi 4 casi favorevoli; la probabilità cercata è $P = \frac{4}{52} = \frac{1}{13}$.

b – Nel mazzo ci sono 4 assi, un 10 di cuori e un 2 di picche, quindi 6 casi favorevoli; la probabilità cercata è $P = \frac{6}{52}$.

**** Esempio 30***
Intorno a un tavolo rotondo si dispongono a caso 5 uomini e 5 donne: calcolare la probabilità che ogni donna si trovi seduta tra due uomini.

Le 10 persone possono disporsi in 10! modi diversi (casi possibili).
Le donne possono disporsi in 5! modi diversi (permutazioni); così anche gli uomini, quindi i casi favorevoli sono $5! \cdot 5!$
La probabilità richiesta vale

$$P = \frac{5! \cdot 5!}{10!} = 0.00397 .$$

**** Esempio 31***
Se su un gruppo di 20 pneumatici, 3 sono difettosi, e si scelgono 4 pneumatici a caso per un controllo di qualità, qual è la probabilità che uno solo di quelli difettosi sia incluso nel gruppo scelto?

I casi possibili sono le combinazioni di 20 oggetti a gruppi di 4; ci sono cioè

$$C_{20,4} = \binom{20}{4} = 4845$$

modi ugualmente probabili di scegliere 4 pneumatici su 20.
Il numero di casi favorevoli è il numero di modi in cui si possono scegliere 3 pneumatici non difettosi e 1 difettoso, cioè

$$C_{17,3} \cdot C_{3,1} = \binom{17}{3} \cdot \binom{3}{1} = 2040$$

Quindi la probabilità è

$$P = \frac{2040}{4845} = \frac{8}{19} \cong 0.42 = 42\%$$

**** Esempio 32***
Determinare la probabilità che, in 4 lanci successivi di un dado, i risultati compaiano in ordine strettamente crescente.

I casi possibili sono le disposizioni con ripetizione di 6 oggetti a gruppi di 4

$$D_{6,4}^{(r)} = 6^4 = 1296$$

I casi favorevoli si hanno quando i risultati dei 4 lanci sono distinti e in ordine crescente.
Il numero di tali casi è dato dal numero delle combinazioni di 6 oggetti a gruppi di 4, perché come gruppo rappresentativo si può scegliere quello in cui i 4 numeri sono disposti in ordine crescente

$$C_{6,4} = \binom{6}{4} = \frac{6!}{4!\,2!} = 15$$

La probabilità cercata è

$$P = \frac{15}{1296} \cong 0.0115$$

**** Esempio 33***
Da un'urna contenente 30 palline, 18 nere e 12 rosse vengono estratte a caso 10 palline.
Determinare la probabilità che 7 fra le palline estratte siano nere.

I casi possibili sono le combinazioni di 30 palline a gruppi di 10

$$C_{30,10} = \binom{30}{10} = 30045015.$$

I casi favorevoli si hanno quando in un gruppo ci sono 7 palline nere e 3 rosse.
Il numero di gruppi di 7 palline nere che si possono formare con 18 palline nere è dato dalle combinazioni

$$C_{18,7} = \binom{18}{7} = 31824.$$

Il numero dei gruppi di 3 palline rosse che si possono formare con 12 palline rosse è dato dalle combinazioni

$$C_{12,3} = \binom{12}{3} = 220.$$

In totale i casi favorevoli sono

$$C_{18,7} \cdot C_{12,3} = \binom{18}{7} \cdot \binom{12}{3} = 7001280.$$

La probabilità cercata è

$$P = \frac{7001280}{30045015} \cong 0.233$$

**** Esempio 34***
Si estraggono 8 palline da un'urna contenente 20 palline numerate da 1 a 20.
Determinare la probabilità che il numero più basso estratto sia 5.
I casi possibili sono le combinazioni di 20 palline a gruppi di 8

$$C_{20,8} = \binom{20}{8} = 125970$$

Se la pallina numerata 5 è la più bassa fra le 8 estratte, allora le rimanenti 7 devono essere numerate da 6 a 20; per trovare i casi favorevoli calcoliamo le combinazioni di 15 elementi a gruppi di 7

$$C_{15,7} = \binom{15}{7} = 6435\,.$$

La probabilità cercata è

$$P = \frac{6435}{125970} \cong 0.051\,.$$

Ci sono molti casi in cui i vari risultati possibili di un esperimento non sono tutti ugualmente probabili. In tal caso si può definire la probabilità per mezzo di una stima frequentistica, possibile solo dopo aver esaminato un gran numero di casi. Si definisce in questo modo la **probabilità a posteriori**, detta anche **probabilità statistica** o **frequentistica.**
Se, dopo aver ripetuto *n* volte un esperimento, con *n* sufficientemente grande, un evento si è verificato *h* volte, si dice che la probabilità di questo evento è $P = \frac{h}{n}$.
Affinché questa definizione sia valida, occorre che tutte le prove avvengano nelle stesse condizioni, cosa che in realtà non è sempre ottenibile quando si analizzano fenomeni statistici.
Se si afferma ad esempio che la probabilità di una nascita di gemelli è $P = \frac{1}{100}$, si intende che la frequenza relativa osservata nell'arco di alcuni anni è stata di 1 su 100; da tale constatazione si può assumere che una nascita futura sarà una nascita di gemelli con probabilità *P* uguale a tale frequenza.

Esempio 35
Si è verificato che su 100 lanci successivi di una moneta, T (testa) si è presentata 56 volte; qual è la probabilità che nel prossimo lancio si presenti C (croce)?

Se T si è presentata 56 volte su 100, allora C si è presentata 44 volte su 100 e la probabilità cercata

è uguale alla frequenza relativa osservata

$$P = \frac{44}{100} = 0.44 .$$

Esempio 36
Si è osservata la durata di un campione di 800 batterie per automobili, ottenendo i dati riportati nella tabella (x indica la durata in anni)

durata	$x < 1$	$1 \le x < 1.5$	$1.5 \le x < 2$	$2 \le x < 2.5$	$2.5 \le x < 3$	$x \ge 3$
numero batterie	61	84	142	247	172	94

Tabella 1

Per una batteria dello stesso tipo e marca si vuole stimare la probabilità relativa a ciascuno dei seguenti eventi
a – Evento A = "la batteria dura almeno tre anni";
b – Evento B = "la batteria dura meno di un anno";
c – Evento C = "la batteria dura almeno due anni".

Se si considera sufficientemente grande il numero di batterie osservate, si può utilizzare il criterio della stima frequentistica della probabilità. Si ottiene così

a – $$P(\mathrm{A}) = \frac{94}{800} = 0.1175 = 11.75\%$$

b – $$P(\mathrm{B}) = \frac{61}{800} = 0.07625 = 7.625\%$$

c – Per calcolare la probabilità dell'evento C occorre considerare il numero delle batterie la cui durata è stata almeno uguale a due anni: 247+172+94 = 513; si ha quindi

$$P(\mathrm{C}) = \frac{513}{800} = 0.6412 = 64.125\%$$

Sia l'approccio classico, sia quello statistico o frequentistico vanno incontro a difficoltà: il primo a causa dell'espressione "ugualmente probabile", il secondo per aver presupposto "n molto grande", concetti di palese vaghezza. A causa di queste difficoltà, si preferisce l'approccio assiomatico alla probabilità, che fa uso degli insiemi.

2.4 Definizione assiomatica di probabilità

Sia S uno spazio campione finito. Ad ogni evento A di S si associa un numero reale $P(\mathrm{A})$, detto **probabilità dell'evento A**, che soddisfa i seguenti assiomi
1 – $0 \le P(\mathrm{A}) \le 1$
2 – $P(\mathrm{S}) = 1$
3 – Se A e B sono eventi mutuamente esclusivi di S (cioè $\mathrm{A} \cap \mathrm{B} = \varnothing$), allora

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}).$$

P è una funzione definita sull'insieme degli eventi di S e a valori reali, detta **funzione di probabilità**, che a ogni sottoinsieme A di S associa un numero reale

$$P : \mathrm{A} \subseteq \mathrm{S} \to P(\mathrm{A}) \in \mathbf{R} .$$

Dal 1° assioma segue che $P(\mathrm{A})$ è un numero reale appartenente all'intervallo [0,1]; dal 2° assioma segue che la probabilità dell'evento certo è 1; dal 3° assioma segue che le funzioni di probabilità sono funzioni additive.
Gli assiomi non devono naturalmente essere dimostrati, ma si può mostrare che essi sono coerenti con la definizione classica di probabilità.

Esempio 37
Un esperimento ha tre soli possibili risultati *a*, *b*, e *c*; in ciascuno dei casi seguenti verificare se i valori assegnati alle probabilità sono accettabili

1 – $P(a) = \frac{1}{3}$, $P(b) = \frac{1}{3}$, $P(c) = \frac{1}{3}$

2 – $P(a) = 0.64$, $P(b) = 0.38$, $P(c) = -0.02$

3 – $P(a) = 0.35$, $P(b) = 0.52$, $P(c) = 0.26$

1 – I valori assegnati alle probabilità sono accettabili, perché sono compresi nell'intervallo [0,1] e la loro somma vale 1.
2 – Il valore di $P(c) = -0.02$ non è accettabile perché negativo.
3 – I valori non sono accettabili perché la loro somma è $0.35+0.52+0.26 = 1.13 > 1$.

Elenchiamo alcuni **teoremi elementari** che seguono dagli assiomi appena enunciati.
Il teorema 7 è una generalizzazione del terzo assioma.

Teorema 7
Se $A_1, A_2, ..., A_n$ sono eventi mutuamente esclusivi di uno spazio campione S, allora

$$P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n) \quad (2.8)$$

Il teorema 8 consente di calcolare la probabilità dell'unione di due eventi qualsiasi, anche nel caso in cui gli eventi non sono necessariamente mutuamente esclusivi.

Teorema 8 – Regola additiva
Se A e B sono due eventi qualsiasi di S, allora

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2.9)$$

Di questo teorema si può dare una semplice rappresentazione grafica con i diagrammi di Venn.

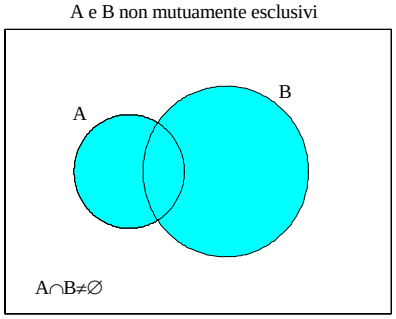


Figura 3

Dal grafico si vede che, sommando semplicemente $P(A)$ e $P(B)$, la probabilità $P(A \cap B)$ viene contata due volte. Se gli eventi sono mutuamente esclusivi, il teorema 8 si riduce al terzo assioma della definizione.

Teorema 9
Se A è un qualunque evento di S, allora

$$P(\overline{A}) = 1 - P(A) \quad (2.10)$$

In particolare l'evento impossibile ha probabilità nulla

$$P(\varnothing) = 0.$$

Esempio 38
Siano A e B due eventi mutuamente esclusivi, con $P(A) = 0.5$ e $P(A \cup B) = 0.6$. Calcolare $P(B)$.

Poiché gli eventi sono mutuamente esclusivi, si ha

$$P(A \cup B) = P(A) + P(B)$$

quindi

$$P(B) = P(A \cup B) - P(A) = 0.6 - 0.5 = 0.1$$

Esempio 39
Una pallina viene estratta da un'urna che ne contiene 6 rosse, 4 bianche e 5 nere. Calcolare la probabilità che la pallina estratta sia
a – rossa;
b – bianca;
c – nera;
d – non rossa;
e – rossa o bianca.

a – Casi possibili: $6 + 4 + 5 = 15$ Casi favorevoli: 6

$$P(\text{rossa}) = \frac{6}{15} = \frac{2}{5}$$

b – $$P(\text{bianca}) = \frac{4}{15}$$

c – $$P(\text{nera}) = \frac{5}{15} = \frac{1}{3}$$

d – $$P(\text{non rossa}) = 1 - P(\text{rossa}) = 1 - \frac{2}{5} = \frac{3}{5}$$

e – $$P(\text{rossa} \cup \text{bianca}) = P(\text{rossa}) + P(\text{bianca}) = \frac{2}{5} + \frac{4}{15} = \frac{10}{15} = \frac{2}{3}$$

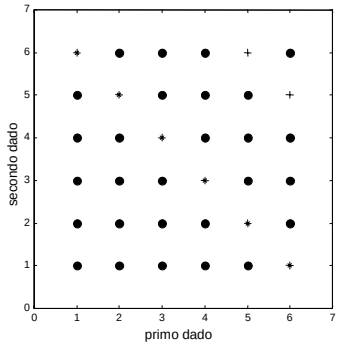
(rossa e bianca sono eventi mutuamente esclusivi)

Esempio 40
Trovare la probabilità di non ottenere come somma del lancio di due dadi né 7 né 11.

Lo spazio campione S è costituito da 36 coppie di numeri, che rappresentano le possibili uscite su ciascuno dei due dadi

$$S = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,1),\ldots,(6,6)\}$$

I punti del grafico che segue rappresentano l'insieme S



* = somma 7
\+ = somma 11

Figura 4

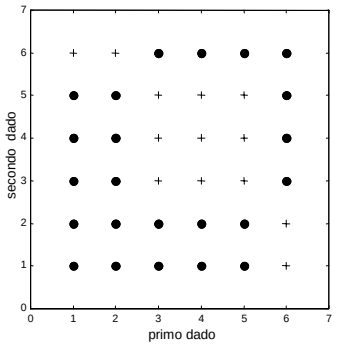
Evento A = " somma uguale a 7 oppure a 11"
Evento $\overline{A}$ = " somma né 7 né 11"

$$P(\overline{A}) = 1 - P(A) = 1 - \frac{8}{36} = \frac{7}{9}$$

Esempio 41
Due dadi hanno le facce numerate nel modo seguente

1 1 2 2 2 3

Trovare la probabilità che il punteggio totale sia
a – uguale a 4;
b – minore di 4;
c – maggiore di 4.



+ = somma 4

Figura 5

a – Casi possibili: 36. Casi favorevoli: 13.

La probabilità che il punteggio totale sia uguale a 4 è $P = \frac{13}{36}$.

b – Casi possibili: 36. Casi favorevoli: 16.

La probabilità che il punteggio totale sia minore di 4 è $P = \frac{16}{36} = \frac{4}{9}$.

c – La probabilità che il punteggio totale sia minore o uguale a 4 è $P = \frac{13}{36} + \frac{4}{9} = \frac{29}{36}$, quindi la probabilità che il punteggio sia maggiore di 4 è $P = 1 - \frac{29}{36} = \frac{7}{36}$.

Esempio 42
Si effettua il lancio di un dado. Calcolare
a – la probabilità che esca un 2 oppure un 5;
b – la probabilità che esca un numero pari;
c – la probabilità che esca un numero divisibile per 3.
d – Dati gli eventi
Evento A_1 = "esce 1 oppure 2" $A_1 = \{1,2\}$
Evento A_2 = "esce 2 oppure 3" $A_2 = \{2,3\}$
calcolare $P(A_1 \cup A_2)$.

a – Si ha

$$P(1) = P(2) = ... = P(6) = \frac{1}{6}$$

L'evento che si verifica quando esce un 2 o un 5 si indica con $2 \cup 5$

$$P(2 \cup 5) = P(2) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

b – $$P(2 \cup 4 \cup 6) = P(2) + P(4) + P(6) = \frac{1}{2}$$

c – $$P(3 \cup 6) = P(3) + P(6) = \frac{1}{3}$$

d – Gli eventi $A_1 = \{1,2\}$ e $A_2 = \{2,3\}$ non sono mutuamente esclusivi, poiché

$$A_1 \cap A_2 = \{2\} \neq \varnothing .$$

Si ha

$$A_1 \cup A_2 = \{1,2,3\}$$

$$P(A_1) = P(A_2) = \frac{1}{3}$$

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = \frac{1}{3} + \frac{1}{3} - \frac{1}{6} = \frac{1}{2} .$$

Esempio 43

Si estrae una carta a caso da un mazzo di 52 carte. Calcolare la probabilità che sia
a – un asso;
b – un fante di cuori;
c – un 3 di picche o un 6 di fiori;
d – un cuori;
e – un seme diverso da cuori;
f – un 10 o un quadri;
g – né un 4 né un picche.

Si usano le notazioni

1 = asso, ... , 11 = fante, 12 = regina, 13 = re,
C = cuori, Q = quadri, P = picche, F = fiori.

a – $$P(1) = \frac{4}{52}$$

b – $$P(11 \cap \mathrm{C}) = \frac{1}{52}$$

c – $$P((13 \cap \mathrm{P}) \cup (6 \cap \mathrm{F})) = P(13 \cap \mathrm{P}) + P(6 \cap \mathrm{F}) = \frac{1}{52} + \frac{1}{52} = \frac{1}{26}$$

d – $$P(\mathrm{C}) = \frac{13}{52} = \frac{1}{4}$$

e – $$P(\overline{\mathrm{F}}) = 1 - P(\mathrm{F}) = 1 - \frac{1}{4} = \frac{3}{4}$$

f – 10 e quadri non sono mutuamente esclusivi, quindi

$$P(10 \cup \mathrm{Q}) = P(10) + P(\mathrm{Q}) - P(10 \cap \mathrm{Q}) = \frac{1}{13} + \frac{1}{4} - \frac{1}{52} = \frac{4}{13}$$

g – P(né 4 né picche) = $P(\overline{4} \cap \overline{\mathrm{P}})$

Per la legge di De Morgan (proprietà 6, pag. 60) si ha

$$P(\overline{4} \cap \overline{\mathrm{P}}) = P(\overline{(4 \cup \mathrm{P})}) = 1 - P(4 \cup \mathrm{P}) =$$

$$= 1 - \left[P(4) + P(\mathrm{P}) - P(4 \cap \mathrm{P})\right] = 1 - \left[\frac{1}{13} + \frac{1}{4} - \frac{1}{52}\right] = \frac{9}{13}$$

(si ricordi che gli eventi 4 e P non sono mutuamente esclusivi).

Esempio 44

Supponiamo che i pezzi prodotti da una certa macchina possano avere due tipi di difetti. E' noto che la probabilità che un pezzo presenti il primo difetto è 0.1, la probabilità che non presenti il secondo difetto è 0.8, la probabilità che li presenti entrambi è 0.01.
Calcolare la probabilità che un pezzo non abbia alcun difetto.
Evento A = "è presente il primo difetto"
Evento B = "è presente il secondo difetto".

Dai dati del problema si ha

$$P(\mathrm{A}) = 0.1 \qquad P(\overline{\mathrm{B}}) = 0.8 \qquad P(\mathrm{A} \cap \mathrm{B}) = 0.01$$

Si deve calcolare $P(\overline{\mathrm{A}} \cap \overline{\mathrm{B}})$.

$$P(\overline{\mathrm{A}}) = 0.9 \qquad P(\mathrm{B}) = 0.2$$

Applicando la regola additiva (teorema 8) si ha

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A} \cap \mathrm{B}) = 0.1 + 0.2 - 0.01 = 0.29$$

Per la legge di De Morgan (proprietà 6, pag. 560) si ha

$$P(\overline{\mathrm{A}} \cap \overline{\mathrm{B}}) = \mathrm{P}(\overline{\mathrm{A} \cup \mathrm{B}}) = 1 - P(\mathrm{A} \cup \mathrm{B}) = 1 - 0.29 = 0.71 .$$

Esempio 45

Se in una stanza sono presenti n persone qual è la probabilità che nessuna di esse festeggi il compleanno nello stesso giorno dell'anno?

Evento A = " tutti compiono gli anni in giorni diversi".

Per calcolare i casi possibili osserviamo che ogni persona può compiere gli anni in uno qualsiasi dei 365 giorni dell'anno (non consideriamo il caso particolare degli anni bisestili), perciò per n persone si hanno complessivamente 365^n casi possibili.

I casi favorevoli si hanno quando tutti compiono gli anni in giorni diversi; la prima persona ha 365 possibilità, la seconda persona 364 possibilità, …, l'n-esima persona ha 365–(n–1) possibilità; complessivamente i casi favorevoli sono

$$365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - (n-1)).$$

Si ha quindi

$$P(\mathrm{A}) = \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - (n-1))}{365^n}.$$

Nella tabella seguente riportiamo i valori della probabilità per vari valori di n

n	10	20	23	30	40	50	60	70	80
P(A)	0.8831	0.5886	0.4927	0.2937	0.1088	0.0296	0.0059	0.0008	0.000085

Tabella 2

Dalla tabella si vede che se n = 23 la probabilità è minore di 0.5; questo significa che se nella stanza ci sono 23 persone, la probabilità che almeno due di esse compiano gli anni nello stesso giorno è maggiore di 0.5; questa probabilità diventa 0.9704 se nella stanza ci sono 50 persone. Questi risultati possono apparire abbastanza sorprendenti.

2.5 Probabilità condizionata

La probabilità di un evento è un numero che misura il grado di fiducia che noi abbiamo circa il realizzarsi di questo evento. E' naturale allora che la probabilità di uno stesso evento possa cambiare, se cambiano le informazioni in nostro possesso.

Il concetto di probabilità condizionata traduce formalmente l'idea intuitiva di probabilità di un evento, calcolata sapendo che si è verificato un altro evento.

Esempio 46

Si effettua un lancio di un dado; consideriamo i seguenti eventi

Evento A = "esce un numero dispari" A = {1,3,5}

Evento B = "esce un numero minore di 4" B = {1,2,3}.

Calcoliamo la probabilità di ottenere un numero minore di 4, sapendo che il risultato è un numero dispari.

La probabilità dell'evento A vale

$$P(\mathrm{A}) = \frac{1}{2}$$

poiché i casi possibili sono 6 e i casi favorevoli sono 3. Analogamente per l'evento B

$$P(\mathrm{B}) = \frac{1}{2}$$

Se sappiamo che l'evento A si è già verificato, i casi possibili per l'evento B non sono più 6, ma si riducono a 3 (ossia la conoscenza del verificarsi dell'evento A riduce lo spazio campione), e i casi favorevoli sono 2, perciò la probabilità di ottenere un numero minore di 4, sapendo che il risultato è dispari, è $\frac{2}{3}$.

La probabilità così ottenuta è detta **probabilità condizionata**

$$P(\mathrm{B}|\mathrm{A}) = \frac{2}{3}$$

(il simbolo | si legge "a condizione che").

Il fatto di aggiungere l'informazione che il numero estratto è dispari, fa aumentare la probabilità di B da $\frac{1}{2}$ a $\frac{2}{3}$.

Osserviamo che si ha

$$\mathrm{A}\cap\mathrm{B} = \{1,3\} \qquad P(\mathrm{A}\cap\mathrm{B}) = \frac{2}{6} = \frac{1}{3}$$

e che

$$P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{A}\cap\mathrm{B})}{P(\mathrm{A})} = \frac{2}{3}$$

Queste considerazioni vengono formalizzate dalla seguente definizione.

Definizione 5

Siano A e B due eventi qualsiasi dello spazio campione S e sia $P(\mathrm{A}) \neq 0$.

La probabilità dell'evento B, nell'ipotesi che si sia già verificato l'evento A, è chiamata **probabilità** di B **condizionata** ad A ed è definita da

$$P(\mathrm{B}\,|\,\mathrm{A}) = \frac{P(\mathrm{A}\cap\mathrm{B})}{P(\mathrm{A})} \tag{2.11}$$

Analogamente, se $P(\mathrm{B}) \neq 0$, la probabilità di A condizionata a B è definita da

$$P(\mathrm{A}\,|\,\mathrm{B}) = \frac{P(\mathrm{A}\cap\mathrm{B})}{P(\mathrm{B})} \tag{2.12}$$

Il seguente risultato è una conseguenza immediata della definizione di probabilità condizionata.

Teorema 10 – Regola di moltiplicazione

$$P(\mathrm{A}\cap\mathrm{B}) = P(\mathrm{A})\cdot P(\mathrm{B}\,|\,\mathrm{A}) \qquad \text{se } P(\mathrm{A}) \neq 0 \tag{2.13}$$

$$P(\mathrm{A}\cap\mathrm{B}) = P(\mathrm{B})\cdot P(\mathrm{A}\,|\,\mathrm{B}) \qquad \text{se } P(\mathrm{B}) \neq 0 \tag{2.14}$$

Questo significa che la probabilità del verificarsi di entrambi gli eventi A e B è uguale alla probabilità di A per la probabilità che B si verifichi, quando si supponga che A si sia già verificato.

Esempio 47

Data un'urna contenente 15 palline rosse e 5 palline nere, indichiamo con A l'evento "estrazione di pallina rossa" e con B l'evento "estrazione di pallina nera". Calcoliamo la probabilità di ottenere in due estrazioni consecutive prima una pallina rossa e poi una nera, nell'ipotesi che la prima pallina estratta non venga rimessa nell'urna.

La probabilità di estrarre una pallina rossa alla prima estrazione è

$$P(\mathrm{A}) = \frac{15}{20} = \frac{3}{4}$$

La probabilità di estrarre una pallina nera dopo aver già estratto una pallina rossa, che non viene rimessa nell'urna prima di effettuare la seconda estrazione, è $\frac{5}{19}$. Infatti ci sono soltanto più 19 palline nell'urna fra le quali estrarre la seconda. Pertanto la probabilità condizionata vale

$$P(\mathrm{B}\,|\,\mathrm{A}) = \frac{5}{19}$$

La probabilità $P(\mathrm{A}\cap\mathrm{B})$ di ottenere in due estrazioni consecutive una pallina rossa e poi una nera, senza rimettere nell'urna la rossa già estratta, in base alla (2.13) è

$$P(\mathrm{A}\cap\mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B}\,|\,\mathrm{A}) = \frac{3}{4} \cdot \frac{5}{19} = \frac{15}{76} = 0.1974\,.$$

Se invece la prima pallina estratta venisse rimessa nell'urna, la probabilità di ottenere in due estrazioni consecutive prima una pallina rossa e poi una nera sarebbe

$$P(\mathrm{A}\cap\mathrm{B}) = \frac{15}{20} \cdot \frac{5}{20} = \frac{3}{16} = 0.1875\,.$$

Esempio 48

Qual è la probabilità che, lanciando una moneta 5 volte, non esca mai "croce"?
Qual è la probabilità dello stesso evento, supponendo di aver già lanciato la moneta 4 volte e di aver ottenuto sempre "testa"?

a – Sia A l'evento "in 5 lanci non esce mai croce"; il numero dei casi possibili, ossia delle possibili sequenze di 5 lanci, è 2^5; c'è un unico caso favorevole, quindi

$$P(\mathrm{A}) = \frac{1}{2^5} = \frac{1}{32} = 0.03125$$

b – Sia B l'evento "nei primi 4 lanci non è mai uscita croce"; come prima si ha

$$P(\mathrm{B}) = \frac{1}{2^4}$$

La probabilità di A, sapendo che si è verificato B, è

$$P(\mathrm{A}\,|\,\mathrm{B}) = \frac{P(\mathrm{A}\cap\mathrm{B})}{P(\mathrm{B})} = \frac{P(\mathrm{A})}{P(\mathrm{B})} = \frac{\frac{1}{2^5}}{\frac{1}{2^4}} = \frac{1}{2}$$

Si noti che $\mathrm{A} \subseteq \mathrm{B}$, perciò $\mathrm{A} \cap \mathrm{B} = \mathrm{A}$.
Possiamo osservare come l'informazione ulteriore in nostro possesso abbia cambiato in modo evidente la valutazione della probabilità di uno stesso evento.

Può però accadere che la probabilità condizionata $P(\mathrm{B}\,|\,\mathrm{A})$ sia uguale alla probabilità $P(\mathrm{B})$; questa condizione significa intuitivamente che sapere che A si è verificato non cambia la valutazione della probabilità di B. In questo caso si dà la seguente definizione.

Definizione 6
Due **eventi** A e B si dicono **indipendenti** se

$$P(\mathrm{B}\,|\,\mathrm{A}) = P(\mathrm{B})$$

In tal caso si ha pure

$$P(\mathrm{A}\,|\,\mathrm{B}) = P(\mathrm{A})$$

Nel caso di due eventi indipendenti, il teorema 10 diventa

Teorema 11 – Regola di moltiplicazione per eventi indipendenti
Se due eventi A e B sono indipendenti, si ha

$$P(A\cap B) = P(A)\cdot P(B) \qquad (2.15)$$

Questa regola viene spesso assunta come definizione di eventi indipendenti; in ogni caso può essere usata per determinare se due eventi sono indipendenti.

Esempio 49
Qual è la probabilità di ottenere due volte testa in due lanci successivi di una moneta?

Poiché la probabilità di ottenere T è $P(T) = \frac{1}{2}$ per ciascun lancio e i due lanci sono indipendenti, la probabilità di ottenere due volte testa è

$$P(TT) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}.$$

Esempio 50
Si lancia due volte un dado. Calcolare la probabilità di ottenere 4, 5 o 6 al primo lancio e 1, 2, 3 o 4 al secondo.

Siano

$$A = \{4,5,6\} \qquad B = \{1,2,3,4\}$$

Si deve calcolare la probabilità $P(A\cap B)$.
Il risultato del secondo lancio è indipendente dal primo, cioè i due eventi A e B sono indipendenti, perciò

$$P(A\cap B) = P(A)\cdot P(B) = \frac{3}{6}\cdot\frac{4}{6} = \frac{1}{3}.$$

Esempio 51
Trovare la probabilità che in due lanci di un dado si presenti almeno una volta il 5.

Evento A = "5 al primo lancio"
Evento B = "5 al secondo lancio"
Evento $A\cup B$ = " 5 al primo oppure al secondo lancio" .
Gli eventi non sono mutuamente esclusivi, perciò per il teorema 8 si ha

$$P(A\cup B) = P(A) + P(B) - P(A\cap B).$$

Per calcolare $P(A\cap B)$ osserviamo che gli eventi A e B sono indipendenti, perciò

$$P(A\cap B) = P(A)\cdot P(B)$$

quindi

$$P(A\cup B) = P(A) + P(B) - P(A\cap B) = P(A) + P(B) - P(A)\cdot P(B) =$$
$$= \frac{1}{6} + \frac{1}{6} - \frac{1}{6}\cdot\frac{1}{6} = \frac{11}{36}$$

Esempio 52
Le probabilità che un marito e una moglie siano vivi tra 20 anni sono rispettivamente 0.8 e 0.9 .
Trovare la probabilità che tra 20 anni
a – entrambi siano vivi;
b – nessuno dei due lo sia;
c – almeno uno dei due sia vivo.

Evento M = "marito vivo"
Evento D = "moglie viva".

Supponiamo che gli eventi siano indipendenti (ipotesi che potrebbe anche non essere ragionevole).

a – P(entrambi vivi) = $P(\mathrm{M} \cap \mathrm{D}) = P(\mathrm{M}) \cdot P(\mathrm{D}) = 0.8 \cdot 0.9 = 0.72$

b – P(nessuno vivo) = $P(\overline{\mathrm{M}} \cap \overline{\mathrm{D}}) = P(\overline{\mathrm{M}}) \cdot P(\overline{\mathrm{D}}) = 0.2 \cdot 0.1 = 0.02$

c – P(almeno uno vivo) = 1 – P(nessuno vivo) = 1 – 0.02 = 0.98

Esempio 53

Si estraggono due carte da un mazzo di 52 carte. Calcolare la probabilità di estrarre due assi se

a – la prima carta viene rimessa nel mazzo prima della seconda estrazione;

b – la prima carta non viene rimessa nel mazzo prima della seconda estrazione.

a – In questo caso gli eventi sono indipendenti; ci sono 4 assi nel mazzo, quindi

$$P = \frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$$

b – In questo caso gli eventi sono dipendenti; fra le 51 carte rimaste dopo l'estrazione del primo asso ci sono solo più 3 assi, quindi la probabilità di estrarre uno di questi è $\frac{3}{51}$; la probabilità richiesta è

$$P = \frac{4}{52} \cdot \frac{3}{51} = \frac{1}{221}$$

**** Esempio 54***

Un'urna contiene 8 palline rosse, 3 palline bianche e 9 palline nere. Si estraggono tre palline a caso senza rimetterle nell'urna dopo ogni estrazione. Determinare le probabilità che siano

a – tre rosse;

b – tre bianche;

c – almeno una bianca;

d – una per ciascun colore, senza tenere conto dell'ordine di estrazione;

e – due rosse e una nera, senza tenere conto dell'ordine di estrazione;

f – una rossa, una bianca e una nera, nell'ordine.

Evento R_1 = "rossa alla prima estrazione"
Evento B_1 = "bianca alla prima estrazione"
Evento N_1 = "nera alla prima estrazione"
Evento R_2 = "rossa alla seconda estrazione"
……

a – Evento $R_1 \cap R_2 \cap R_3$ = "tre rosse"

$$P(\mathrm{R}_1 \cap \mathrm{R}_2 \cap \mathrm{R}_3) = P(\mathrm{R}_1) \cdot P(\mathrm{R}_2 \mid \mathrm{R}_1) \cdot P(\mathrm{R}_3 \mid \mathrm{R}_1 \cap \mathrm{R}_2) =$$

$$= \frac{8}{20} \cdot \frac{7}{19} \cdot \frac{6}{18} = \frac{14}{285} \cong 0.049$$

b – Evento $B_1 \cap B_2 \cap B_3$ = "tre bianche"

$$P(\mathrm{B}_1 \cap \mathrm{B}_2 \cap \mathrm{B}_3) = P(\mathrm{B}_1) \cdot P(\mathrm{B}_2 \mid \mathrm{B}_1) \cdot P(\mathrm{B}_3 \mid \mathrm{B}_1 \cap \mathrm{B}_2) =$$

$$= \frac{3}{20} \cdot \frac{2}{19} \cdot \frac{1}{18} = \frac{1}{1140} \cong 0.00088$$

c – P("almeno una bianca") = 1 – P("nessuna bianca")

$$P(\text{"nessuna bianca"}) = \frac{\mathrm{C}_{17,3}}{\mathrm{C}_{20,3}} = \frac{\binom{17}{3}}{\binom{20}{3}} = \frac{34}{57} \cong 0.596$$

$$P(\text{“almeno una bianca”}) = 1 - \frac{34}{57} = \frac{23}{57} \cong 0.404$$

d – Non si tiene conto dell’ordine di estrazione

$$P(\text{“una rossa, una bianca e una nera”}) = \frac{C_{8,1} \cdot C_{3,1} \cdot C_{9,1}}{C_{20,3}} = \frac{18}{95} \cong 0.189$$

e – Non si tiene conto dell’ordine di estrazione

$$P(\text{“due rossa e una nera”}) = \frac{C_{8,2} \cdot C_{9,1}}{C_{20,3}} = \frac{21}{95} \cong 0.221$$

f – Si tiene conto dell’ordine di estrazione

$$P(R_1 \cap B_2 \cap N_3) = P(R_1) \cdot P(B_2 \mid R_1) \cdot P(N_3 \mid R_1 \cap B_2) =$$
$$= \frac{8}{20} \cdot \frac{3}{19} \cdot \frac{9}{18} = \frac{3}{95} \cong 0.0316$$

Si noti che i quesiti d, ed e non possono essere risolti con la tecnica del quesito f, perché non è noto l’ordine di estrazione dei colori; ad esempio nel quesito e non si sa se le rosse siano le prime due estratte, quindi è sbagliato calcolare $P(R_1 \cap R_2 \cap N_3) = P(R_1) \cdot P(R_2 \mid R_1) \cdot P(N_3 \mid R_1 \cap R_2)$.

Esempio 55

Si lancia un dado; sia A l'evento “esce un numero pari” e B l'evento “esce un numero maggiore di 3” . Verificare se A e B sono indipendenti.

Si ha

$A = \{2,4,6\}$ $B = \{4,5,6\}$ $A \cap B = \{4,6\}$

$$P(A) = P(B) = \frac{3}{6} = \frac{1}{2}$$

$$P(A) \cdot P(B) = \frac{1}{4}$$

$$P(A \cap B) = \frac{2}{6} = \frac{1}{3}$$

Dunque gli eventi non sono indipendenti, essendo

$$P(A \cap B) \neq P(A) \cdot P(B)$$

In altre parole, sapere che il numero uscito è maggiore di 3 non lascia inalterata la valutazione della probabilità che il numero uscito sia pari; infatti

$$P(A) = \frac{1}{2} \qquad P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} \ .$$

Esempio 56

Data la tabella

	$P(A)$	$P(B)$	$P(A \cup B)$
caso 1	0.1	0.9	0.91
caso 2	0.4	0.6	0.76
caso 3	0.5	0.3	0.73

Tabella 3

esaminare in quali casi gli eventi sono indipendenti.

Ricordando che (teorema 8)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

si ottiene

	$P(A\cap B)$	$P(A)\cdot P(B)$	indipendenza
caso 1	0.09	0.09	sì
caso 2	0.24	0.24	sì
caso 3	0.07	0.15	no

Tabella 4

Esempio 57
Si effettuano due lanci di un dado. Sia
Evento A = "primo lancio pari"
Evento B = "secondo lancio ≤ 2".
Stabilire se gli eventi A e B sono indipendenti.

Lo spazio campione S ha 36 elementi, che sono le seguenti coppie

$S = \{(1,1),(1,2),...(1,6),(2,1),...(5,6),(6,6)\}$.

$A = \{2,4,6\}$ $B = \{1,2\}$

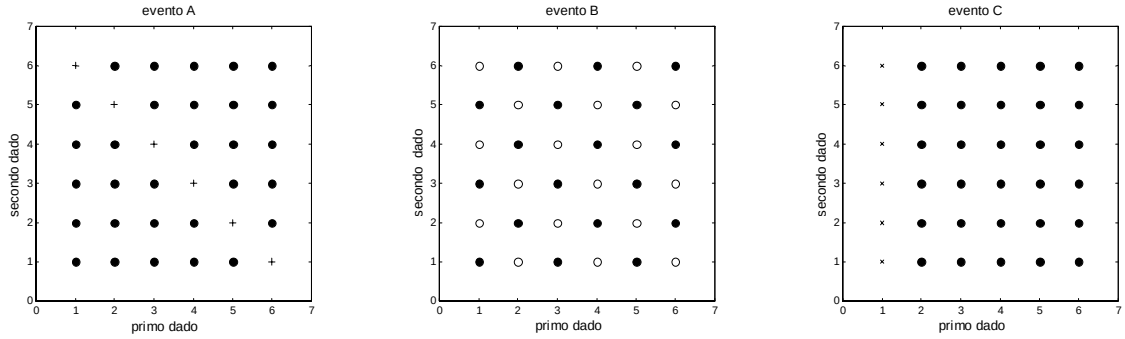
A e B sono indipendenti: infatti

$$P(A) = \frac{3}{6} \qquad P(B) = \frac{2}{6}$$

$A\cap B = \{(2,1),(2,2),(4,1),(4,2),(6,1),(6,2)\}$

$$P(A\cap B) = \frac{6}{36} = \frac{1}{6} = P(A)\cdot P(B)$$

Esempio 58
Si effettua il lancio di due dadi. Sia
Evento A = "somma uguale a 7"
Evento B = "somma dispari"
Evento C = "1 sul primo dado"
Verificare se sono indipendenti le coppie di eventi
a – A e B
b – A e C
c – B e C



\+ = somma 7 o = somma dispari × = 1 sul primo dado

Figura 6

Casi possibili: 36
Casi favorevoli per l'evento A: 6. Casi favorevoli per l'evento B: 18.
Casi favorevoli per l'evento C: 6.

$$P(A) = \frac{6}{36} = \frac{1}{6} \qquad P(B) = \frac{18}{36} = \frac{1}{2} \qquad P(C) = \frac{6}{36} = \frac{1}{6}$$

a – $P(A\cap B) = \frac{1}{6}$ $P(A)\cdot P(B) = \frac{1}{6}\cdot\frac{1}{2} = \frac{1}{12}$

$P(A\cap B) \neq P(A)\cdot P(B) \Rightarrow$ A e B non sono indipendenti

b – $P(A\cap C) = \frac{1}{36}$ $P(A)\cdot P(C) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$

$P(A\cap C) = P(A)\cdot P(C) \Rightarrow$ A e B sono indipendenti

c – $P(B\cap C) = \frac{3}{36} = \frac{1}{12}$ $P(B)\cdot P(C) = \frac{1}{2}\cdot\frac{1}{6} = \frac{1}{12}$

$P(B\cap C) = P(B)\cdot P(C) \Rightarrow$ B e C sono indipendenti

Esempio 59
Un dado è lanciato quattro volte. Calcolare la probabilità di ottenere almeno un 6 in quattro lanci.

Evento A = "almeno un 6 in 4 lanci"
Evento $\overline{A}$ = "nessun 6 in quattro lanci".
La probabilità di non ottenere 6 in un singolo lancio è $\frac{5}{6}$, quindi la probabilità di non ottenere nessun 6 in quattro lanci (eventi indipendenti) è

$$P(\overline{A}) = \left(\frac{5}{6}\right)^4 .$$

Pertanto

$$P(A) = 1 - P(\overline{A}) = 1 - \left(\frac{5}{6}\right)^4 \cong 0.518 .$$

Si osservi che **eventi mutuamente esclusivi,** (ossia **disgiunti**), **non sono indipendenti.**
Infatti per ogni coppia di eventi disgiunti A e B si ha $A\cap B = \varnothing$; se A e B fossero indipendenti dovrebbe essere

$$P(A\cap B) = P(\varnothing) = 0 = P(A)\cdot P(B)$$

quindi almeno uno dei due eventi dovrebbe avere probabilità 0, cioè essere impossibile.
In realtà due eventi disgiunti sono fortemente dipendenti, perché disgiunti significa che se uno si realizza, allora l'altro non si può realizzare.

2.6 Il teorema di Bayes

Consideriamo la situazione illustrata con il seguente diagramma di Venn

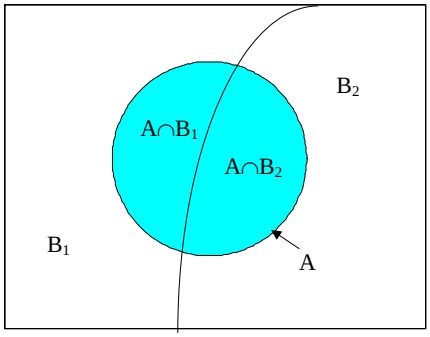


Figura 7

Gli eventi B_1 e B_2 sono tali che

$$B_1 \cap B_2 = \varnothing \qquad \text{e} \qquad B_1 \cup B_2 = S$$

dove S è lo spazio campione. Gli insiemi $A \cap B_1$ e $A \cap B_2$ sono mutuamente esclusivi, perciò

$$P(A) = P(A \cap B_1) + P(A \cap B_2).$$

Applicando la regola di moltiplicazione (2.14) si ottiene

$$P(A) = P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2).$$

Questa formula esprime la regola della probabilità totale nel caso particolare di due eventi B_1 e B_2. La regola può essere generalizzata al caso di una famiglia di *n* eventi $B_1, B_2, ..., B_n$ mutuamente esclusivi ed esaustivi[2].
Si può dimostrare il seguente teorema.

Teorema 12 – Teorema della probabilità totale
Sia A un evento e $\{B_1, B_2, ..., B_n\}$ una famiglia di eventi dello spazio campione S mutuamente esclusivi e tali che uno e uno solo di essi si verifichi, ossia tali che

$B_i \cap B_j = \varnothing$ per $i \neq j$ (mutuamente esclusivi)

$B_1 \cup B_2 \cup ... \cup B_n = S$ (esaustivi)

$P(B_i) \neq 0$ per ogni i

Allora si dimostra che

$$P(A) = P(A \mid B_1) \cdot P(B_1) + P(A \mid B_2) \cdot P(B_2) + ... + P(A \mid B_n) \cdot P(B_n) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i) \qquad (2.16)$$

Per dimostrare questo risultato è sufficiente osservare che se A si verifica, esso deve verificarsi insieme ad uno e uno solo degli eventi $B_1, B_2, ..., B_n$, perciò

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_n).$$

Applicando il teorema 10 si ha

$$P(A \cap B_i) = P(B_i) \cdot P(A \mid B_i)$$

Sostituendo questa relazione nella precedente si ottiene la tesi.

L'utilità del teorema sta nel fatto che talvolta $P(A)$ è difficile da calcolare direttamente, mentre è più facile calcolare le probabilità $P(A \mid B_i)$ e poi ricostruire $P(A)$ dalla formula (2.16).

Esempio 60
Siano date due urne che contengono rispettivamente
urna I 2 palline rosse e 1 nera
urna II 3 palline rosse e 2 nere.
Scegliamo a caso un'urna ed estraiamo a caso una pallina dall'urna scelta. Qual è la probabilità di estrarre una pallina nera?

Evento B_1 = "è stata scelta l'urna I"
Evento B_2 = "è stata scelta l'urna II"

$$B_1 \cap B_2 = \varnothing \qquad B_1 \cup B_2 = S$$

Evento A = "è stata estratta una pallina nera "
Applicando il teorema della probabilità totale si ha

$$P(A) = P(A \mid B_1) \cdot P(B_1) + P(A \mid B_2) \cdot P(B_2)$$

Si ha

$$P(B_1) = \frac{1}{2} \qquad P(B_2) = \frac{1}{2}$$

$$P(A \mid B_1) = \frac{1}{3} \qquad P(A \mid B_2) = \frac{2}{5}$$

[2] Gli eventi $B_1, B_2, ..., B_n$ si dicono **esaustivi**, se la loro unione è tutto lo spazio campione.

quindi

$$P(\mathrm{A}) = \frac{1}{3}\cdot\frac{1}{2} + \frac{2}{5}\cdot\frac{1}{2} = \frac{11}{30} \cong 0.367\,.$$

Si osservi che la probabilità è diversa da quella che si avrebbe se tutte le palline fossero contenute in un'unica urna: in questo caso la probabilità di estrarre una pallina nera sarebbe

$$P(\mathrm{A}) = \frac{3}{8} = 0.375\,.$$

La differenza fra i due risultati dipende dal fatto che le due urne contengono un numero diverso di palline, quindi una pallina dell'urna I non ha la stessa probabilità di essere estratta di una pallina dell'urna II.

Esempio 61

Riferendoci all'esempio 60 possiamo ora porre il seguente quesito: se è stata estratta una pallina nera, qual è la probabilità di aver scelto l'urna I?

Per rispondere a questa domanda bisogna calcolare la probabilità $P(\mathrm{B}_1 \mid \mathrm{A})$.

Dal teorema 10 si ricava la relazione

$$P(\mathrm{B}_1 \mid \mathrm{A})\cdot P(\mathrm{A}) = P(\mathrm{A} \mid \mathrm{B}_1)\cdot P(\mathrm{B}_1)$$

da cui segue

$$P(\mathrm{B}_1 \mid \mathrm{A}) = \frac{P(\mathrm{A} \mid \mathrm{B}_1)\cdot P(\mathrm{B}_1)}{P(\mathrm{A})} = \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{11}{30}} = \frac{5}{11} \cong 0.455\,.$$

Generalizzando il procedimento seguito nell'esempio 61 si può ottenere il seguente importante risultato.

Teorema 13 – Teorema di Bayes

Sia A un evento con $P(\mathrm{A}) > 0$ e $\{\mathrm{B}_1, \mathrm{B}_2, \ldots\ldots, \mathrm{B}_n\}$ una famiglia di eventi dello spazio campione S soddisfacenti le ipotesi del teorema precedente.

Allora

$$P(\mathrm{B}_k \mid \mathrm{A}) = \frac{P(\mathrm{A} \mid \mathrm{B}_k)\cdot P(\mathrm{B}_k)}{\sum_{i=1}^{n} P(\mathrm{A} \mid \mathrm{B}_i)\cdot P(\mathrm{B}_i)} \qquad \text{per ogni } k \qquad (2.17)$$

Questo teorema ci permette di trovare le probabilità degli eventi B_k che possono essere la causa del verificarsi dell'evento A, in altre parole che l'effetto A sia stato provocato dalla causa B_k; per questo motivo è detto anche **teorema della probabilità delle cause**.

Esempio 62

Siano date due urne contenenti delle palline bianche e nere; nell'urna I il 70% delle palline sono nere; nell'urna II il 40% delle palline sono nere.

La probabilità di scegliere l'urna I sia 0.1; la probabilità di scegliere l'urna II sia invece 0.9. Calcolare la probabilità che una pallina nera estratta a caso provenga dall'urna I.

Evento A = "pallina estratta nera";

Evento B_1 = "la pallina proviene dall'urna I";

Evento B_2 = "la pallina proviene dall'urna II".

$P(\mathrm{B}_1) = 0.1$ $\qquad$ $P(\mathrm{B}_2) = 0.9$

$P(\mathrm{A} \mid \mathrm{B}_1) = 0.7$ $\qquad$ $P(\mathrm{A} \mid \mathrm{B}_2) = 0.4$

Dal teorema di Bayes segue

$$P(\text{dall'urna I} \mid \text{nera}) = P(B_1 \mid A) =$$
$$= \frac{P(B_1) \cdot P(A \mid B_1)}{P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2)} = \frac{0.1 \cdot 0.7}{0.1 \cdot 0.7 + 0.9 \cdot 0.4} = 0.163 = 16.3\%$$

Il risultato può essere interpretato come segue: effettuando numerose prove, nel 16.3% dei casi in cui si è estratta una pallina nera, essa proviene dall'urna I.

Esempio 63

Un problema di collaudo in un processo produttivo. Un'industria ha installato un sistema automatico per il controllo di qualità, che garantisce che, se un pezzo è difettoso, viene eliminato con probabilità 0.995. C'è una probabilità pari a 0.001 che anche un pezzo non difettoso venga eliminato. Si sa anche che la probabilità che un pezzo sia difettoso è 0.2.
Calcoliamo la probabilità che un pezzo che non sia stato eliminato al controllo di qualità sia difettoso.

Evento E = "il pezzo viene eliminato"
Evento D = "il pezzo è difettoso"
Sappiamo che

$$P(E \mid D) = 0.995 \qquad P(E \mid \overline{D}) = 0.001 \qquad P(D) = 0.2$$

Con il teorema di Bayes vogliamo calcolare

$$P(D \mid \overline{E}) = \frac{P(\overline{E} \mid D) \cdot P(D)}{P(\overline{E} \mid D) \cdot P(D) + P(\overline{E} \mid \overline{D}) \cdot P(\overline{D})}$$

Abbiamo

$$P(\overline{E} \mid D) = 1 - P(E \mid D) = 1 - 0.995 = 0.005$$
$$P(\overline{E} \mid \overline{D}) = 1 - P(E \mid \overline{D}) = 1 - 0.001 = 0.999$$
$$P(\overline{D}) = 1 - P(D) = 1 - 0.2 = 0.8$$

Calcoliamo perciò

$$P(D \mid \overline{E}) = \frac{P(\overline{E} \mid D) \cdot P(D)}{P(\overline{E} \mid D) \cdot P(D) + P(\overline{E} \mid \overline{D}) \cdot P(\overline{D})} =$$
$$= \frac{0.005 \cdot 0.2}{0.005 \cdot 0.2 + 0.999 \cdot 0.8} \cong 0.00125 = 0.125\%$$

Esempio 64

Un problema di marketing. Il responsabile marketing di una società che produce giocattoli sta analizzando le probabilità di successo sul mercato di un nuovo gioco. Nell'esperienza passata della ditta il 65% dei nuovi giocattoli ha avuto successo di mercato, mentre il restante 35% non l'ha ottenuto. Si sa inoltre che l'80% dei giocattoli di successo avevano ricevuto un giudizio positivo da parte degli esperti di marketing della società prima dell'immissione del prodotto sul mercato, mentre lo stesso giudizio era stato attribuito solo al 30% dei giocattoli che si sarebbero poi rivelati un insuccesso di mercato.
Il responsabile è interessato a calcolare la probabilità che il nuovo giocattolo sia premiato dal mercato, sapendo che gli esperti della società lo hanno valutato positivamente.

Evento S = "giocattolo di successo"
Evento $\overline{S}$ = "giocattolo non di successo"
Evento Pos = "giudizio positivo degli esperti di marketing"
Evento Neg = " giudizio negativo degli esperti di marketing ".

Sappiamo che

$$P(\mathrm{S}) = 0.65 \qquad P(\mathrm{Pos} \mid \mathrm{S}) = 0.80$$
$$P(\bar{\mathrm{S}}) = 0.35 \qquad P(\mathrm{Pos} \mid \bar{\mathrm{S}}) = 0.30$$

Con il teorema di Bayes calcoliamo

$$P(\mathrm{S} \mid \mathrm{Pos}) = \frac{P(\mathrm{Pos} \mid \mathrm{S}) \cdot P(\mathrm{S})}{P(\mathrm{Pos} \mid \mathrm{S}) \cdot P(\mathrm{S}) + P(\mathrm{Pos} \mid \bar{\mathrm{S}}) \cdot P(\bar{\mathrm{S}})} =$$
$$= \frac{0.80 \cdot 0.65}{0.80 \cdot 0.65 + 0.30 \cdot 0.35} = 0.832 = 83.2\%$$

La probabilità dell'evento complementare, ossia che il giocattolo valutato positivamente dagli esperti della società non abbia poi successo di mercato, vale

$$P(\bar{\mathrm{S}} \mid \mathrm{Pos}) = 1 - P(\mathrm{S} \mid \mathrm{Pos}) = 1 - 0.832 = 0.168 = 16.8\% .$$

Esempio 65

Quattro tecnici si occupano delle riparazioni dei guasti che accadono in una linea automatica di produzione.

Il primo tecnico effettua il 20% delle riparazioni e in un caso su 20 non esegue correttamente il lavoro; il secondo tecnico effettua il 60% delle riparazioni e in un caso su 10 non esegue correttamente il lavoro; il terzo tecnico effettua il 15% delle riparazioni e in un caso su 10 non esegue correttamente il lavoro; il quarto tecnico effettua il 5% delle riparazioni e in un caso su 20 non esegue correttamente il lavoro.

Il successivo guasto viene ritenuto una conseguenza della precedente riparazione imperfetta; qual è la probabilità che la precedente riparazione sia stata fatta dal primo tecnico?

Evento B_1 = "riparazione eseguita dal 1° tecnico" $\qquad P(\mathrm{B}_1) = 0.20 \qquad P(\mathrm{A} \mid \mathrm{B}_1) = 0.05$

Evento B_2 = "riparazione eseguita dal 2° tecnico" $\qquad P(\mathrm{B}_2) = 0.60 \qquad P(\mathrm{A} \mid \mathrm{B}_2) = 0.10$

Evento B_3 = "riparazione eseguita dal 3° tecnico" $\qquad P(\mathrm{B}_3) = 0.15 \qquad P(\mathrm{A} \mid \mathrm{B}_3) = 0.10$

Evento B_4 = "riparazione eseguita dal 4° tecnico" $\qquad P(\mathrm{B}_4) = 0.05 \qquad P(\mathrm{A} \mid \mathrm{B}_4) = 0.05$

Applicando il teorema di Bayes si trova

$$P(\mathrm{B}_1 \mid \mathrm{A}) = \frac{(0.20)(0.05)}{(0.20)(0.05) + (0.60)(0.10) + (0.15)(0.10) + (0.05)(0.05)} = 0.114 .$$

E' interessante notare che, sebbene il primo tecnico svolga un lavoro imperfetto solo nel 5% dei casi, tuttavia più dell'11% delle riparazioni non perfette sono una sua responsabilità.

Esempio 66

Per produrre uno stesso tipo di prodotto sono impiegate tre diverse macchine, M_1, M_2, M_3, che producono pezzi difettosi con le rispettive probabilità: 1%, 2% e 0.1%.

Le tre macchine producono rispettivamente il 30%, il 50% e il 20% della produzione totale.

a – Qual è la probabilità che un pezzo uscito dalla fabbrica sia difettoso?

b – Qual è la probabilità che un pezzo difettoso sia stato prodotto dalla macchina M_2?

Evento D = "pezzo difettoso".

Si hanno le seguenti probabilità

$$P(\mathrm{M}_1) = 30\% = 0.3 \qquad P(\mathrm{D} \mid \mathrm{M}_1) = 1\% = 0.01$$
$$P(\mathrm{M}_2) = 50\% = 0.5 \qquad P(\mathrm{D} \mid \mathrm{M}_2) = 2\% = 0.02$$
$$P(\mathrm{M}_3) = 20\% = 0.2 \qquad P(\mathrm{D} \mid \mathrm{M}_3) = 0.1\% = 0.001$$

a – Applicando il teorema della probabilità totale si trova la probabilità che un pezzo sia difettoso, non importa da quale macchina sia stato prodotto

$$P(\mathrm{D}) = P(\mathrm{D}\mid \mathrm{M}_1)\cdot P(\mathrm{M}_1) + P(\mathrm{D}\mid \mathrm{M}_2)\cdot P(\mathrm{M}_2) + P(\mathrm{D}\mid \mathrm{M}_3)\cdot P(\mathrm{M}_3) =$$
$$= 0.01\cdot 0.3 + 0.02\cdot 0.5 + 0.001\cdot 0.2 = 0.0132 = 1.32\%$$

b – Applicando il teorema di Bayes si trova la probabilità che il pezzo difettoso sia stato prodotto dalla macchina M_2

$$P(\mathrm{M}_2 \mid D) = \frac{P(\mathrm{D}\mid \mathrm{M}_2)\cdot P(\mathrm{M}_2)}{P(\mathrm{D}\mid \mathrm{M}_1)\cdot P(\mathrm{M}_1) + P(\mathrm{D}\mid \mathrm{M}_2)\cdot P(\mathrm{M}_2) + P(\mathrm{D}\mid \mathrm{M}_3)\cdot P(\mathrm{M}_3)} =$$
$$= \frac{0.02\cdot 0.5}{0.0132} \cong 0.76 = 76\%$$

Quindi in circa $\frac{3}{4}$ dei casi si può ritenere che la causa di un pezzo difettoso sia la macchina M_2.

Nel caso in cui gli eventi della famiglia $\{B_1, B_2, ...,B_n\}$ hanno la stessa probabilità $P(\mathrm{B}_i) = \frac{1}{n}$, la formula del teorema di Bayes si semplifica e diventa

$$P(\mathrm{B}_k \mid \mathrm{A}) = \frac{P(\mathrm{A}\mid \mathrm{B}_k)}{\sum_{i=1}^{n} (P\mathrm{A}\mid \mathrm{B}_i)} \qquad \text{per ogni } k \tag{2.18}$$

Esempio 67
Quattro tiratori di una stessa squadra vengono classificati in base alle probabilità di fare centro con un tiro; al tiratore T_1 viene attribuita una probabilità dell'80%, al tiratore T_2 una probabilità del 50%, al tiratore T_3 una probabilità del 20% e al tiratore T_4 una probabilità del 10%. I quattro tiratori sparano contemporaneamente un colpo ciascuno e solo uno ha fatto centro: qual è la probabilità che il centro sia stato colpito da T_1?

Evento T_i = "centro colpito da T_i " $\qquad P(\mathrm{T}_i) = \frac{1}{4}$.

Evento C = "il tiratore ha fatto centro".

Applicando la formula di Bayes nella forma semplificata (2.18) si ha

$$P(\mathrm{T}_1 \mid \mathrm{C}) = \frac{P(\mathrm{C}\mid \mathrm{T}_1)}{P(\mathrm{C}\mid \mathrm{T}_1) + P(\mathrm{C}\mid \mathrm{T}_2) + P(\mathrm{C}\mid \mathrm{T}_3) + P(\mathrm{C}\mid \mathrm{T}_4)} = \frac{0.8}{0.8 + 0.5 + 0.2 + 0.1} = 0.5 = 50\%$$

Applicazione del teorema di Bayes a un problema di diagnosi medica.
Il teorema di Bayes trova un'importante applicazione in ambito sanitario.
In un test clinico, un individuo viene sottoposto ad un certo esame di laboratorio, per stabilire se ha o non ha una data malattia. Il test può avere esito positivo (il che indica la presenza della malattia) o negativo (il che indica che l'individuo è sano). C'è però sempre una possibilità di errore: può darsi che qualcuno degli individui risultati positivi siano in realtà sani ("**falsi positivi**"), e che qualcuno degli individui risultati negativi siano in realtà malati ("**falsi negativi**").
Prima di applicare il test nei laboratori su larga scala, è quindi opportuno valutarne la bontà. Per far questo si possono sottoporre al test un campione di persone di cui sappiamo già se sono sane o malate, e vedere se la risposta del test è corretta.
Gli eventi a cui siamo interessati sono
Evento M = "l'individuo è malato"
Evento S = "l'individuo è sano"
Evento Pos = "il test è positivo"
Evento Neg = "il test è negativo".

Utilizzando la nozione di probabilità condizionata si danno le seguenti definizioni.

Definizione 7
La probabilità condizionata $P(\mathbf{Pos}|\mathbf{M})$ viene detta **sensibilità** del test.

Definizione 8
La probabilità condizionata $P(\mathbf{Neg}|\mathbf{S})$ viene detta **specificità** del test.

Il test è tanto più sensibile quanto più è probabile che un malato risulti positivo, ed è tanto più specifico quanto più è probabile che un sano risulti negativo, ovvero che solo i malati risultino positivi. Pertanto un buon test è un test con sensibilità e specificità molto vicine a 1.
Supponiamo ora che il test venga effettivamente applicato per scoprire se una persona è malata o meno. Calcoliamo la probabilità che un individuo che risulta positivo al test sia effettivamente malato. Questa è una probabilità condizionata e si definisce nel modo seguente.

Definizione 9
La probabilità che un individuo che risulta positivo al test sia effettivamente malato $P(\mathbf{M}\,|\,\mathbf{Pos})$ viene detta **valore predittivo** del test.

Per il teorema di Bayes il valore predittivo del test è

$$P(\mathrm{M}\,|\,\mathrm{Pos}) = \frac{P(\mathrm{Pos}\,|\,\mathrm{M})\cdot P(\mathrm{M})}{P(\mathrm{Pos}\,|\,\mathrm{M})\cdot P(\mathrm{M}) + P(\mathrm{Pos}\,|\,\mathrm{S})\cdot P(\mathrm{S})}$$

Si può quindi notare che per calcolare il valore predittivo del test non basta conoscerne la sensibilità e la specificità, ma occorre conoscere anche la probabilità $P(\mathrm{M})$ con cui la malattia colpisce la popolazione complessiva.

Esempio 68
Supponiamo che la probabilità che una persona abbia una certa malattia sia uguale a 0.03. La diagnosi della malattia viene fatta con un test che ha le seguenti caratteristiche: applicato a un individuo affetto dalla malattia dà risultato positivo con probabilità pari a 0.9; applicato a un individuo sano dà esito positivo con probabilità pari a 0.02.
Supponiamo che su un individuo il test abbia dato risultato positivo: qual è la probabilità che sia effettivamente malato?

Con le notazioni sopra suggerite si ha

$P(\mathrm{M}) = 0.03$ $\qquad$ $P(\mathrm{S}) = 1 - P(\mathrm{M}) = 0.97$
$P(\mathrm{Pos}\,|\,\mathrm{M}) = 0.9$ (sensibilità)
$P(\mathrm{Pos}\,|\,\mathrm{S}) = 0.02$

La probabilità che l'individuo sia malato, sapendo che il test è positivo, è il valore predittivo e si calcola con il teorema di Bayes

$$P(\mathrm{M}\,|\,\mathrm{Pos}) = \frac{P(\mathrm{Pos}\,|\,\mathrm{M})\cdot P(\mathrm{M})}{P(\mathrm{Pos}\,|\,\mathrm{M})\cdot P(\mathrm{M}) + P(\mathrm{Pos}\,|\,\mathrm{S})\cdot P(\mathrm{S})} = \frac{0.9\cdot 0.03}{0.9\cdot 0.03 + 0.02\cdot 0.97} = 0.582$$

In base a questo risultato possiamo dire che solo il 58% circa di coloro che risultano positivi al test è effettivamente malato, il restante 42% sono falsi positivi.
Osserviamo che la probabilità che una persona sia malata, sapendo che è risultata positiva al test, è comunque maggiore della probabilità che aveva prima di sottoporsi al test.
La probabilità che il test dia esito positivo si calcola con il teorema della probabilità totale, ed è uguale al denominatore della frazione nel teorema di Bayes

$$P(\mathrm{Pos}) = P(\mathrm{Pos}\,|\,\mathrm{M})\cdot P(\mathrm{M}) + P(\mathrm{Pos}\,|\,\mathrm{S})\cdot P(\mathrm{S}) = 0.9\cdot 0.03 + 0.02\cdot 0.97 = 0.0464$$

Supponiamo ora che il test abbia dato risultato negativo: qual è la probabilità che l'individuo sia sano?

Anche questa probabilità si calcola con il teorema di Bayes

$$P(\mathrm{S} \mid \mathrm{Neg}) = \frac{P(\mathrm{Neg} \mid \mathrm{S}) \cdot P(\mathrm{S})}{P(\mathrm{Neg} \mid \mathrm{S}) \cdot P(\mathrm{S}) + P(\mathrm{Neg} \mid \mathrm{M}) \cdot P(\mathrm{M})}$$

Osserviamo che

$$P(\mathrm{Neg} \mid \mathrm{S}) = 1 - P(\mathrm{Pos} \mid \mathrm{S}) = 1 - 0.02 = 0.98 \quad \text{(specificità)}$$
$$P(\mathrm{Neg} \mid \mathrm{M}) = 1 - P(\mathrm{Pos} \mid \mathrm{M}) = 1 - 0.9 = 0.1$$

Pertanto

$$P(\mathrm{S} \mid \mathrm{Neg}) = \frac{0.98 \cdot 0.97}{0.98 \cdot 0.97 + 0.1 \cdot 0.03} = 0.997$$

In conclusione, se il test è risultato negativo, abbiamo una probabilità molto alta che la persona sia sana, quindi il test è altamente predittivo negativamente, mentre non è molto predittivo in senso positivo (solo il 58% circa). In altre parole i falsi negativi sono pochissimi, mentre i falsi positivi sono piuttosto numerosi (il 42%).

Esempio 69

Caso di una malattia rara.

La sensibilità del test per una data malattia rara (ad esempio l'HIV) sia circa uguale a 0.993: la specificità del test sia circa 0.9999. La probabilità di contrarre la malattia nella popolazione sia circa 0.000025.

$$P(\mathrm{Pos} \mid \mathrm{M}) = 0.993 \qquad \text{(sensibilità)}$$
$$P(\mathrm{Neg} \mid \mathrm{S}) = 0.9999 \qquad \text{(specificità)}$$
$$P(\mathrm{M}) = 0.000025$$

La probabilità che una persona risultata positiva a questo test sia effettivamente malata è, con il teorema di Bayes

$$P(\mathrm{M} \mid \mathrm{Pos}) = \frac{P(\mathrm{Pos} \mid \mathrm{M}) \cdot P(\mathrm{M})}{P(\mathrm{Pos} \mid \mathrm{M}) \cdot P(\mathrm{M}) + P(\mathrm{Pos} \mid \mathrm{S}) \cdot P(\mathrm{S})} =$$
$$= \frac{0.993 \cdot 0.000025}{0.993 \cdot 0.000025 + (1 - 0.9999) \cdot (1 - 0.000025)} = 0.19888 \cong 20\%$$

Questo significa che solo il 20% circa di coloro che risultano positivi al test sono effettivamente malati; in altre parole l'80% sono "falsi positivi". Il risultato, apparentemente sorprendente, dipende dal fatto che la malattia che si cerca è molto rara sulla popolazione complessiva.

Si osservi che si sta supponendo di sottoporre al test persone di cui a priori non si sa nulla; se si applicasse il test a persone scelte non casualmente, ma in qualche "categoria a rischio" (ad esempio per l'HIV fra i tossicodipendenti), la probabilità P(M) andrebbe sostituita con la probabilità della malattia in quella classe di persone, e sarebbe più elevata; risulterebbe più elevato di conseguenza il valore predittivo del test.

Si noti ancora che la probabilità che una persona sia malata, sapendo che è risultata positiva al test, è comunque molto maggiore della probabilità che aveva prima di sottoporsi al test

$$\frac{P(\mathrm{M} \mid \mathrm{Pos})}{P(\mathrm{M})} = \frac{0.19888}{0.000025} \cong 7955 \ .$$

(la probabilità è cresciuta di circa 8000 volte).

Se calcoliamo la probabilità che una persona risultata negativa al test sia sana, otteniamo

$$P(\mathrm{S} \mid \mathrm{Neg}) = \frac{P(\mathrm{Neg} \mid \mathrm{S}) \cdot P(\mathrm{S})}{P(\mathrm{Neg} \mid \mathrm{S}) \cdot P(\mathrm{s}) + P(\mathrm{Neg} \mid \mathrm{M}) \cdot P(\mathrm{M})} =$$
$$= \frac{0.9999 \cdot (1 - 0.000025)}{0.9999 \cdot (1 - 0.000025) + (1 - 0.993) \cdot 0.000025} = 0.9999998$$

Il numero dei falsi negativi è quindi molto basso.

3. Variabili aleatorie e distribuzioni di probabilità

3.1 Variabili aleatorie

Una variabile aleatoria è una variabile che può assumere valori diversi in dipendenza da qualche fenomeno casuale; la sua definizione rigorosa è la seguente.

Definizione 1
Una **variabile aleatoria** (o **casuale**) è una funzione reale X definita sullo spazio campione S e a valori reali

$$X : S \to \mathbf{R}$$

Essa associa ad ogni possibile risultato di un esperimento, cioè ad ogni elemento dello spazio campione S, un numero reale.

In alcuni casi gli eventi elementari sono già numeri reali, ad esempio i numeri da 1 a 6 nel lancio di un dado, e allora sono essi stessi valori di una variabile aleatoria. In altri casi è necessaria un'opportuna codifica.

Esempio 1
Si effettua il lancio di una moneta. Lo spazio campione è

$$S = \{T,C\}$$

Ponendo

$$X(C) = m \qquad X(T) = n \qquad m, n \in \mathbf{R} \qquad m \neq n$$

si definisce una variabile aleatoria X.

Esempio 2
Si effettuano due lanci di una moneta. Lo spazio campione è

$$S = \{TT,CC,TC,CT\}$$

Ad ogni elemento dello spazio campione possiamo associare un numero reale che rappresenta il numero delle volte che esce T, secondo la seguente tabella

Elementi di S	TT	TC	CT	CC
X	2	1	1	0

Tabella 1

ossia

$$X(TT) = 2 \qquad X(TC) = 1 \qquad X(CT) = 1 \qquad X(CC) = 0$$

X è una variabile aleatoria.
Si osservi che si possono definire altre variabili aleatorie su questo spazio campione: ad esempio il quadrato del numero delle teste, anziché il numero delle teste, o il numero delle teste meno il numero delle croci.

Definizione 2
Una variabile aleatoria che può assumere solo un numero finito di valori o un'infinità numerabile[1] di valori è detta **variabile aleatoria discreta**, mentre una variabile aleatoria che assume un'infinità non numerabile di valori è detta **continua**.

Le variabili aleatorie definite negli esempi 1 e 2 sono variabili aleatorie discrete.
Di solito quello che interessa di una variabile aleatoria è calcolare la probabilità che essa assuma certi valori; nel caso dei due lanci di una moneta ci potrebbe ad esempio interessare la probabilità che la variabile aleatoria assuma il valore 1 oppure che assuma un valore minore o uguale a 1.

[1] Vedere nota pag. 4.

Osservazione
In generale, se X è una variabile aleatoria, si usano notazioni del tipo seguente
Evento "X assume il valore a" $X = a$
Evento "X assume valori compresi nell'intervallo (a , b)" $a < X < b$
Evento "X assume valori minori o uguali a c" $X \le c$.
Indichiamo con $P(X = a)$, $P(a < X < b)$, $P(X \le c)$ le probabilità dei precedenti eventi.
Per il teorema 9, pag. 72, si ha

$$P(X > c) = 1 - P(X \le c) \qquad c \in \mathbf{R}$$

dove $P(X > c)$ indica la probabilità che X assuma un valore maggiore di c.

Esempio 3
Si consideri la variabile aleatoria discreta X, definita come il numero di teste T in due lanci di una moneta; si ha ad esempio

$$P(X = 2) = \frac{1}{4} \qquad P(X = 1) = \frac{2}{4} = \frac{1}{2}$$

$$P(1 < X < 2) = 0 \qquad P(1 < X \le 2) = \frac{1}{4} \qquad P(0 \le X \le 2) = 1$$

Esempio 4
Si consideri la variabile aleatoria discreta X, definita come il numero ottenuto nel lancio di un dado; si ha ad esempio

$$P(5 < X < 6) = 0 \qquad P(5 \le X < 6) = \frac{1}{6} \qquad P(1 \le X \le 6) = 1$$

$$P(X > 2) = 1 - P(X \le 2) = 1 - \frac{2}{6} = \frac{2}{3}$$

3.2 Distribuzioni di probabilità discrete

Sia X una variabile aleatoria discreta e siano $x_1, x_2, \ldots$ i valori che essa può assumere; si supponga per semplicità che la variabile aleatoria assuma un numero finito di valori[2] e inoltre che questi valori siano assunti con probabilità

$$P(X = x_i) \qquad i = 1,2,\ldots$$

Definizione 3
La funzione

$$f(x_i) = P(X = x_i) \qquad i = 1,2,\ldots \tag{3.1}$$

che ad ogni valore assunto dalla variabile aleatoria discreta X associa la corrispondente probabilità è detta **distribuzione di probabilità** della variabile aleatoria X.

La rappresentazione grafica di $f(x)$ può essere fatta con un **diagramma a barre** o con un **istogramma**.

Definizione 4
Si definisce **funzione di distribuzione** o **funzione di ripartizione** di una variabile aleatoria X la funzione

$$F(x) = P(X \le x) \qquad x \in \mathbf{R} \tag{3.2}$$

La funzione F associa ad ogni valore reale x la probabilità che la variabile aleatoria X assuma un valore minore o uguale a x. Essa è definita su $\mathbf{R}$, monotona crescente da 0 a 1; il suo grafico è una funzione a gradino.

[2] Si può facilmente generalizzare al caso di un numero non finito di valori.

Esempio 5
Si effettuano due lanci consecutivi di una moneta. La variabile aleatoria X è il numero di volte che esce T ed è descritta dalla tabella 1 (esempio 2).
Si ha

$$P(\text{TT}) = \frac{1}{4} \qquad P(\text{TC}) = \frac{1}{4} \qquad P(\text{CT}) = \frac{1}{4} \qquad P(\text{CC}) = \frac{1}{4}$$

quindi

$$P(X=0) = P(\text{CC}) = \frac{1}{4}$$

$$P(X=1) = P(\text{TC} \cup \text{CT}) = P(\text{TC}) + P(\text{CT}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(X=2) = P(\text{TT}) = \frac{1}{4}$$

La distribuzione di probabilità è assegnata dalla tabella 2

x_i	0	1	2
$f(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Tabella 2

La funzione $f(x)$ può essere rappresentata con un diagramma a barre (figura 1), o con un istogramma (figura 2).

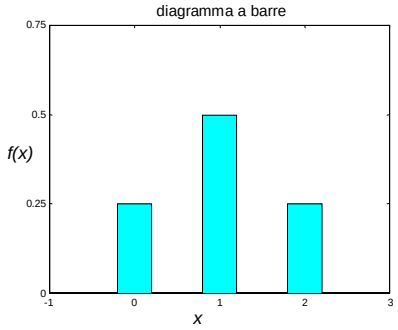


Figura 1

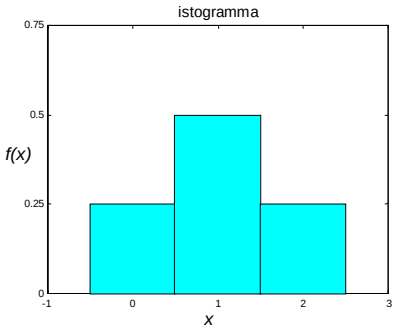


Figura 2

Nel grafico della figura 1 la somma delle ordinate è 1; nel grafico della figura 2 la somma delle aree dei tre rettangoli è 1.
Ricaviamo ora la funzione di distribuzione $F(x)$

x	$F(x) = P(X \le x)$
$x < 0$	0
$0 \le x < 1$	$\frac{1}{4}$
$1 \le x < 2$	$\frac{1}{4} + \frac{1}{2} = \frac{3}{4}$
$x \ge 2$	$\frac{3}{4} + \frac{1}{4} = 1$

Tabella 3

La funzione di distribuzione della variabile aleatoria X è quindi

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \le x < 1 \\ \frac{3}{4} & 1 \le x < 2 \\ 1 & x \ge 2 \end{cases}$$

$F(x)$ è una funzione a gradino con salto non costante; il grafico è il seguente

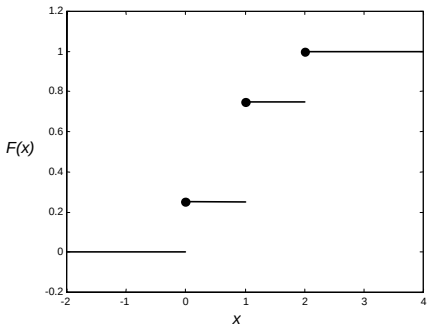


Figura 3

Nell'esempio precedente si può osservare che la funzione di distribuzione $F(x)$ è uguale alla somma delle probabilità

$$f(x_i) = P(X = x_i)$$

per tutti gli $x_i \le x$ .Questo risultato è vero per ogni variabile aleatoria discreta.
Per una variabile aleatoria discreta si ha quindi la seguente relazione tra funzione di distribuzione e distribuzione di probabilità

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i) \qquad (3.3)$$

In generale, nel caso di una variabile aleatoria discreta, una funzione $f(x)$ è una distribuzione di probabilità se

1) $f(x_i) \ge 0 \qquad \forall x_i \qquad (3.4)$

2) $\sum_{x_i} f(x_i) = 1 \qquad (3.5)$

dove la sommatoria è estesa a tutti i possibili valori x_i assunti dalla variabile aleatoria *X*.

Esempio 6
Sia data la funzione

$$f(x) = \frac{x+3}{15} \qquad x = 1,2,3 .$$

Verificare se $f(x)$ è una distribuzione di probabilità di una data variabile aleatoria discreta *X*.

Sostituendo $x_1 = 1,\ x_2 = 2,\ x_3 = 3$ si ottiene

$$f(1) = \frac{4}{15} \qquad f(2) = \frac{1}{3} \qquad f(3) = \frac{6}{15} .$$

Questi valori sono tutti compresi fra 0 e 1; inoltre la loro somma vale 1, perciò la funzione assegnata è una distribuzione di probabilità discreta.

Esempio 7
Trovare il valore della costante $k \in \mathbf{R}$ in modo che la funzione

$$f(x) = \begin{cases} \left(\frac{1}{2}\right)^x & x = 1,2,3,4,5 \\ k & x = 6 \end{cases}$$

sia una distribuzione di probabilità discreta.
Trovare la funzione di distribuzione.

Deve essere

1) $f(x_i) \geq 0 \qquad \forall x_i \qquad \Rightarrow \quad k \geq 0$

2) $\sum_{x_i} f(x_i) = 1$

$$\sum_{x_i} f(x_i) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + k = 1$$

$$k = 1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} \right) = 1 - \frac{31}{32} = \frac{1}{32}$$

$$f(x) = \begin{cases} \left(\frac{1}{2}\right)^x & x = 1,2,3,4,5 \\ \frac{1}{32} & x = 6 \end{cases}$$

La distribuzione di probabilità può essere scritta anche sotto forma di tabella (tabella 4) ed è rappresentata nella figura 4.

x_i	1	2	3	4	5	6
$f(x_i)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$

Tabella 4

La funzione di distribuzione è definita dalla tabella 5 ed è rappresentata nella figura 5.

x	$F(x) = P(X \leq x)$
$x < 1$	0
$1 \leq x < 2$	$\frac{1}{2}$
$2 \leq x < 3$	$\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$
$3 \leq x < 4$	$\frac{3}{4} + \frac{1}{8} = \frac{7}{8}$
$4 \leq x < 5$	$\frac{7}{8} + \frac{1}{16} = \frac{15}{16}$
$5 \leq x < 6$	$\frac{15}{16} + \frac{1}{32} = \frac{31}{32}$
$x \geq 6$	$\frac{31}{32} + \frac{1}{32} = 1$

Tabella 5

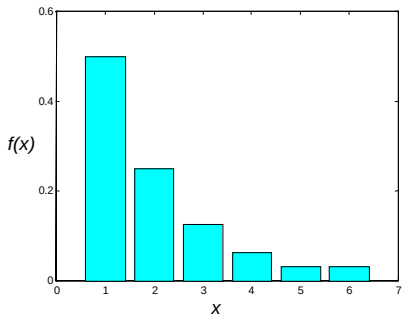


Figura 4

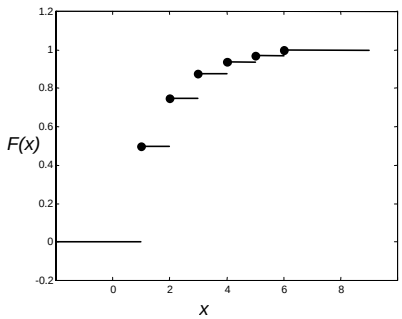


Figura 5

Esempio 8

Sia data la funzione di distribuzione $F(x)$ della variabile aleatoria discreta X

$$F(x)=\begin{cases}0 & x<-2\\ 0.2 & -2\le x<0\\ 0.7 & 0\le x<2\\ 1 & x\ge 2\end{cases}$$

Determinare la distribuzione di probabilità $f(x)$.

Il grafico di $F(x)$ è il seguente

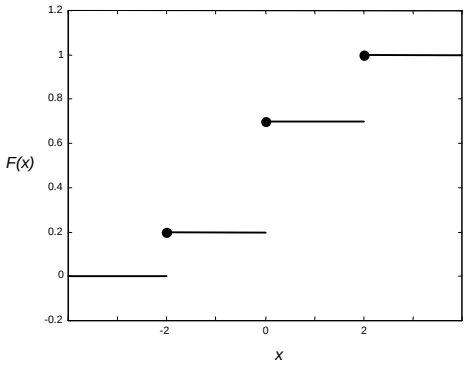


Figura 6

Si ha

$$f(-2)=0.2-0=0.2$$
$$f(0)=0.7-0.2=0.5$$
$$f(2)=1-0.7=0.3$$

La distribuzione di probabilità $f(x)$ è la seguente

$$f(x)=\begin{cases}0.2 & x=-2\\ 0.5 & x=0\\ 0.3 & x=2\end{cases}$$

Esempio 9

Si consideri la variabile aleatoria discreta X = numero ottenuto nel lancio di un dado; i valori che X può assumere sono i numeri 1,2,...,6.

La distribuzione di probabilità è definita dalla tabella 6; il grafico è rappresentato nella figura 7

x_i	1	2	3	4	5	6
$f(x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tabella 6

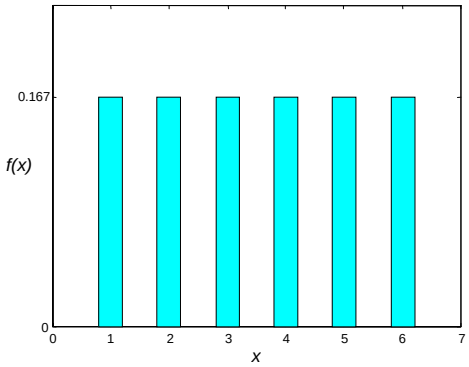


Figura 7

La funzione di distribuzione $F(x)$ è definita dalla tabella 7. $F(x)$ è una funzione a gradino; il salto fra i gradini è costante e vale sempre $\frac{1}{6}$, il grafico è rappresentato nella figura 8.

x	$F(x) = P(X \le x)$
$x < 1$	0
$1 \le x < 2$	$\frac{1}{6}$
$2 \le x < 3$	$\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$
$3 \le x < 4$	$\frac{1}{6} + \frac{1}{3} = \frac{1}{2}$
$4 \le x < 5$	$\frac{1}{6} + \frac{1}{2} = \frac{2}{3}$
$5 \le x < 6$	$\frac{1}{6} + \frac{2}{3} = \frac{5}{6}$
$x \ge 6$	$\frac{1}{6} + \frac{5}{6} = 1$

Tabella 7

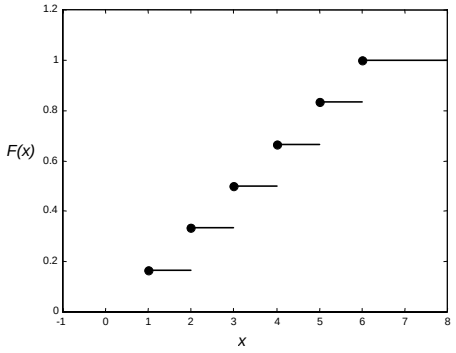


Figura 8

Esempio 10
Si effettua il lancio di due dadi. La variabile aleatoria X è la somma dei risultati dei due dadi. Determinare la distribuzione di probabilità $f(x)$ e la funzione di distribuzione $F(x)$ e disegnarne i grafici.

Lo spazio campione S è illustrato dalla figura 9

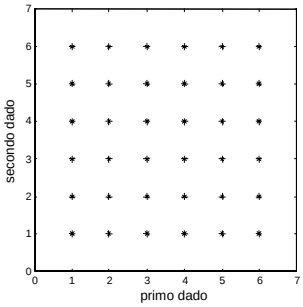


Figura 9

La distribuzione di probabilità $f(x)$ è data dalla tabella 8; il grafico è rappresentato nella figura 10 (pag. seguente)

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tabella 8

La funzione di distribuzione $F(x)$ è definita dalla tabella 9 ed è una funzione a gradino con salto non costante (figura 11)

x	$F(x)=P(X\le x)$	x	$F(x)=P(X\le x)$
$x<2$	0	$7\le x<8$	$\frac{7}{12}$
$2\le x<3$	$\frac{1}{36}$	$8\le x<9$	$\frac{13}{18}$
$3\le x<4$	$\frac{1}{12}$	$9\le x<10$	$\frac{15}{18}$
$4\le x<5$	$\frac{1}{6}$	$10\le x<11$	$\frac{11}{12}$
$5\le x<6$	$\frac{5}{18}$	$11\le x<12$	$\frac{35}{36}$
$6\le x<7$	$\frac{15}{36}$	$x\ge 12$	1

Tabella 9

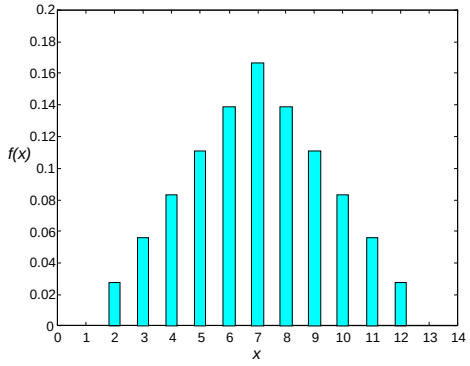


Figura 10

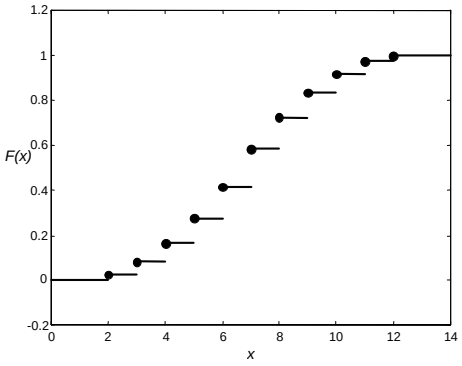
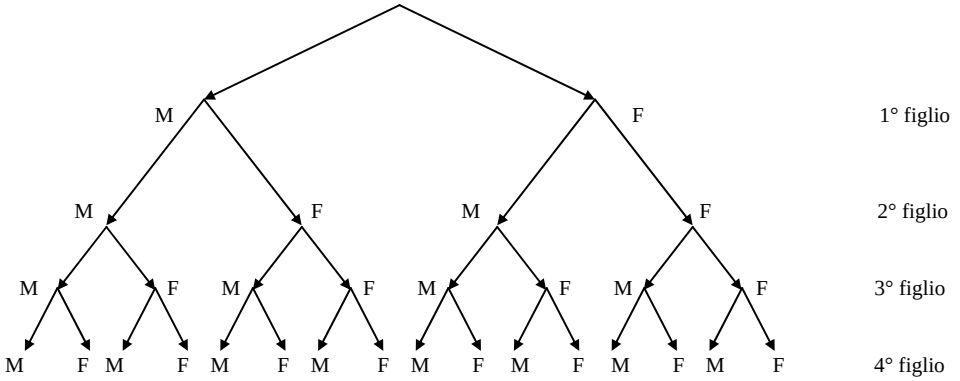


Figura 11

Esempio 11

Si considerino le famiglie con 4 figli; la composizione delle famiglie, tenendo conto del sesso dei figli e dell'ordine di nascita, si può rappresentare con il seguente diagramma ad albero



Il numero dei casi possibili è $2^4=16$. Se si trascura l'ordine di nascita, i 16 casi si riducono ai 5 seguenti

MMMM MMMF MMFF MFFF FFFF

Supponendo che gli eventi “nascita di un maschio” e “nascita di una femmina” siano equiprobabili, si costruisce la seguente tabella della distribuzione di probabilità

evento	MMMM	MMMF	MMFF	MFFF	FFFF
n° casi favorevoli	1	4	6	4	1
probabilità	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

Tabella 10

Scegliendo come variabile aleatoria X il numero delle figlie femmine, la tabella della distribuzione di probabilità $f(x)$ può essere riscritta nel modo seguente

x_i	0	1	2	3	4
$f(x_i)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

Tabella 11

La funzione di distribuzione è la seguente

x	$F(x) = P(X \le x)$
$x < 0$	0
$0 \le x < 1$	$\frac{1}{16}$
$1 \le x < 2$	$\frac{5}{16}$
$2 \le x < 3$	$\frac{11}{16}$
$3 \le x < 4$	$\frac{15}{16}$
$x \ge 4$	1

Tabella 12

I grafici di $f(x)$ e di $F(x)$ sono rappresentati nelle figure 12 e 13.

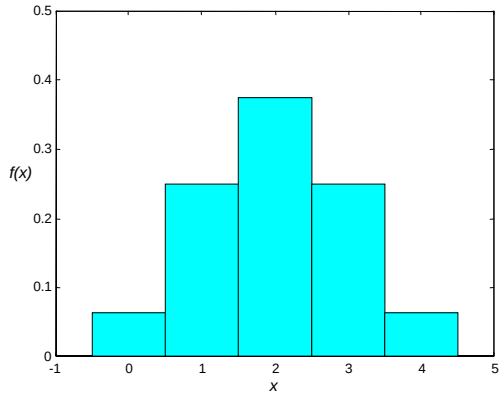


Figura 12

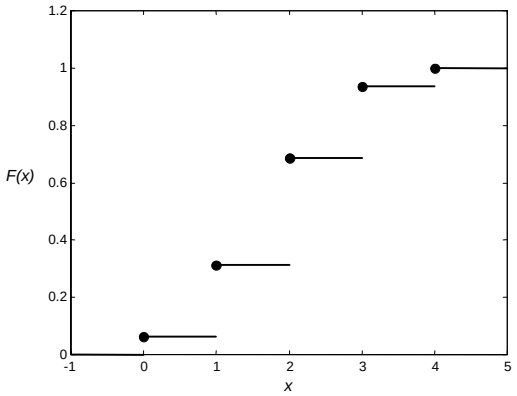


Figura 13

3.3 Densità di probabilità

Se X è una variabile aleatoria continua, la probabilità che X assuma un certo valore x fissato è in generale zero (si veda anche l'osservazione al termine di questo §, pag. 101), quindi non ha senso definire una distribuzione di probabilità con lo stesso procedimento seguito per una variabile aleatoria discreta. Nel caso di una variabile aleatoria continua ha senso invece calcolare la probabilità che X sia compresa fra a e b, dove a e b sono costanti, con $a \le b$.

Esempio 12
Se si sceglie a caso un adulto da una popolazione e si misura la sua altezza, la probabilità che l'altezza X sia esattamente 175 cm è uguale a zero, perché la misura viene fatta con uno strumento avente precisione finita. Tuttavia si ha una certa probabilità non nulla che X sia compresa ad esempio fra 174.9 cm e 175.1 cm.

In base a queste considerazioni, e in analogia con le proprietà (3.4) e (3.5) valide per le variabili discrete, si presuppone l'esistenza di una funzione $f(x)$ tale che

1)	$f(x) \ge 0 \qquad \forall x \in \mathbf{R}$	(3.6)
2)	$\int_{-\infty}^{\infty} f(x)dx = 1$	(3.7)

Si definisce poi la probabilità che X sia compresa fra a e b nel modo seguente

$$P(a < X < b) = \int_a^b f(x)dx$$

Si può dimostrare che questa definizione soddisfa gli assiomi della teoria della probabilità.
Una funzione $f(x)$ che soddisfi le condizioni (3.6) e (3.7) è detta **densità di probabilità**.

Esempio 13
Sia data la funzione

$$f(x) = \begin{cases} \dfrac{x}{8} & 0 \le x \le 4 \\ 0 & \text{altrimenti} \end{cases}.$$

Verificare che $f(x)$ è una densità di probabilità di una variabile aleatoria continua X e calcolare la probabilità che la variabile aleatoria X avente densità di probabilità $f(x)$ sia
a – minore di 2;
b – compresa fra 1 e 3.

Deve essere
1) $f(x) \ge 0 \qquad \forall x \in \mathbf{R}$
2) $\int_{-\infty}^{\infty} f(x)dx = 1$

La prima condizione è verificata $\forall x \in \mathbf{R}$. Inoltre si ha

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^4 \frac{x}{8}dx = \left.\frac{x^2}{16}\right|_0^4 = 1.$$

a – $$P(X < 2) = \int_{-\infty}^{2} f(x)dx = \int_0^2 \frac{x}{8}dx = \left.\frac{x^2}{16}\right|_0^2 = \frac{1}{4}$$

b – $$P(1<X<3)=\int_1^3 f(x)dx=\int_1^3 \frac{x}{8}dx=\left.\frac{x^2}{16}\right|_1^3=\frac{9}{16}-\frac{1}{16}=\frac{1}{2}$$

In analogia al caso della variabile aleatoria discreta, la funzione di distribuzione $F(x)$ è definita mediante l'integrazione della funzione $f(x)$.

Definizione 5
Si definisce **funzione di distribuzione** o **funzione di ripartizione** della variabile aleatoria continua X la funzione

$$F(x)=P(X\le x)=\int_{-\infty}^{x} f(t)dt \qquad (3.8)$$

Affinché la definizione abbia senso basta che $f(x)$ sia integrabile; come si vedrà negli esempi seguenti non è necessario che $f(x)$ sia continua.
Dalla definizione di $F(x)$ come funzione integrale, segue che $F(x)$ è una funzione continua; inoltre, per il teorema fondamentale del calcolo integrale, in tutti i punti in cui $f(x)$ è continua, la derivata della funzione di distribuzione $F(x)$ è la densità di probabilità $f(x)$

$$\frac{dF(x)}{dx}=f(x).$$

La densità di probabilità $f(x)$ di una variabile aleatoria X può essere rappresentata graficamente mediante una curva come nella figura 14 (in questo grafico è rappresentata una densità $f(x)$ continua particolarmente importante, la densità normale, che sarà trattata nel cap. 5). Per le proprietà (3.6) e (3.7) la curva non può andare sotto l'asse delle x e l'intera area compresa fra la curva e l'asse x è uguale a 1. Geometricamente la probabilità che X sia compresa fra a e b è rappresentata dall'area colorata.
La funzione di distribuzione $F(x)$ è una funzione continua, monotona crescente da 0 a 1 ed è rappresentata da una curva del tipo della figura 15.

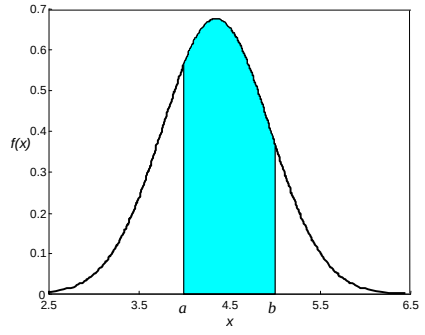


Figura 14

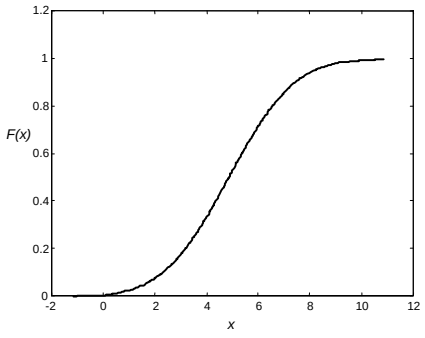


Figura 15

Osservazione – Eventi di probabilità nulla.
La definizione di probabilità nel caso continuo presuppone l'esistenza di un'opportuna funzione $f(x)$, il cui integrale sull'intervallo (a,b) fornisce la probabilità che la variabile aleatoria continua X assuma valori appartenenti ad (a,b); se l'intervallo si riduce a un solo punto l'integrale è nullo.
Pertanto, se X è una variabile aleatoria continua, la probabilità che essa assuma un valore fissato è sempre zero

$$P(X=x)=0 \qquad \forall x\in\mathbf{R}.$$

Questo fatto è importante per più motivi.

1 – Nel continuo l'espressione "evento di probabilità nulla" non è sinonimo di "evento impossibile", come invece accade nel discreto. Dunque nel continuo è significativo soltanto calcolare la probabilità che X assuma valori in un dato intervallo: questa è una prima sostanziale differenza tra variabili discrete e continue.

2 – Quanto detto al punto 1 significa che, se X è una variabile aleatoria continua, allora

$$P(X \le a) = P(X < a)$$
$$P(X \ge a) = P(X > a)$$
$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)\ .$$

3 – Da questo segue anche che la densità $f(x)$ non rappresenta la probabilità $P(X = x)$. Infatti la probabilità $P(X = x)$ è sempre nulla per ogni $x \in \mathbf{R}$, mentre $f(x)$ non è dappertutto nulla.

La funzione $f(x)$ non è una probabilità, è solo il suo integrale su un intervallo che ha il significato di probabilità. Nel caso discreto invece, la distribuzione di probabilità $f(x_k)$ è per definizione la probabilità $P(X = x_k)$.

In conclusione distribuzioni discrete e densità continue sono oggetti matematici di tipo diverso, non confrontabili tra loro; lo strumento che consente di confrontare variabili aleatorie discrete e continue sono invece le rispettive funzioni di distribuzione.

Esempio 14

Definiamo la funzione $f(x)$ (figura 16)

$$f(x) = \begin{cases} \dfrac{1}{2}x & 0 \le x \le 2 \\ 0 & \text{altrimenti} \end{cases}$$

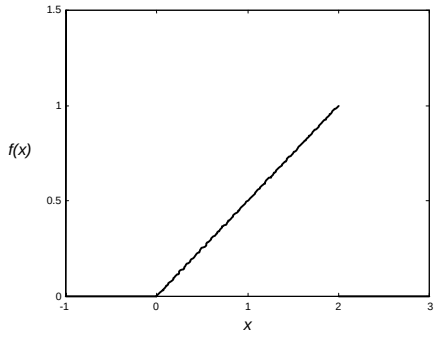


Figura 16

Si può verificare che $f(x)$ è una densità di probabilità; infatti

1) $f(x) \ge 0 \qquad \forall x \in \mathbf{R}$

2) $$\int_{-\infty}^{\infty} f(x)dx = \int_0^2 \frac{1}{2}x\,dx = \frac{1}{2}\frac{x^2}{2}\Big|_0^2 = 1$$

Troviamo la funzione di distribuzione (figura 17)

Per $x < 0$ $\qquad F(x) = 0$

Per $0 \le x \le 2$

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_0^x f(t)\,dt = \int_0^x \frac{1}{2}t\,dt = \frac{1}{4}x^2$$

Per $x > 2$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_0^2 \frac{1}{2}tdt = 1$$

$$F(x)=\begin{cases}0 & x<0\\ \dfrac{1}{4}x^2 & 0\le x\le 2\\ 1 & x>2\end{cases}$$

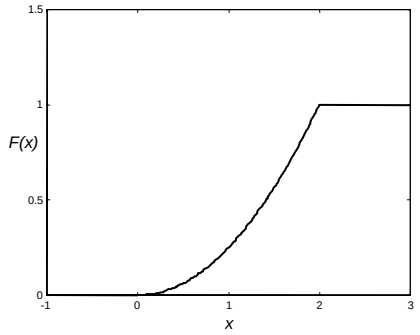


Figura 17

Esempio 15

Trovare la probabilità che una variabile aleatoria X avente la densità di probabilità

$$f(x)=\begin{cases}x & 0<x<1\\ 2-x & 1\le x<2\\ 0 & \text{altrimenti}\end{cases}$$

assuma valori compresi

a – fra 0.2 e 0.8

b – fra 0.6 e 1.2

c – maggiori di 1.8 .

a – $$P(0.2<X<0.8)=\int_{0.2}^{0.8} f(x)dx=\int_{0.2}^{0.8} x dx=\frac{x^2}{2}\Big|_{0.2}^{0.8}=\frac{0.64}{2}-\frac{0.04}{2}=0.3$$

b – $$P(0.6<X<1.2)=\int_{0.6}^{1.2} f(x)dx=\int_{0.6}^{1} x dx+\int_{1}^{1.2}(2-x)dx=$$

$$=\frac{x^2}{2}\Big|_{0.6}^{1}+\left(-\frac{(2-x)^2}{2}\right)\Big|_{1}^{1.2}=\frac{1}{2}-\frac{0.36}{2}-\frac{0.64}{2}+\frac{1}{2}=0.5$$

c – $$P(X>1.8)=\int_{1.8}^{2} f(x)dx=\int_{1.8}^{2}(2-x)dx=\left(-\frac{(2-x)^2}{2}\right)\Big|_{1.8}^{2}=\frac{0.04}{2}=0.02$$

Gli stessi risultati si possono ottenere ricavando la funzione di distribuzione $F(x)$

$$F(x)=\int_{-\infty}^{x} f(t)dt$$

Per $x\le 0$ $\quad F(x)=0$

Per $0<x<1$ $$F(x)=\int_{0}^{x} t dt=\frac{x^2}{2}$$

Per $1\le x<2$ $$F(x)=\int_{0}^{x} f(t)dt=\int_{0}^{1} t dt+\int_{1}^{x}(2-t)dt=\frac{1}{2}+\left(2t-\frac{t^2}{2}\right)\Big|_{1}^{x}=-\frac{x^2}{2}+2x-1$$

Per $x\ge 2$ $\quad F(x)=1$

$$F(x)=\begin{cases}0 & x\le 0\\ \dfrac{x^2}{2} & 0<x<1\\ -\dfrac{x^2}{2}+2x-1 & 1\le x<2\\ 1 & x\ge 2\end{cases}$$

$$P(0.2<X<0.8)=F(0.8)-F(0.2)=\frac{0.64}{2}-\frac{0.04}{2}=0.3$$

$$P(0.6<X<1.2)=F(1.2)-F(0.6)=-\frac{1.44}{2}+2.4-1-\frac{0.36}{2}=0.5$$

$$P(X>1.8)=1-P(X<1.8)=1-F(1.8)=1-\left(-\frac{1.8^2}{2}+3.6-1\right)=1-0.98=0.02$$

Esempio 16

La funzione di distribuzione di una variabile aleatoria X è

$$F(x)=\begin{cases}1-e^{-2x} & x>0\\ 0 & x\le 0\end{cases}$$

a – Calcolare le probabilità $P(X>2)$ e $P(-3<X\le 4)$.

b – Determinare la densità di probabilità $f(x)$.

a – $$P(X>2)=1-P(x\le 2)=1-F(2)=1-(1-e^{-4})=e^{-4}\cong 0.0183$$

$$P(-3<X\le 4)=F(4)-F(-3)=1-e^{-8}-0\cong 0.9997$$

b – $$f(x)=\frac{dF(x)}{dx}=\begin{cases}2e^{-2x} & x>0\\ 0 & x\le 0\end{cases}$$

Esempio 17

Sapendo che la funzione di distribuzione di una variabile aleatoria continua X è

$$F(x)=\begin{cases}0 & x<0\\ \left(\dfrac{x}{3}\right)^3 & 0\le x\le 3\\ 1 & x>3\end{cases}$$

calcolare le probabilità $P\left(1<X<\frac{3}{2}\right)$ e $P(2<X<4)$.

a – $$P\left(1<X<\frac{3}{2}\right)=F\left(\frac{3}{2}\right)-F(1)=\left(\frac{3}{6}\right)^3-\left(\frac{1}{3}\right)^3=\frac{1}{8}-\frac{1}{27}=\frac{19}{216}\equiv 0.088=8.8\%$$

b – $$P(2<X<4)=F(4)-F(2)=1-\left(\frac{2}{3}\right)^3=\frac{19}{27}\cong 0.704=70.4\%$$

Esempio 18

Trovare il valore della costante $c\in\mathbf{R}$ in modo che la funzione

$$f(x)=\begin{cases}cx^2 & 0<x<3\\ 0 & \text{altrimenti}\end{cases}$$

sia una densità di probabilità.

Trovare la funzione di distribuzione $F(x)$ e calcolare la probabilità $P(1<X<2)$.

Deve essere

1) $f(x) \geq 0 \qquad \forall x \in \mathbf{R} \qquad \Rightarrow \; c \geq 0$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{3} cx^2 dx = c\frac{x^3}{3}\Big|_0^3 = 9c$$

$$9c = 1 \;\Rightarrow\; c = \frac{1}{9}$$

La densità di probabilità è pertanto

$$f(x) = \begin{cases} \frac{1}{9}x^2 & 0 < x < 3 \\ 0 & \text{altrimenti} \end{cases}$$

Troviamo la funzione di distribuzione

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt\,.$$

Per $x \leq 0$ $\qquad F(x) = 0$

Per $0 < x < 3$ $\qquad F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} f(t)dt = \int_{0}^{x} \frac{1}{9}t^2 dt = \frac{1}{27}x^3$

Per $x \geq 3$ $\qquad F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{3} f(t)dt + \int_{3}^{x} f(t)dt = \int_{0}^{3} \frac{1}{9}t^2 dt = 1$

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{27}x^3 & 0 < x < 3 \\ 1 & x \geq 3 \end{cases}$$

$$P(1 < X < 2) = F(2) - F(1) = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

Oppure

$$P(1 < X < 2) = \int_{1}^{2} \frac{1}{9}x^2 dx = \frac{1}{9}\frac{x^3}{3}\Big|_1^2 = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

Osserviamo esplicitamente che

$$P(1 < X < 2) = P(1 \leq X < 2) = P(1 < X \leq 2) = P(1 \leq X \leq 2) = \frac{7}{27}$$

Esempio 19

Trovare il valore della costante $c \in \mathbf{R}$ tale che la funzione

$$f(x) = \begin{cases} \frac{c}{x^2} & 1 \leq x \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

sia una densità di probabilità e disegnare il grafico di $f(x)$.

Trovare la probabilità che la variabile aleatoria *X*, avente densità di probabilità $f(x)$, sia compresa fra 1.5 e 2 .

Deve essere

1) $f(x) \geq 0 \qquad \forall x \in \mathbf{R} \quad \Rightarrow \quad c \geq 0$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{1}^{2} \frac{c}{x^2} dx = c\left(-\frac{1}{x}\right)\Big|_1^2 = c\left(-\frac{1}{2}+1\right) = \frac{c}{2}$$

$$\frac{c}{2} = 1 \Rightarrow c = 2$$

$$f(x) = \begin{cases} \dfrac{2}{x^2} & 1 \leq x \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

Il grafico di $f(x)$ è il seguente

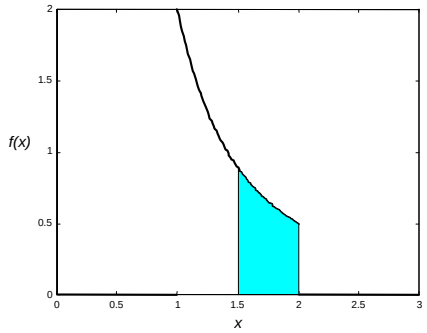


Figura 18

$$P\left(\frac{3}{2} < X < 2\right) = \int_{\frac{3}{2}}^{2} \frac{2}{x^2} dx = 2\left(-\frac{1}{x}\right)\Big|_{\frac{3}{2}}^{2} = 2\left(-\frac{1}{2}+\frac{2}{3}\right) = \frac{1}{3}$$

Nella figura 18 l'area colorata rappresenta la probabilità $P\left(\frac{3}{2} < X < 2\right)$.

Esempio 20

Trovare il valore della costante $k \in \mathbf{R}$ in modo che la funzione

$$f(x) = \begin{cases} k - 5 + x & 5 \leq x \leq 6 \\ 0 & \text{altrimenti} \end{cases}$$

sia una densità di probabilità.

Trovare la funzione di distribuzione $F(x)$ e calcolare le probabilità

a – $P(5 < X < 5.5)$

b – $P(5 < X < 6)$

c – $P(5.5 < X < 7)$.

Deve essere

1) $f(x) \geq 0 \qquad \forall x \in \mathbf{R} \quad \Rightarrow \; k \geq 0$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{5}^{6} (k - 5 + x)\, dx = (k-5)x + \frac{x^2}{2}\Big|_5^6 = k + \frac{1}{2}$$

$$k+\frac{1}{2}=1 \;\Rightarrow\; k=\frac{1}{2}$$

La densità di probabilità è pertanto

$$f(x)=\begin{cases} x-\dfrac{9}{2} & 5\le x\le 6 \\ 0 & \text{altrimenti} \end{cases}$$

Troviamo la funzione di distribuzione

$$F(x)=P(X\le x)=\int_{-\infty}^{x} f(t)dt$$

Per $x<5$ $\qquad F(x)=0$

Per $5\le x\le 6$ $\qquad F(x)=\int_{-\infty}^{x} f(t)dt=\int_{5}^{x} f(t)dt=\int_{5}^{x}\left(t-\frac{9}{2}\right)dt=\frac{1}{2}t^2-\frac{9}{2}t\Big|_5^x=\frac{1}{2}x^2-\frac{9}{2}x+10$

Per $x>6$ $\qquad F(x)=\int_{-\infty}^{x} f(t)dt=\int_{5}^{6}\left(t-\frac{9}{2}\right)dt=\frac{1}{2}t^2-\frac{9}{2}t\Big|_5^6=1$

$$F(x)=\begin{cases} 0 & x<5 \\ \dfrac{1}{2}x^2-\dfrac{9}{2}x+10 & 5\le x\le 6 \\ 1 & x>6 \end{cases}$$

a – $\qquad P(5<X<5.5)=F(5.5)-F(5)=F(5.5)=\frac{3}{8}$

b – $\qquad P(5<X<6)=1$

c –

$$P(5.5<X<7)=P(5.5<X<6)=1-P(5<X<5.5)=1-F(5.5)=1-\frac{3}{8}=\frac{5}{8}$$

Sia il valore di k che le probabilità possono essere ricavate anche per via geometrica.
Il grafico della funzione $f(x)$ assegnata è del tipo rappresentato nella figura 19. Il valore di k può essere trovato imponendo che l'area del trapezio nella figura 19 sia uguale a 1.

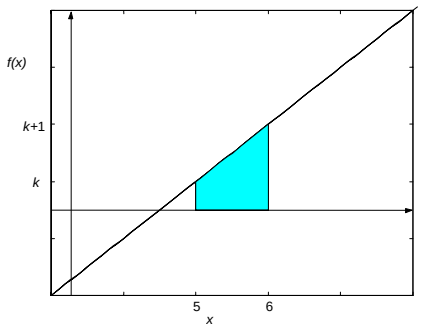


Figura 19

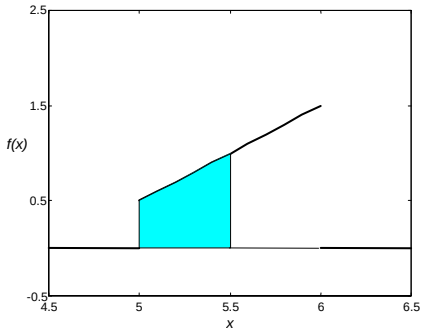


Figura 20

$$\text{Area trapezio}=\frac{(k+k+1)\cdot 1}{2}=k+\frac{1}{2}=1 \qquad \Rightarrow\; k=\frac{1}{2}.$$

In modo analogo si possono calcolare le probabilità.
Ad esempio la probabilità $P(5<X<5.5)$ è uguale all'area del trapezio colorato nella figura 20

$$P(5<X<5.5)=\frac{1}{2}\left(\frac{1}{2}+1\right)\cdot\frac{1}{2}=\frac{3}{8}$$

3.4 Parametri di una distribuzione

Nel Capitolo 1 abbiamo introdotto il concetto di valor medio di un insieme di dati, che consiste semplicemente nella media aritmetica di *n* valori assunti da una variabile numerica; introduciamo ora un concetto simile, che riguarda le variabili aleatorie.
Data una variabile aleatoria *X*, alla sua distribuzione o densità di probabilità $f(x)$ sono associati alcuni numeri, detti **parametri della distribuzione** o **della densità di probabilità**, aventi lo stesso significato degli indici di posizione e di dispersione, introdotti per un insieme di dati.

Valor medio – Caso discreto

Sia data una variabile aleatoria discreta *X*, i cui valori possibili sono $x_1, x_2, ..., x_n$, con probabilità rispettivamente

$$P(X = x_1) = f(x_1),\ \ P(X = x_2) = f(x_2), ...,\ P(X = x_n) = f(x_n).$$

Definizione 6
Si definisce **valor medio** o **speranza matematica** di una variabile aleatoria discreta X la quantità

$$\mu = E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + ... + x_n P(X = x_n) = \\ = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i f(x_i) \qquad (3.9)$$

Un caso particolare si ha quando le probabilità $f(x_i)$ sono tutte uguali

$$f(x_i) = P(X = x_i) = \frac{1}{n} \qquad i = 1,2,....,n$$

in tal caso μ è la **media aritmetica** di $x_1, x_2, ..., x_n$

$$\mu = E(X) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Il valor medio di *X* è un numero che indica dove è "centrata" la variabile aleatoria *X*, ossia attorno a quale valore ci aspettiamo che cadano i valori di *X*; esso rappresenta quindi una misura di tendenza centrale. Il valor medio di *X* può non essere un valore effettivamente assunto da *X*.

Esempio 21
Se la variabile aleatoria X è il punteggio ottenuto nel lancio di un dado, poiché i 6 risultati possibili sono ugualmente probabili, si ha

$$\mu = E(X) = \frac{1+2+3+4+5+6}{6} = 3.5$$

Esempio 22
La variabile aleatoria *X* indica la somma dei punti ottenuti con il lancio di due dadi.
La tabella della distribuzione di probabilità $f(x)$ è la seguente (vedere esempio 10)

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tabella 13

Per il valor medio si ottiene

$$\mu = \sum_{i=2}^{12} x_i \cdot f(x_i) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + ... + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$$

Si noti che in questo esempio i valori x_i non sono ugualmente probabili.

Esempio 23
Trovare il valor medio della variabile aleatoria X definita come il numero di teste ottenute con tre lanci successivi di una moneta.

I casi possibili sono $2^3 = 8$

CCC	nessuna testa	$X = 0$
CCT, CTC, TCC	1 testa	$X = 1$
CTT, TCT, TTC	2 teste	$X = 2$
TTT	3 teste	$X = 3$

La distribuzione di probabilità $f(x)$ è definita dalla tabella 14

x_i	0	1	2	3
$f(x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Tabella 14

Il valor medio è

$$\mu = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}.$$

Esempio 24
Si lancia un dado: un giocatore vince € 2000 se esce il 2, € 4000 se esce il 4, perde € 3000 se esce il 6; se esce un numero dispari non vince né perde nulla.
Determinare il guadagno medio del giocatore.

La variabile aleatoria X indica il guadagno/perdita del giocatore.
Nella tabella 15 si riportano le probabilità associate ai guadagni/perdite

x_i	0	+ 2000	0	+ 4000	0	–3000
$f(x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tabella 15

Il valor medio è

$$\mu = 0 \cdot \frac{1}{6} + 2000 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 4000 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} - 3000 \cdot \frac{1}{6} = 500$$

Il guadagno medio è di € 500.

Gli esempi seguenti illustrano un'interpretazione del concetto di valor medio. Sia X una variabile aleatoria e consideriamo un gioco in cui si paga una somma fissa S per partecipare e si riceve una vincita variabile X. Il valor medio μ può essere visto come il valore da assegnare ad S affinché il **gioco** sia **equo**. Se $S > \mu$ il gioco è iniquo a favore del banco.

Esempio 25
Un giocatore acquista un biglietto di una lotteria: può vincere il primo premio di € 5000 con probabilità 0.001 e il secondo premio di € 2000 con probabilità 0.003. Quale dovrebbe essere il giusto prezzo del biglietto?

Calcoliamo il valor medio (speranza matematica)

$$\mu = 5000 \cdot 0.001 + 2000 \cdot 0.003 = 11.$$

Affinché il gioco sia equo, il prezzo giusto per il biglietto dovrebbe essere € 11.

Esempio 26
In una lotteria nazionale vengono messi in palio i seguenti premi

1° premio	€ 3.000.000
2° premio	€ 2.000.000
3° premio	€ 1.000.000
5 premi da	€ 100.000
20 premi da	€ 10.000
100 premi da	€ 1.000

Vengono venduti 2 milioni di biglietti; qual è il valor medio della vincita per chi acquista un biglietto? Se il biglietto costa € 5, il gioco è equo, ossia conviene partecipare alla lotteria?

Sia X la variabile aleatoria "premio vinto con un biglietto"; la distribuzione di probabilità è la seguente

x_i	3.000.000	2.000.000	1.000.000	100.000	10.000	1.000
$f(x_i)$	$\frac{1}{2.000.000}$	$\frac{1}{2.000.000}$	$\frac{1}{2.000.000}$	$\frac{5}{2.000.000}$	$\frac{20}{2.000.000}$	$\frac{100}{2.000.000}$

Tabella 16

Il valor medio della vincita con un biglietto è

$$\mu = 3.000.000 \cdot \frac{1}{2.000.000} + 2.000.000 \cdot \frac{1}{2.000.000} + 1.000.000 \cdot \frac{1}{2.000.000}$$
$$+100.000 \cdot \frac{5}{2.000.000} + 10.000 \cdot \frac{20}{2.000.000} + 1.000 \cdot \frac{100}{2.000.000} =$$
$$= 1.5 + 1 + 0.5 + 0.25 + 0.1 + 0.05 = 3.4$$

Poiché il prezzo del biglietto è di € 5, il gioco non è equo, ma è a sfavore di chi compra i biglietti. Se il gioco fosse equo, il biglietto della lotteria dovrebbe costare € 3.4 .

Valor medio – Caso continuo

Sia X una variabile aleatoria continua avente densità di probabilità $f(x)$.

Definizione 7
Si definisce **valor medio** di X la quantità

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \qquad (3.10)$$

Esempio 27
Sia data la densità di probabilità

$$f(x) = \begin{cases} \frac{1}{2}x & 0 \le x \le 2 \\ 0 & \text{altrimenti} \end{cases}$$

Il valor medio di X è

$$\mu = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{2} x \cdot \frac{1}{2} x dx = \left. \frac{x^3}{6} \right|_0^2 = \frac{4}{3}$$

Varianza e scarto quadratico medio

Definizione 8
Si definisce **varianza** della variabile aleatoria X la quantità

$$\sigma^2 = \text{var}(X) = E[(X-\mu)^2] \tag{3.11}$$

dove μ è il valor medio di X.

Definizione 9
La radice quadrata non negativa

$$\sigma = \sqrt{\text{var}(X)} = \sqrt{E[(X-\mu)^2]} \tag{3.12}$$

è detta **scarto quadratico medio** o **deviazione standard** di X.

La varianza (o la deviazione standard) è una misura della dispersione dei valori della variabile aleatoria X attorno al valor medio μ.
Se i valori sono concentrati vicino alla media, la varianza è piccola, mentre se i valori sono dispersi lontano dal valor medio, la varianza è grande.
Il grafico della figura 21 illustra la situazione nel caso di due densità di probabilità continue aventi lo stesso valor medio μ e varianza diversa.

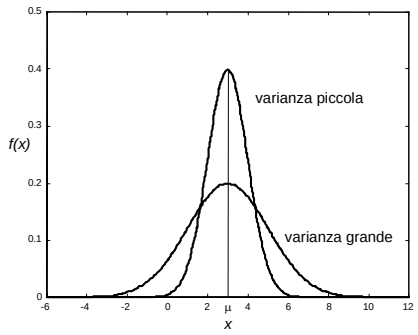


Figura 21

Varianza e scarto quadratico medio – Caso discreto

Sia data una variabile aleatoria discreta X, i cui valori possibili sono $x_1, x_2, ..., x_n$, con probabilità rispettivamente $f(x_1), f(x_2), ..., f(x_n)$.

Definizione 10
Si definisce **varianza** della variabile aleatoria discreta X, avente valor medio μ, la quantità

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \, f(x_i) \tag{3.13}$$

Definizione 11
Si definisce **deviazione standard** o **scarto quadratico medio** della variabile aleatoria discreta X, avente valor medio μ, la quantità

$$\sigma = \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 \, f(x_i)} \tag{3.14}$$

La varianza può anche essere calcolata con la seguente formula, alternativa alla (3.13)

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 \, f(x_i) - \mu^2 \tag{3.15}$$

Esempio 28
Trovare la varianza della variabile aleatoria X definita come il numero di teste ottenute con tre lanci successivi di una moneta .

Nell'esempio 23 è stato calcolato il valor medio $\mu = \frac{3}{2}$ della variabile X.

Per la varianza, con la (3.13) si ha

$$\sigma^2 = \sum_{i=1}^{4}\left(x_i - \frac{3}{2}\right)^2 f(x_i) = \left(-\frac{3}{2}\right)^2 \frac{1}{8} + \left(1-\frac{3}{2}\right)^2 \frac{3}{8} + \left(2-\frac{3}{2}\right)^2 \frac{3}{8} + \left(3-\frac{3}{2}\right)^2 \frac{1}{8} = \frac{3}{4}.$$

Esempio 29
Trovare la varianza della variabile aleatoria X definita come la somma dei punti ottenuti con il lancio di due dadi.

Nell'esempio 22 è stato calcolato il valor medio $\mu = 7$ della variabile X.
Per la varianza, con la (3.15) si ha

$$\sigma^2 = \sum_{i=1}^{11} x_i^2 f(x_i) - 49 = 4\cdot\frac{1}{36} + 9\cdot\frac{2}{36} + \ldots + 121\cdot\frac{2}{36} + 144\cdot\frac{1}{36} - 49 = \frac{35}{6}$$

Esempio 30
Sia data la funzione

$$f(x) = \begin{cases} \left(\frac{1}{2}\right)^x & x = 1,2,3,4 \\ k & x = 5 \end{cases}$$

a – Trovare il valore della costante $k \in \mathbf{R}$ in modo che la funzione sia una distribuzione di probabilità discreta.
b – Calcolare il valor medio e la varianza della variabile aleatoria discreta avente la distribuzione di probabilità $f(x)$.

a – Deve essere

1) $f(x_i) \geq 0 \qquad \forall x_i \qquad \Rightarrow \quad k \geq 0$

2) $\sum_{x_i} f(x_i) = 1$

$$\sum_{x_i} f(x_i) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + k = 1$$

$$k = 1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}\right) = 1 - \frac{15}{16} = \frac{1}{16}$$

$$f(x) = \begin{cases} \left(\frac{1}{2}\right)^x & x = 1,2,3,4 \\ \frac{1}{16} & x = 5 \end{cases}$$

La distribuzione di probabilità può essere scritta anche sotto forma di tabella (tabella 17) ed è rappresentata nella figura 22.

x_i	$f(x_i)$
1	$\frac{1}{2}$
2	$\frac{1}{4}$
3	$\frac{1}{8}$
4	$\frac{1}{16}$
5	$\frac{1}{16}$

Tabella 17

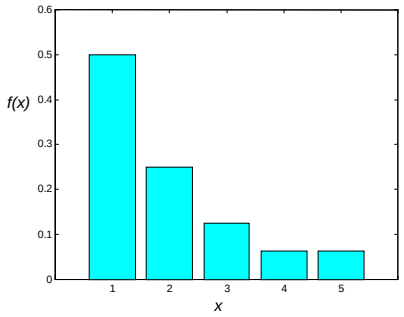


Figura 22

b – Valor medio

$$\mu = \sum_{i=1}^{5} x_i f(x_i) = 1\cdot\frac{1}{2} + 2\cdot\frac{1}{4} + 3\cdot\frac{1}{8} + 4\cdot\frac{1}{16} + 5\cdot\frac{1}{16} = \frac{31}{16}$$

Varianza

$$\sigma^2 = \sum_{i=1}^{5} (x_i - \mu)^2 f(x_i) = \left(1-\frac{31}{16}\right)^2\cdot\frac{1}{2} + \left(2-\frac{31}{16}\right)^2\cdot\frac{1}{4} +$$

$$+\left(3-\frac{31}{16}\right)^2\cdot\frac{1}{8} + \left(4-\frac{31}{16}\right)^2\cdot\frac{1}{16} + \left(5-\frac{31}{16}\right)^2\cdot\frac{1}{16} = \frac{367}{256} \cong 1.4336$$

Varianza e scarto quadratico medio – Caso continuo

Sia X una variabile aleatoria continua avente densità di probabilità $f(x)$.

Definizione 12
Si definisce **varianza** della variabile aleatoria continua X la quantità

$$\sigma^2 = E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \qquad (3.16)$$

Definizione 13
Si definisce **deviazione standard** o **scarto quadratico medio** della variabile aleatoria continua X la quantità

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx} \qquad (3.17)$$

La varianza può anche essere calcolata con la seguente formula, alternativa alla (3.16)

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \qquad (3.18)$$

Esempio 31
Calcolare varianza e deviazione standard della densità di probabilità

$$f(x) = \begin{cases} \frac{1}{2}x & 0 \le x \le 2 \\ 0 & \text{altrimenti} \end{cases}$$

Il valor medio è $\mu = \frac{4}{3}$ (vedere esempio 27). Per la varianza, con la (3.16) si ha

$$\sigma^2 = \int_{-\infty}^{\infty} \left(x - \frac{4}{3} \right)^2 f(x)dx = \int_0^2 \left(x - \frac{4}{3} \right)^2 \frac{x}{2} dx = \frac{2}{9}$$

$$\sigma = \sqrt{\frac{2}{9}} = \frac{\sqrt{2}}{3}.$$

Applicando in alternativa la (3.18) si ha

$$\sigma^2 = \int_0^2 x^2 \frac{1}{2} x dx - \frac{16}{9} = \left| \frac{1}{8} x^4 \right|_0^2 - \frac{16}{9} = 2 - \frac{16}{9} = \frac{2}{9}.$$

Esempio 32

Data la densità di probabilità

$$f(x) = \begin{cases} \frac{2(x+1)}{3} & 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

trovare il valor medio e la varianza.

$$\mu = \int_0^1 x \cdot \frac{2(x+1)}{3} dx = \frac{2}{3} \int_0^1 \left(x^2 + x \right) dx = \frac{2}{3} \cdot \left| \frac{x^3}{3} + \frac{x^2}{2} \right|_0^1 = \frac{2}{3} \left(\frac{1}{3} + \frac{1}{2} \right) = \frac{5}{9}$$

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 = \int_0^1 x^2 \cdot \frac{2(x+1)}{3} dx - \left(\frac{5}{9} \right)^2 = \frac{2}{3} \int_0^1 \left(x^3 + x^2 \right) dx - \frac{25}{81} = \frac{13}{162}$$

Esempio 33

Sia data la funzione

$$f(x) = \begin{cases} \frac{3\left(1 - x^2\right)}{4} & -1 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

a – Verificare che è una densità di probabilità; disegnare il grafico di $f(x)$.

b – Trovare la funzione di distribuzione e disegnarne il grafico.

c – Calcolare la probabilità che la variabile aleatoria *X* avente densità di probabilità $f(x)$ assuma valori maggiori di $\frac{1}{4}$.

d – Calcolare il valor medio e la varianza della densità di probabilità $f(x)$.

a – Deve essere

1) $f(x) \geq 0 \qquad \forall x \in \mathbf{R}$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{1} \frac{3\left(1 - x^2\right)}{4} dx = \frac{3}{4} \left| x - \frac{x^3}{3} \right|_{-1}^{1} = \frac{3}{4} \left(2 - \frac{2}{3} \right) = 1$$

La figura 23 illustra il grafico di $f(x)$.

b – Troviamo la funzione di distribuzione $F(x)$ (figura 24)

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Per $x \le -1$ $\qquad F(x) = 0$

Per $-1 < x < 1$ $\qquad F(x) = \int_{-1}^{x} \frac{3\left(1-t^2\right)}{4} dt = \frac{3}{4}\left| t - \frac{t^3}{3} \right|_{-1}^{x} = \frac{3}{4}\left(-\frac{x^3}{3} + x + \frac{2}{3} \right)$

Per $x \ge 1$ $\qquad F(x) = 1$

$$F(x) = \begin{cases} 0 & x \le -1 \\ \frac{3}{4}\left(-\frac{x^3}{3} + x + \frac{2}{3} \right) & -1 < x < 1 \\ 1 & x \ge 1 \end{cases}$$

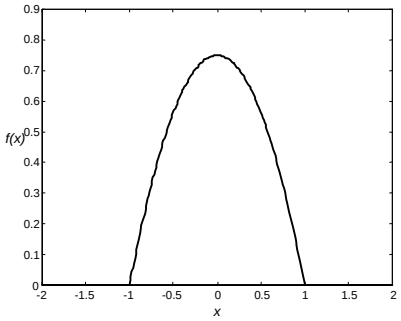


Figura 23

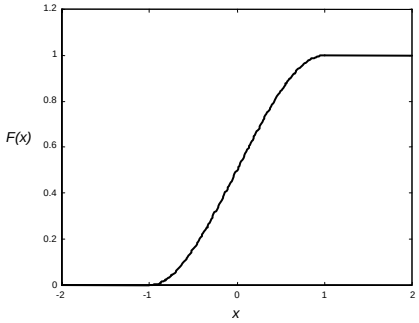


Figura 24

c – $\qquad P\left(X > \frac{1}{4} \right) = 1 - F\left(\frac{1}{4} \right) = 1 - \frac{3}{4}\left(-\frac{1}{3}\frac{1}{64} + \frac{1}{4} + \frac{2}{3} \right) = \frac{81}{256} \cong 0.3164$

d – Dalla figura 23 si osserva che $f(x)$ è simmetrica rispetto alla retta $x = 0$; in tal caso il valor medio è $\mu = 0$.

Per la varianza con la formula (3.18) si ottiene

$$\sigma^2 = \int_{-1}^{1} x^2 f(x) dx - \mu^2 = \int_{-1}^{1} x^2 \cdot \frac{3\left(1-x^2\right)}{4} dx = \frac{3}{4}\int_{-1}^{1} \left(x^2 - x^4 \right) dx =$$

$$= \frac{3}{4}\left| \frac{x^3}{3} - \frac{x^5}{5} \right|_{-1}^{1} = \frac{3}{4}\left(\frac{1}{3} - \frac{1}{5} + \frac{1}{3} - \frac{1}{5} \right) = \frac{1}{5}$$

Esempio 34

Trovare il valore della costante $a \in \mathbf{R}$ in modo che la funzione

$$f(x) = \begin{cases} \frac{1}{2}x & 0 \le x \le 1 \\ a - x & 1 < x \le 2 \\ 0 & \text{altrimenti} \end{cases}$$

sia una densità di probabilità.

Trovare il valor medio μ e la varianza σ^2.

Calcolare la probabilità che la variabile aleatoria X avente densità di probabilità $f(x)$ sia

a – compresa fra 0.5 e 1;

b – compresa fra 0 e 1;

c – compresa fra 0.5 e 2.

Deve essere

1) $f(x) \geq 0 \qquad \forall x \in \mathbf{R} \quad \Rightarrow \quad a \geq 2$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} \frac{1}{2} x dx + \int_{1}^{2} (a-x)dx = \frac{1}{2}\cdot\frac{1}{2} + a + \left(-\frac{x^2}{2}\right)\Big|_1^2 = \frac{1}{4} + a - 2 + \frac{1}{2} = a - \frac{5}{4}$$

$$a - \frac{5}{4} = 1 \;\Rightarrow\; a = \frac{9}{4}$$

La densità di probabilità è pertanto

$$f(x) = \begin{cases} \frac{1}{2}x & 0 \leq x \leq 1 \\ \frac{9}{4} - x & 1 < x \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

Il grafico di $f(x)$ è rappresentato nella figura 25.

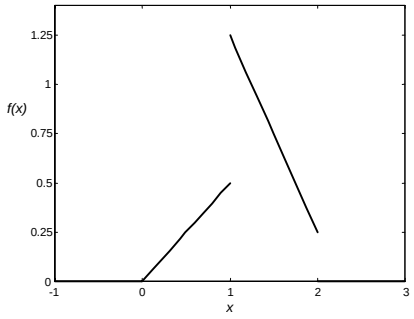


Figura 25

Valor medio

$$\mu = \int_{-\infty}^{\infty} x f(x)dx = \int_{0}^{2} x f(x)dx = \int_{0}^{1} \frac{1}{2} x^2 dx + \int_{1}^{2} x\left(\frac{9}{4} - x\right)dx =$$

$$= \frac{1}{2}\cdot\frac{1}{3} + \frac{9}{4}\frac{x^2}{2}\Big|_1^2 - \frac{x^3}{3}\Big|_1^2 = \frac{29}{24} \cong 1.2$$

Varianza (con la (3.18))

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 = \int_{0}^{1} \frac{1}{2} x^3 dx + \int_{1}^{2} \left(\frac{9}{4} - x\right)x^2 dx - \left(\frac{29}{24}\right)^2 =$$

$$= \frac{1}{8} + \frac{21}{4} - \frac{15}{4} - \left(\frac{29}{24}\right)^2 = \frac{95}{576} \cong 0.1649$$

Probabilità

a – $$P\left(\frac{1}{2} < X < 1\right) = \int_{\frac{1}{2}}^{1} \frac{1}{2} x dx = \frac{1}{4} x^2 \Big|_{\frac{1}{2}}^{1} = \frac{3}{16}$$

b – $$P(0 < X < 1) = \int_{0}^{1} \frac{1}{2} x dx = \frac{1}{4} x^2 \Big|_0^1 = \frac{1}{4}$$

c – $$P\left(\frac{1}{2} < X < 2\right) = \frac{3}{16} + \int_1^2 \left(\frac{9}{4} - x\right) dx = \frac{3}{16} + \frac{9}{4}x - \frac{1}{2}x^2 \Big|_1^2 = \frac{15}{16}$$

Queste probabilità possono anche essere trovate per via geometrica (vedere esempio 20).

Per il valor medio e la varianza valgono alcune proprietà.

Proprietà 1
Sia X una variabile aleatoria con valor medio $E(X)$; si ha

$$E(aX + b) = aE(X) + b \qquad a,b \in \mathbf{R} \tag{3.19}$$

$$\text{var}(aX + b) = a^2 \text{var}(X) \qquad a,b \in \mathbf{R} \tag{3.20}$$

Proprietà 2
Siano X e Y variabili aleatorie con valori medi $E(X)$ e $E(Y)$; si ha

$$E(aX + bY) = aE(X) + bE(\text{Y}) \qquad a,b \in \mathbf{R} \tag{3.21}$$

Proprietà 3
Siano X e Y variabili aleatorie indipendenti (ciò avviene se gli eventi $X = x$ e $Y = y$ sono indipendenti per ogni x e y); si ha

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) \qquad a,b \in \mathbf{R} \tag{3.22}$$

Un caso particolare delle proprietà 2 e 3, degno di nota, è il seguente

$$\begin{aligned} &E(X - Y) = E(X) - E(\text{Y}) \\ &\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) \end{aligned} \tag{3.23}$$

Definizione 14
Sia X una variabile aleatoria con valor medio μ e deviazione standard σ. Si definisce **variabile aleatoria standardizzata *Z*** associata a X la variabile aleatoria

$$Z = \frac{X - \mu}{\sigma} \tag{3.24}$$

Proprietà 4
La variabile aleatoria standardizzata Z ha valor medio 0 e varianza 1

$$\mu = E(Z) = 0 \qquad \sigma^2 = \text{var}(Z) = 1 \tag{3.25}$$

Esempio 35
Una variabile aleatoria discreta X ha la seguente distribuzione di probabilità

x_i	0	1	2
$f(x_i)$	0.4	0.4	0.2

Tabella 18

Calcolare il valor medio e la varianza della variabile aleatoria $Y = 2X - 1$.

Calcoliamo il valor medio e la varianza della variabile aleatoria X con le formule (3.9) e (3.15)

$$E(X) = 0 \cdot 0.4 + 1 \cdot 0.4 + 2 \cdot 0.2 = 0.8$$

$$\text{var}(X) = 0 \cdot 0.4 + 1 \cdot 0.4 + 4 \cdot 0.2 - 0.8^2 = 0.56$$

Calcoliamo ora il valor medio e la varianza della variabile aleatoria Y con le formule (3.19) e (3.20)

$$E(Y) = E(2X - 1) = 2E(X) - 1 = 2 \cdot 0.8 - 1 = 0.6$$

$$\text{var}(Y) = \text{var}(2X - 1) = 4 \cdot \text{var}(X) = 4 \cdot 0.56 = 2.24$$

Esempio 36
Determinare il valor medio e la varianza della somma dei punti ottenuti nel lancio di una coppia di dadi.

a – Valor medio. Per la proprietà 2 si ha

$$E(X+Y) = E(X) + E(Y)$$

$$E(X) = E(Y) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = \frac{7}{2}$$

$$E(X+Y) = \frac{7}{2} + \frac{7}{2} = 7$$

b – Varianza. Per la proprietà 3 si ha

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$$

Per il calcolo della varianza di X ci serviamo della formula (3.15)

$$\text{var}(X) = \sum_{i=1}^{n} x_i^2 f(x_i) - E(X)^2 = 1^2\cdot\frac{1}{6} + 2^2\cdot\frac{1}{6} + 3^2\cdot\frac{1}{6}$$

$$+ 4^2\cdot\frac{1}{6} + 5^2\cdot\frac{1}{6} + 6^2\cdot\frac{1}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

$$\text{var}(X) = \text{var}(Y) = \frac{35}{12}$$

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) = \frac{35}{12} + \frac{35}{12} = \frac{35}{6}$$

Altre misure di tendenza centrale – Moda e mediana

Come abbiamo già visto, il valor medio di una variabile aleatoria X fornisce una misura di tendenza centrale per i valori della distribuzione. Sebbene il valor medio sia la misura più usata per questo scopo, esistono anche altre misure.

Definizione 15
La **moda** $\tilde{x}$ è il valore che si verifica il maggior numero di volte, ossia che ha la maggior probabilità di verificarsi.

In corrispondenza a questo valore di x, $f(x)$ ha un massimo. A volte ci sono due o più valori di questo tipo: in tal caso la **distribuzione** si dice **bimodale** o **multimodale**.

Definizione 16
La **mediana** è il valore M per il quale si ha

$$P(X \le M) = P(X \ge M) = \frac{1}{2}$$

Nel caso di una distribuzione continua, la mediana corrisponde a un punto che separa la regione sottesa dalla curva $f(x)$ in due parti, entrambe di area uguale a $\frac{1}{2}$.

Esempio 37
Sia data la distribuzione (vedere esempio 7)

$$f(x)=\begin{cases}\left(\frac{1}{2}\right)^x & x=1,2,3,4,5\\ \frac{1}{32} & x=6\end{cases}$$

Calcolare il valor medio, la varianza, la moda e la mediana.

Valor medio (formula (3.9))

$$\mu=1\cdot\frac{1}{2}+2\cdot\frac{1}{4}+3\cdot\frac{1}{8}+4\cdot\frac{1}{16}+5\cdot\frac{1}{32}+6\cdot\frac{1}{32}=\frac{63}{32}\cong 1.97$$

Varianza (formula (3.15))

$$\sigma^2=1\cdot\frac{1}{2}+4\cdot\frac{1}{4}+9\cdot\frac{1}{8}+16\cdot\frac{1}{16}+25\cdot\frac{1}{32}+36\cdot\frac{1}{32}-\left(\frac{63}{32}\right)^2\cong 1.6553$$

Moda (vedere figura 4)

$$\tilde{x}=1$$

Mediana

$$M=1.5\ .$$

Infatti

$$P(X\le 1.5)=P(X=1)=\frac{1}{2}$$

$$P(X\ge 1.5)=P(X=2)+P(X=3)+\ldots+P(X=6)=\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\frac{1}{32}=\frac{1}{2}$$

Esempio 38
Trovare il valore della costante $k\in\mathbf{R}$ in modo che la funzione

$$f(x)=\begin{cases}kx^2 & -2\le x\le 1\\ 0 & \text{altrimenti}\end{cases}$$

sia una densità di probabilità.
Calcolare il valor medio μ, la moda $\tilde{x}$ e la mediana M.

Deve essere

1) $f(x)\ge 0 \qquad \forall x\in\mathbf{R} \quad\Rightarrow\quad k\ge 0$

2) $$\int_{-\infty}^{\infty} f(x)dx=1$$

$$\int_{-\infty}^{\infty} f(x)dx=\int_{-2}^{1} kx^2dx=k\left.\frac{x^3}{3}\right|_{-2}^{1}=k\left(\frac{1}{3}+\frac{8}{3}\right)=3k$$

$$3k=1\ \Rightarrow\ k=\frac{1}{3}$$

$$f(x)=\begin{cases}\frac{1}{3}x^2 & -2\le x\le 1\\ 0 & \text{altrimenti}\end{cases}$$

Valor medio

$$\mu=\int_{-\infty}^{\infty} xf(x)\,dx=\int_{-2}^{1} xf(x)\,dx=\int_{-2}^{1}\frac{1}{3}x^3dx=\left.\frac{1}{3}\frac{x^4}{4}\right|_{-2}^{1}=-\frac{5}{4}=-1.25$$

Moda

$$\tilde{x} = -2$$

In base alla definizione 16, la mediana è il valore *M* per il quale si verifica

$$P(X \le M) = \int_{-2}^{M} f(x)dx = \frac{1}{2}$$

$$\int_{-2}^{M} f(x)dx = \int_{-2}^{M} \frac{1}{3}x^2 dx = \frac{1}{3}\cdot\frac{x^3}{3}\Big|_{-2}^{M} = \frac{1}{9}\left(M^3+8\right)$$

$$\frac{1}{9}\left(M^3+8\right) = \frac{1}{2} \qquad \Rightarrow \qquad M = -\sqrt[3]{\frac{7}{2}} \cong -1.52$$

Nella figura 26 l'area ombreggiata vale $\frac{1}{2}$, ed è la metà dell'area totale sottesa da $f(x)$ nell'intervallo [–2,1].

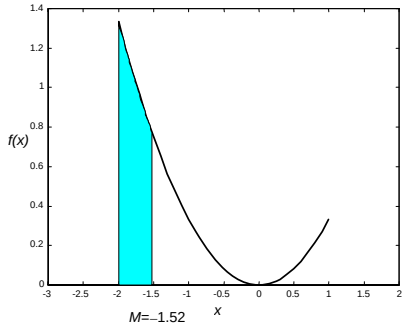


Figura 26

3.5 Disuguaglianza di Chebishev

Come già osservato, la varianza (o lo scarto quadratico medio) misura la dispersione di una distribuzione di probabilità.
Se la varianza σ^2 è piccola, c'è un'alta probabilità di ottenere valori della variabile aleatoria vicini al valor medio; se invece σ^2 è grande, c'è una maggior probabilità di ottenere valori lontani dal valor medio.
Queste considerazioni sono formalizzate dal seguente risultato.

Teorema 1 – Disuguaglianza di Chebishev
Sia *X* una variabile aleatoria con valor medio μ e varianza σ^2; allora per ogni $\varepsilon > 0$ si ha

$$P\left(|X-\mu| \ge \varepsilon\right) \le \frac{\sigma^2}{\varepsilon^2} \tag{3.26}$$

La relazione $|X-\mu| \ge \varepsilon$ equivale alle disuguaglianze

$$X \le \mu - \varepsilon \qquad X \ge \mu + \varepsilon,$$

quindi la disuguaglianza di Chebishev afferma che la probabilità che la variabile aleatoria *X* assuma un valore fuori dall'intervallo $(\mu - \varepsilon, \mu + \varepsilon)$ è minore o uguale a $\frac{\sigma^2}{\varepsilon^2}$; concludiamo perciò che più è piccola la varianza, minore è la probabilità che *X* assuma valori fuori dall'intervallo $(\mu-\varepsilon, \mu+\varepsilon)$.

La disuguaglianza di Chebishev viene spesso presentata anche nella seguente forma, che si ottiene dalla (3.26), osservando che l'evento $|X-\mu| \geq \varepsilon$ è il complementare dell'evento $|X-\mu| < \varepsilon$

$$P(|X-\mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2} \qquad (3.27)$$

Dal punto di vista teorico la caratteristica più importante della disuguaglianza di Chebishev è che si applica ad ogni distribuzione di probabilità di cui siano noti valor medio e varianza μ e σ^2.
Tuttavia questo è anche il suo limite, perché fornisce solo una stima, a volte assai poco precisa, della probabilità di ottenere un valore di X che differisce da μ di una quantità minore o uguale a ε.

Esempio 39
Una variabile aleatoria X ha valor medio $\mu = 3$ e varianza $\sigma^2 = 2$. Mediante la disuguaglianza di Chebishev determinare una maggiorazione per le seguenti probabilità

a – $P(|X-3| \geq 2)$

b – $P(|X-3| \geq 1)$

c – $P(|X-3| \leq 1.5)$

Le tre probabilità che si vogliono stimare sono date dalle aree colorate, rispettivamente nelle figure 27, 28, 29, dove è rappresentata una generica distribuzione di probabilità.
Con la disuguaglianza di Chebishev nella forma (3.26) si ottiene

a – $P(|X-3| \geq 2) \leq \frac{2}{4} = \frac{1}{2}$ (figura 27)

b – $P(|X-3| \geq 1) \leq \frac{2}{1} = 2$ (figura 28)

Quest'ultima stima è priva di interesse, perché troppo grossolana.

c – Con la disuguaglianza di Chebishev nella forma (3.27) si ottiene

$$P(|X-3| \leq 1.5) \geq 1 - \frac{2}{1.5^2} = \frac{1}{9} \quad \text{(figura 29)}$$

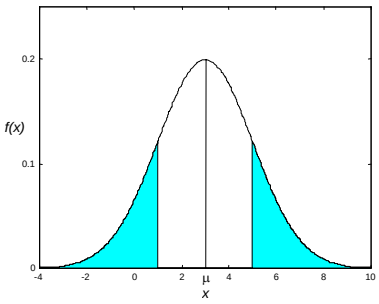


Figura 27

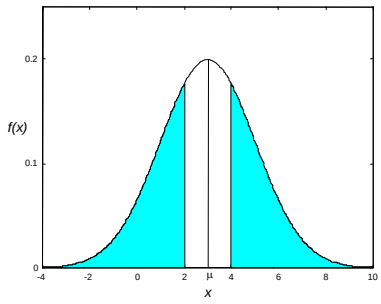


Figura 28

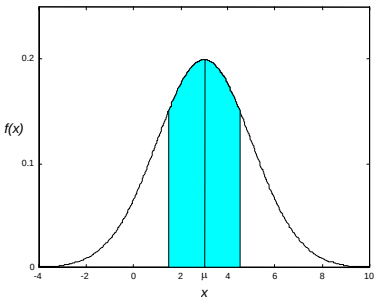


Figura 29

Esempio 40
Il numero di automobili prodotte da una fabbrica in una settimana è una variabile aleatoria X con valor medio $\mu = 500$ e varianza $\sigma^2 = 100$.
Qual è la probabilità che questa settimana la produzione sia compresa fra 400 e 600 automobili?

Per calcolare la probabilità utilizziamo la disuguaglianza di Chebishev (3.27)

$\mu = 500 \qquad \sigma^2 = 100$

$400 \leq X \leq 600 \Rightarrow |X-\mu| = |X-500| \leq 100 \Rightarrow \varepsilon = 100$

$$P\left(|X-500|\le 100\right)\ge 1-\frac{100}{100^2}=0.99$$

Esempio 41

Il numero di clienti che visitano un concessionario di auto al sabato mattina è una variabile aleatoria X con valor medio $\mu = 18$ e deviazione standard $\sigma = 2.5$.
Con quale probabilità si può asserire che il numero di clienti è compreso fra 8 e 28?

Si applica la disuguaglianza di Chebishev (3.27)

$$\mu = 18 \qquad \sigma = 2.5 \qquad 8 \le X \le 28 \Rightarrow \varepsilon = 10$$

$$P\left(|X-18|\le 10\right)\ge 1-\frac{2.5^2}{10^2}=0.9375$$

Esempio 42

Una variabile aleatoria X ha valor medio $\mu = 6$ e deviazione standard $\sigma = \sqrt{2}$; trovare una stima della probabilità che la variabile aleatoria X assuma valori compresi fra 4.5 e 7.5 .

Si applica la disuguaglianza di Chebyshev (3.27)

$$\mu = 6 \quad \sigma = \sqrt{2} \qquad 4.5 \le X \le 7.5 \Rightarrow \varepsilon = 1.5$$

$$P\left(|X-8|\le 1.5\right)\ge 1-\frac{2}{(1.5)^2}=1-\frac{8}{9}=\frac{1}{9}$$

Esercizio 43

Una variabile aleatoria ha densità di probabilità

$$f(x)=\begin{cases}2e^{-2x} & x\ge 0\\ 0 & x<0\end{cases}$$

Sapendo che il valor medio e la varianza valgono $\mu=\frac{1}{2}$ e $\sigma^2=\frac{1}{4}$,

a – calcolare $P\left(|X-\mu|\ge 1\right)$;

b – trovare una stima per $P\left(|X-\mu|\ge 1\right)$ con la disuguaglianza di Chebyshev, e confrontare questa stima con il risultato esatto ottenuto al punto a.

a –
$$P\left(\left|X-\frac{1}{2}\right|\ge 1\right)=1-P\left(\left|X-\frac{1}{2}\right|\le 1\right)$$

$$P\left(\left|X-\frac{1}{2}\right|\le 1\right)=P\left(-\frac{1}{2}\le X\le\frac{3}{2}\right)=\int_0^{\frac{3}{2}}2e^{-2x}dx=-e^{-2x}\Big|_0^{\frac{3}{2}}=1-e^{-3}$$

$$P\left(\left|X-\frac{1}{2}\right|\ge 1\right)=1-(1-e^{-3})=e^{-3}\cong 0.04979$$

b – Con la disuguaglianza di Chebyshev (3.26) si trova

$$P\left(\left|X-\frac{1}{2}\right|\ge 1\right)\le\frac{1}{4}=0.25$$

Il confronto con il risultato esatto trovato al punto a ci permette di concludere che la stima fornita dalla disuguaglianza di Chebyshev è in questo caso molto grossolana. In pratica la disuguaglianza di Chebyshev è usata solo quando non sia nota la densità di probabilità, ma se ne conoscano solo valor medio e varianza.

4. Distribuzioni di probabilità discrete

4.1 Distribuzione binomiale o di Bernoulli

Il concetto di variabile aleatoria permette di formulare modelli utili allo studio di molti fenomeni aleatori. Un primo importante esempio di modello probabilistico è la distribuzione di Bernoulli, così chiamata in onore del matematico svizzero James Bernoulli (1654-1705), che diede importanti contributi nel campo della probabilità..
Alcuni esperimenti consistono nell'eseguire ripetutamente una data prova. Ad esempio vogliamo conoscere la probabilità che 45 su 300 guidatori fermati a un blocco stradale indossino la cintura di sicurezza, oppure la probabilità che 9 su 10 lampadine durino almeno 1000 ore.
In ciascuno di questi esempi si cerca la probabilità di ottenere ***x* successi in *n* prove** o, in altre parole, x successi e $n - x$ insuccessi.
Una sequenza di **prove bernoulliane** costituisce un **processo di Bernoulli** sotto le seguenti ipotesi:
1 – ci sono solo due possibili risultati mutuamente esclusivi per ogni prova, chiamati arbitrariamente "successo" e "insuccesso";
2 – la probabilità di successo p è la stessa per ogni prova;
3 – tutte le prove sono indipendenti; l'indipendenza significa che il risultato di una prova non è influenzato dal risultato di qualunque altra prova; ad esempio, l'evento "alla terza prova si ha successo" è indipendente dall'evento "alla prima prova si ha successo".

Esempio 1
Il lancio di una moneta è una prova bernoulliana: si può considerare successo l'evento "esce testa" e insuccesso l'evento "esce croce". In questo caso la probabilità di successo vale $p = \frac{1}{2}$.
Nel lancio di due dadi si può considerare successo ad esempio l'evento "la somma dei punti è 7" e insuccesso l'evento complementare: in questo caso si tratta di una prova bernoulliana e la probabilità di successo è $p = \frac{1}{6}$.

Sia p la probabilità di successo in una prova bernoulliana.
La variabile aleatoria X che conta il numero di successi in n prove si dice **variabile aleatoria binomiale di parametri *n* e *p***; X può assumere come valori gli interi compresi fra 0 e n.
Si dimostra il seguente risultato[1].

Teorema 1
La probabilità che in n prove la variabile aleatoria X assuma il valore x, ossia che il successo si verifichi x volte in n prove, è data dalla **distribuzione di probabilità binomiale o di Bernoulli**

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{per } x = 0,1,2,\ldots,n \tag{4.1}$$

La **funzione di distribuzione binomiale** è data da

$$F(x) = P(X \le x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \tag{4.2}$$

[1] Si ricordi che

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

(Vedere anche la definizione di combinazioni e coefficienti binomiali, Cap. 2, pag. 66)

La distribuzione binomiale si indica anche con il simbolo

$$b(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x} \qquad \text{per } x=0,1,2,\ldots,n$$

Si osservi che $n-x$ è il numero di insuccessi, e $q=1-p$ la probabilità di insuccesso.
La funzione di distribuzione binomiale si indica anche con il simbolo

$$B(x;n,p)=\sum_{k=0}^{x}\binom{n}{k}p^k(1-p)^{n-k} \ .$$

La media e la varianza di una distribuzione binomiale dipendono solo da *n* e *p*; si dimostra la seguente proprietà.

Proprietà 1
Se *X* è una variabile aleatoria avente distribuzione binomiale con parametri *n* e *p*, allora il **valor medio** è

$$\mu = np \tag{4.3}$$

e la **varianza** è

$$\sigma^2 = np(1-p) \tag{4.4}$$

Nel calcolo della probabilità con la distribuzione binomiale e con la funzione di ripartizione binomiale sono utili le seguenti relazioni.

Proprietà 2

$$P(X<x)=P(X\le x-1) \tag{4.5}$$
$$P(X>x)=1-P(X\le x) \tag{4.6}$$
$$P(X\ge x)=1-P(X\le x-1) \tag{4.7}$$
$$P(X=x)=P(X\le x)-P(X\le x-1) \tag{4.8}$$

Si presti attenzione a non confondere le probabilità $P(X<x)$ e $P(X\le x)$: nel caso delle distribuzioni discrete queste due probabilità non sono uguali.

Esempio 2
Calcolare la probabilità di ottenere 2 volte testa, effettuando 6 lanci di una moneta.

numero prove $n=6$
numero successi $x=2$
probabilità di successo $p=\frac{1}{2}$

$$P(X=2)=\binom{6}{2}\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^{6-2}=\frac{6!}{2!\,4!}\cdot\frac{1}{4}\cdot\frac{1}{16}=\frac{15}{64}\cong 0.2344$$

Esempio 3
Si effettuano 20 lanci di un dado; il successo sia di ottenere 3. Calcolare la probabilità di ottenere 2 volte il caso di successo.

$$n=20 \qquad x=2 \qquad p=\frac{1}{6}$$

$$P(X=2)=\binom{20}{2}\left(\frac{1}{6}\right)^2\left(1-\frac{1}{6}\right)^{20-2}=\binom{20}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{18}\cong 0.1982$$

Esempio 4
Calcolare la probabilità che, effettuando quattro estrazioni con reimbussolamento da un'urna contenente 20 palline bianche e 30 nere, venga estratta per tre volte una pallina bianca.

La probabilità di successo (estrazione di pallina bianca) è

$$p = \frac{20}{50} = \frac{2}{5}$$

$$n = 4 \qquad x = 3$$

$$P(X = 3) = \binom{4}{3}\left(\frac{2}{5}\right)^3\left(1-\frac{2}{5}\right)^1 = 4 \cdot \frac{8}{125} \cdot \frac{3}{5} \cong 0.1536$$

Esempio 5
Si effettuano 10 lanci successivi di una moneta; calcolare la probabilità che per metà delle volte esca croce e per metà testa.

In questo caso si ha

$$n = 10 \qquad x = 5 \qquad p = \frac{1}{2} \qquad 1 - p = \frac{1}{2}$$

$$P(X = 5) = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 = \binom{10}{5}\left(\frac{1}{2}\right)^{10} \cong 0.2461$$

Esempio 6
Calcolare la probabilità che effettuando 6 lanci di due dadi si ottenga la somma 9
a – 2 volte;
b – almeno 2 volte.

Il successo sia di ottenere come somma 9; calcoliamo la probabilità di successo. Servendosi del grafico riprodotto nella figura 9, pag. 97, si deduce facilmente che i casi possibili sono 36 e i casi favorevoli sono 4; questi ultimi sono dati dalle coppie

(3, 6) (4, 5) (5, 4) (6, 3).

Pertanto la probabilità di successo è

$$p = \frac{4}{36} = \frac{1}{9}.$$

a – $n = 6 \qquad x = 2 \qquad p = \frac{1}{9}$

$$P(X = 2) = \binom{6}{2}\left(\frac{1}{9}\right)^2\left(\frac{8}{9}\right)^4 = 0.1156$$

b –
$$P(X \geq 2) = 1 - P(X < 2) = 1 - \left[P(X = 0) + P(X = 1)\right] =$$

$$= 1 - \left[\binom{6}{0}\left(\frac{1}{9}\right)^0\left(\frac{8}{9}\right)^6 + \binom{6}{1}\left(\frac{1}{9}\right)^1\left(\frac{8}{9}\right)^5\right] = 1 - (0.4933 + 0.3700) = 0.1367$$

Esempio 7
La probabilità di laurearsi di uno studente che si iscrive all'Università è $p = 0.4$. Calcolare la probabilità che su 5 studenti
a – nessuno si laurei;
b – uno si laurei;
c – almeno uno si laurei;
d – tutti si laureino.

Il successo è che lo studente si laurei; la variabile aleatoria X indica il numero di laureati.

a – $n = 5 \qquad x = 0 \qquad p = 0.4$

$$P(X = 0) = \binom{5}{0}(0.4)^0(0.6)^5 = 0.07776$$

b – $$P(X = 1) = \binom{5}{1}(0.4)^1(0.6)^4 = 0.2592$$

c – $$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - 0.07776 = 0.9222$$

d – $$P(X = 5) = \binom{5}{5}(0.4)^5(0.6)^0 = 0.01024$$

Esempio 8

La ditta produttrice sostiene che nel 60% degli impianti a pannelli solari installati si è verificata una riduzione di un terzo del costo della fattura dell'energia elettrica. Calcolare la probabilità che questa riduzione si verifichi

a – in 4 su 5 installazioni;

b – in almeno 4 installazioni.

a – $n = 5 \qquad x = 4 \qquad p = 0.60$

$$P(X = 4) = \binom{5}{4}(0.60)^4(1 - 0.60)^{5-4} = 0.2592$$

b – $n = 5 \qquad x = 5 \qquad p = 0.60$

$$P(X = 5) = \binom{5}{5}(0.60)^5(1 - 0.60)^{5-5} = 0.07776$$

$$P(X \geq 4) = P(X = 4) + P(X = 5) = 0.2592 + 0.07776 = 0.3370$$

Esempio 9

Un test è costituito da 10 domande a risposta multipla: ci sono 4 risposte possibili per ogni domanda, di cui una sola esatta. Per superare il test occorre rispondere esattamente ad almeno 8 domande. Rispondendo a caso alle domande, qual è la probabilità di superare il test?

La variabile aleatoria X indica il numero delle risposte esatte.

$$n = 10 \qquad x = 8 \qquad p = \frac{1}{4}$$

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) =$$

$$= \binom{10}{8} \cdot \left(\frac{1}{4}\right)^8 \left(1 - \frac{1}{4}\right)^2 + \binom{10}{9} \cdot \left(\frac{1}{4}\right)^9 \left(1 - \frac{1}{4}\right)^1 + \binom{10}{10} \cdot \left(\frac{1}{4}\right)^{10} \left(1 - \frac{1}{4}\right)^0$$

$$= 45 \cdot \left(\frac{1}{4}\right)^8 \left(\frac{3}{4}\right)^2 + 10 \cdot \left(\frac{1}{4}\right)^9 \left(\frac{3}{4}\right) + \left(\frac{1}{4}\right)^{10} = 0.0004158 \cong 0.04\%$$

Esempio 10

La probabilità che un apparecchio subisca un certo tipo di guasto è $p = 0.05$; calcolare la probabilità che su 16 di tali apparecchi

a – al più 2 si guastino;

b – almeno 2 si guastino;

c – almeno 4 si guastino.

La variabile aleatoria X indica il numero dei guasti.

$$n = 16 \qquad p = 0.05 \qquad 1 - p = 0.95$$

a – $P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$$P(X = 0) = \binom{16}{0}(0.05)^0(0.95)^{16} = 0.4401$$

$$P(X = 1) = \binom{16}{1}(0.05)^1(0.95)^{15} = 0.3706$$

$$P(X = 2) = \binom{16}{2}(0.05)^2(0.95)^{14} = 0.1463$$

$$P(X \le 2) = 0.4401 + 0.3706 + 0.1463 = 0.9570$$

b – $P(X \ge 2) = 1 - P(X < 2) = 1 - \left[P(X = 0) + P(X = 1)\right] =$

$$= 1 - 0.4401 - 0.3706 = 0.1893$$

c – $P(X \ge 4) = 1 - P(X < 4)$

$$P(X < 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$P(X = 3) = \binom{16}{3}(0.05)^3(0.95)^{13} = 0.0359$$

$$P(X \ge 4) = 1 - (0.4401 + 0.3706 + 0.1463 + 0.0359) = 0.0071$$

Esempio 11

Determinare la probabilità che lanciando 3 volte una moneta si verifichi
a – 3 volte T;
b – 2 volte C e una volta T;
c – almeno una volta T;
d – al più una volta C.

La variabile aleatoria X indica il numero di teste.

$$n = 3 \qquad p = 0.5$$

a – $P(X = 3) = \binom{3}{3}(0.5)^3(0.5)^0 = \frac{1}{8}$

b – $P(X = 1) = \binom{3}{1}(0.5)^1(0.5)^2 = \frac{3}{8}$

c – $P(X \ge 1) = P(X = 1) + P(X = 2) + P(X = 3) =$

$$= \frac{3}{8} + \binom{3}{2}(0.5)^2(0.5)^1 + \frac{1}{8} = \frac{7}{8}$$

d – $P(\text{al più 1 C}) = P(\text{nessuna C}) + P(1\,\text{C}) =$

$$= P(X = 3) + P(X = 2) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

Esempio 12

Determinare la probabilità che in 5 lanci di un dado il numero 3 esca
a – 2 volte;
b – al più una volta;
c – almeno 2 volte.

La variabile aleatoria indica il numero di volte che esce 3.

$$n = 5 \qquad p = \frac{1}{6}$$

a – $P(X=2)=\binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 \cong 0.1608$

b – $P(X\le 1)=P(X=0)+P(X=1)=$

$$=\binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5+\binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 \cong 0.8038$$

c – $P(X\ge 2)=1-P(X<2)=1-P(X\le 1)=1-0.8038=0.1962$

Esempio 13

Determinare la probabilità che in una famiglia con 4 figli ci sia

a – almeno un maschio;

b – almeno un maschio e una femmina.

c – Su 2000 famiglie con 4 figli ciascuna, quante famiglie hanno in media almeno un figlio maschio? E quante famiglie hanno in media due maschi?

Si supponga che le probabilità di nascita di un maschio e di una femmina siano uguali.

La variabile aleatoria X indica il numero dei maschi e p è la probabilità di nascita di un maschio.

$n = 4$ $p = 0.5$

a – $P(X\ge 1)=P(X=1)+P(X=2)+P(X=3)+P(X=4)$

$$P(X=1)=\binom{4}{1}(0.5)^1(0.5)^3=\frac{1}{4} \qquad P(X=2)=\binom{4}{2}(0.5)^2(0.5)^2=\frac{3}{8}$$

$$P(X=3)=\binom{4}{3}(0.5)^3(0.5)^1=\frac{1}{4} \qquad P(X=4)=\binom{4}{4}(0.5)^4(0.5)^0=\frac{1}{16}$$

$$P(X\ge 1)=\frac{1}{4}+\frac{3}{8}+\frac{1}{4}+\frac{1}{16}=\frac{15}{16}$$

La probabilità $P(X\ge 1)$ può anche essere calcolata più brevemente come segue

$$P(X\ge 1)=1-P(X<1)=1-P(X=0)=1-\binom{4}{0}(0.5)^0(0.5)^4=1-\frac{1}{16}=\frac{15}{16}$$

b – P(almeno un M e una F) = 1–[P(nessun M)+P(nessuna F)]

$$P(\text{almeno un M e una F})=1-\left[P(X=0)+P(X=4)\right]=1-\frac{1}{16}-\frac{1}{16}=\frac{7}{8}$$

c – Ricordiamo i risultati trovati al punto a

$$P(X\ge 1)=\frac{15}{16} \qquad P(X=2)=\frac{3}{8}.$$

Il numero medio di famiglie con almeno un maschio è

$$N_1 = 2000\cdot\frac{15}{16}=1875$$

Il numero medio di famiglie con due maschi è

$$N_2 = 2000\cdot\frac{3}{8}=750$$

Esempio 14

Se il 5% dei chip di memoria prodotti da una macchina sono difettosi, determinare la probabilità che su 4 chip scelti a caso

a – 1 sia difettoso;

b – nessuno sia difettoso;

c – meno di 2 siano difettosi.

Calcolare la media e la deviazione standard del numero di chip difettosi su un totale di 400 chip.

La variabile aleatoria X indica il numero di chip difettosi.

$n = 4 \qquad p = 0.05$

a – $P(X = 1) = \binom{4}{1}(0.05)^1(0.95)^3 = 0.1715$

b – $P(X = 0) = \binom{4}{0}(0.05)^0(0.95)^4 = 0.8145$

c – $P(X < 2) = P(X = 0) + P(X = 1) = 0.1715 + 0.8145 = 0.9860$

d – $n = 400 \qquad p = 0.05$

$\mu = np = 400 \cdot 0.05 = 20$

$\sigma^2 = np(1 - p) = 400 \cdot 0.05 \cdot 0.95 = 19$

$\sigma = \sqrt{19} \cong 4.36$

Esempio 15

Data una distribuzione binomiale con $n = 9$ e $\sigma = 0.9$, ricavare i possibili valori di p; per ciascun valore di p calcolare $P(X = 4)$.

Per la proprietà 1, si ha

$$\sigma^2 = np(1-p) = 0.81 = \frac{81}{100}$$

$$9p(1-p) = \frac{81}{100}$$

$$100p^2 - 100p + 9 = 0$$

I possibili valori di p sono

$p = 0.1 \qquad p = 0.9$.

Per $n = 9$ e $p = 0.1$ si ha

$$P(X = 4) = \binom{9}{4}(0.1)^4(0.9)^5 = 0.007440$$

Per $n = 9$ e $p = 0.9$ si ha

$$P(X = 4) = \binom{9}{4}(0.9)^4(0.1)^5 = 0.0008267$$

Esempio 16

La variabile aleatoria X ha distribuzione binomiale ed è tale che

$$\frac{\sigma^2}{\mu} = 0.3 \qquad \mu = 10.5$$

Trovare i valori di n e p.

Per la proprietà 1 si ha

$$\sigma^2 = np(1-p) = 0.3 \cdot 10.5 = 3.15$$

$$\mu = np = 10.5$$

Risolvendo il sistema

$$\begin{cases} np(1-p) = 3.15 \\ np = 10.5 \end{cases}$$

si trova

$p = 0.7 \qquad n = 15$.

4.2 Uso delle tavole della distribuzione binomiale

Il calcolo dei valori della distribuzione binomiale può essere lungo, specialmente per valori di n non piccoli; in tali casi, se non si dispone di un opportuno software statistico, si possono usare delle tavole di approssimazione numerica che agevolano il calcolo.
Sono disponibili delle **tavole della funzione di distribuzione $B(x; n, p)$**, per valori di n da 2 a 20 e per $p = 0.05, 0.10, 0.15, ...,0.95$, che riportiamo nell'Appendice A.
Sono state tabulate le funzioni di distribuzione $B(x; n, p)$, anziché le distribuzioni di probabilità $b(x; n, p)$, perché sono più frequentemente utilizzate nelle applicazioni statistiche.
Per l'uso delle tavole sono utili le relazioni (4.5), (4.6), (4.7), (4.8), elencate nella proprietà 2. In particolare per ricavare i valori di $b(x; n, p)$ si utilizza la relazione (4.8), che può anche essere scritta nella seguente forma

$$b(x;n,p) = B(x;n,p) - B(x-1;n,p)$$

Le tavole non possono fornire i valori della funzione di distribuzione per ogni combinazione di valori di n e p (i motivi sono evidenti); se il valore di p non è reperibile sulle tavole, è preferibile calcolare la probabilità direttamente con la formula, anziché ricorrere ad un'approssimazione (ottenibile interpolando sulle tavole), perché il valore approssimato potrebbe essere poco accurato. In casi di questo tipo può essere utile la relazione di ricorrenza che verrà trattata nel § 4.3.

Esempio 17
Riprendiamo l'esempio 10; con l'uso delle tavole si ottiene

a – $P(X \le 2) = B(2;16, 0.05) = 0.9571$

b – $P(X \ge 2) = 1 - P(X < 2) = 1 - P(X \le 1) =$
$= 1 - B(1;16,0.05) = 1 - 0.8108 = 0.1892$

c – $P(X \ge 4) = 1 - P(X < 4)$
$P(X < 4) = P(X \le 3) = B(3;16, 0.05) = 0.9930$
$P(X \ge 4) = 1 - 0.9930 = 0.0070$

d – $P(X = 3) = b(3;16, 0.05) = B(3;16, 0.05) - B(2;16, 0.05) =$
$= 0.9930 - 0.9571 = 0.0359$

Esempio 18
Se la probabilità che una persona non gradisca il gusto di un nuovo dentifricio è $p = 0.20$, qual è la probabilità che 5 su 18 persone scelte a caso non lo gradiscano?

La variabile X indica il numero di persone che non gradiscono il nuovo gusto. Con l'uso delle tavole si ottiene

$$n = 18 \qquad x = 5 \qquad p = 0.20$$
$$P(X = 5) = B(5;18, 0.20) - B(4;18, 0.20) = 0.8671 - 0.7164 = 0.1507$$

Esempio 19
Supponiamo che il 75% degli incidenti sul lavoro in un'azienda possano essere evitati con il rigoroso rispetto delle norme di sicurezza; trovare le probabilità che possano essere evitati
a – meno di 16 incidenti su 20;
b – 12 incidenti su 15.

La variabile aleatoria X indica il numero di incidenti. Utilizzando le tavole si ottiene

a – $n = 20 \qquad p = 0.75$
$P(X \le 15) = B(15; 20, 0.75) = 0.5852$

b – $n = 15 \qquad p = 0.75$
$P(X = 12) = b(12;15, 0.75) = B(12;15, 0.75) - B(11;15, 0.75) =$
$= 0.7639 - 0.5387 = 0.2252$

Esempio 20
Una variabile aleatoria X ha distribuzione binomiale con media $\mu = 14$ e varianza $\sigma^2 = 4.2$; calcolare la probabilità $P(X \geq 13)$.

Per la proprietà 1 si ha

$$\begin{cases} \mu = np \\ \sigma^2 = np(1-p) \end{cases} \qquad \begin{cases} np = 14 \\ np(1-p) = 4.2 \end{cases}$$

Risolvendo questo sistema si trova

$$n = 20 \qquad p = 0.7$$

Sulle tavole si trova

$$P(X \geq 13) = 1 - P(X \leq 12) = 1 - 0.2277 = 0.7723$$

4.3 Relazione di ricorrenza per la distribuzione binomiale

In certi casi si devono calcolare valori della distribuzione binomiale per $n > 20$ e/o per valori di p che non compaiono sulle tavole. Per valori di n sufficientemente grandi si può usare la distribuzione normale per approssimare la distribuzione binomiale, come vedremo nel capitolo 5; in tal caso si usano le tavole della distribuzione normale e questo modo di procedere è più veloce e meno noioso del calcolo delle probabilità direttamente con la distribuzione binomiale.
Se però il valore di n non è sufficientemente grande per poter usare l'approssimazione con la distribuzione normale, e se il valore di p non compare sulle tavole, allora si può usare una relazione di ricorrenza che agevola i calcoli. Questa relazione è particolarmente utile se si devono calcolare molti valori delle probabilità con la distribuzione binomiale per gli stessi valori di n e p.
Si dimostra che vale la relazione seguente.

Proprietà 3 – Relazione di ricorrenza per la binomiale

$$P(X = x+1) = \frac{n-x}{x+1} \cdot \frac{p}{1-p} \cdot P(X = x) \tag{4.9}$$

Usando questa relazione, dopo aver calcolato $P(X = 0)$, le probabilità $P(X = 1)$, $P(X = 2)$,… possono essere facilmente ottenute senza dover fare lunghi calcoli coinvolgenti i coefficienti binomiali (vedere esempi 21 e 22).

4.4 Rappresentazione grafica della distribuzione binomiale

La distribuzione binomiale viene rappresentata graficamente per mezzo di un istogramma o di un diagramma a barre. La forma della distribuzione dipende dal valore della probabilità di successo p.

Nel caso $p = \frac{1}{2}$, è anche $1 - p = \frac{1}{2}$: ciò significa che il successo e l'insuccesso sono ugualmente probabili; da questo segue che la probabilità di avere ad esempio 2 successi (e quindi $n - 2$ insuccessi) è uguale alla probabilità di avere $n - 2$ successi (e quindi 2 insuccessi): l'istogramma della distribuzione è quindi simmetrico (figura 1, pagina seguente).

Se invece $p < \frac{1}{2}$ oppure $p > \frac{1}{2}$, l'istogramma è asimmetrico; nel primo caso l'asimmetria è positiva, la distribuzione è obliqua a destra (figura 2), nel secondo caso l'asimmetria è negativa, la distribuzione è obliqua a sinistra (figura 3).

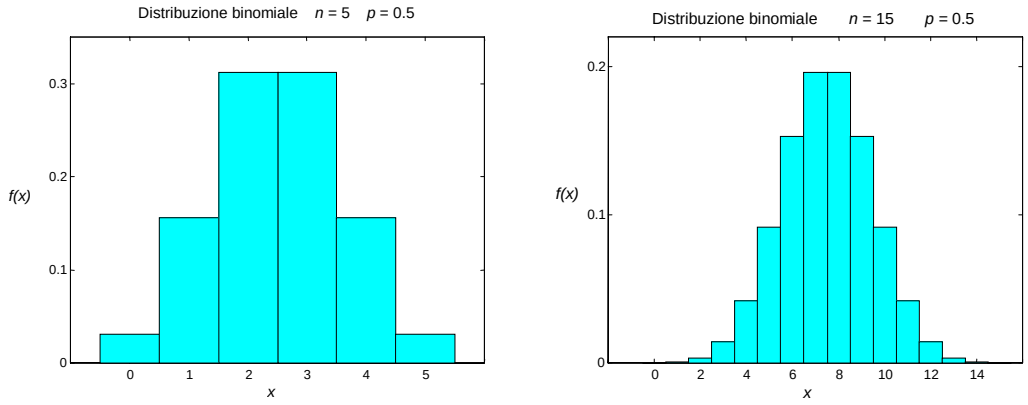


Figura 1

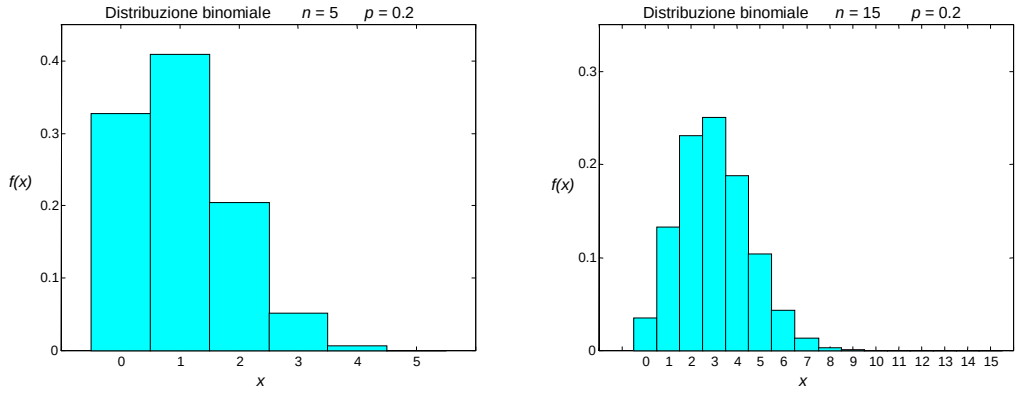


Figura 2

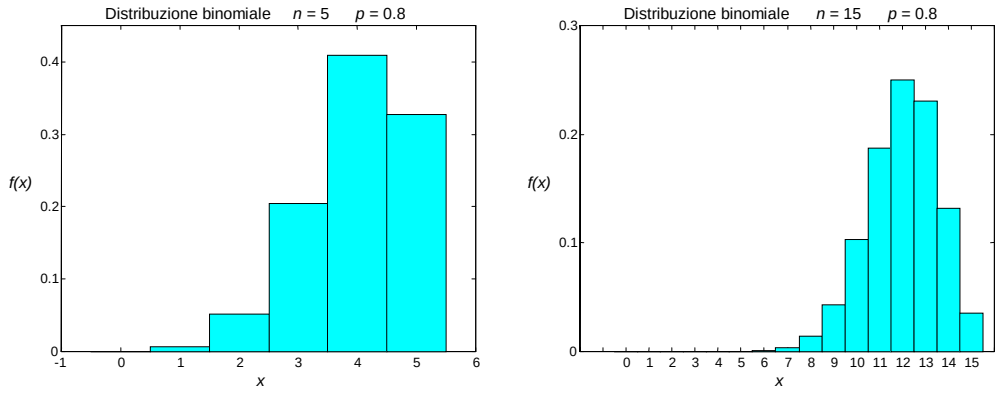


Figura 3

Esempio 21

Si effettuano 6 lanci di una moneta; studiare la distribuzione di probabilità della variabile aleatoria binomiale X = numero di teste T uscite nei 6 lanci.

Il successo è dato dall'uscita T e la probabilità di successo è $p = \frac{1}{2}$.

Calcoliamo con la formula della distribuzione binomiale la probabilità di ottenere 0 volte l'uscita T

$$P(X = 0) = \binom{6}{0}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^0 = 0.015625 \cong 0.015625$$

Applicando la formula di ricorrenza (4.9) si calcolano gli altri valori delle probabilità.

$$P(X=1)=\frac{6}{1}\cdot\frac{0.5}{0.5}\cdot P(X=0)=6\cdot 0.015625=0.09375$$

$$P(X=2)=\frac{5}{2}\cdot 0.09375=0.2344$$

$$P(X=3)=\frac{4}{3}\cdot 0.2344=0.3125$$

$$P(X=4)=0.2344 \qquad P(X=5)=0.09375 \qquad P(X=6)=0.015625$$

Questi valori potrebbero essere cercati direttamente sulle tavole, dove compare sia il valore $n = 6$ che $p=\frac{1}{2}$.

Il grafico della distribuzione di probabilità è rappresentato dal seguente istogramma (figura 4); si noti la simmetria, dovuta al fatto che $p=\frac{1}{2}$. Data la simmetria, non è necessario ripetere il calcolo degli ultimi tre valori delle probabilità $P(X=4),\ P(X=5),\ P(X=6)$, che sono rispettivamente uguali a quelli già calcolati $P(X=2),\ P(X=1),\ P(X=0)$.

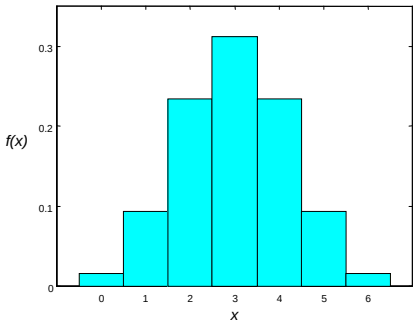


Figura 4

Esempio 22

Si effettuano 10 lanci di un dado.

a – Studiare la distribuzione di probabilità della variabile aleatoria binomiale X = numero di uscite del numero 3.

b – Studiare la distribuzione di probabilità della variabile aleatoria binomiale X = numero di uscite di un numero diverso da 3.

a – Il successo è dato dall'uscita del numero 3 e la probabilità di successo è $p = \frac{1}{6}$ (questo valore non compare sulle tavole).

Calcoliamo con la formula della distribuzione binomiale la probabilità di ottenere 0 volte l'uscita del numero 3, e ricaviamo gli altri valori delle probabilità con la formula di ricorrenza (4.9).

$$P(X=0)=\binom{10}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{10}=0.1615$$

$$P(X=1)=10\cdot\frac{1}{6}\cdot\frac{6}{5}0.1615=0.3230$$

$$P(X=2)=\frac{9}{2}\cdot\frac{1}{5}\cdot 0.3230=0.2907$$

$$P(X=3)=\frac{8}{3}\cdot\frac{1}{5}\cdot 0.2907=0.1550$$

$$P(X=4)=\frac{7}{4}\cdot\frac{1}{5}\cdot 0.1550=0.05425$$

.....

Il grafico della distribuzione di probabilità è rappresentato dal seguente istogramma; si noti l'asimmetria positiva del grafico, la distribuzione è obliqua verso destra.

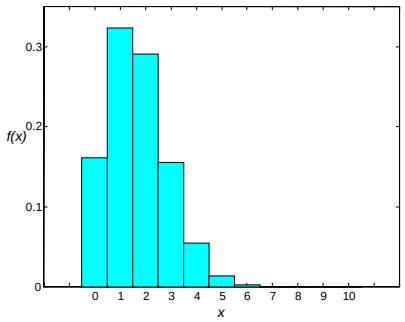


Figura 5

b – Le probabilità di ottenere un numero diverso da 3 si ricavano per simmetria dai valori ottenuti al punto a: infatti in questo caso il successo, l'uscita di un numero diverso da 3, coincide con l'insuccesso del caso precedente, l'uscita del numero 3; quindi si ha

.....

$P(X = 6) = 0.05425$

$P(X = 7) = 0.1550$

$P(X = 8) = 0.2907$

$P(X = 9) = 0.3230$

$P(X = 10) = 0.1615$

Il grafico della distribuzione di probabilità è rappresentato dal seguente istogramma; si noti l'asimmetria negativa del grafico, la distribuzione è obliqua verso sinistra.

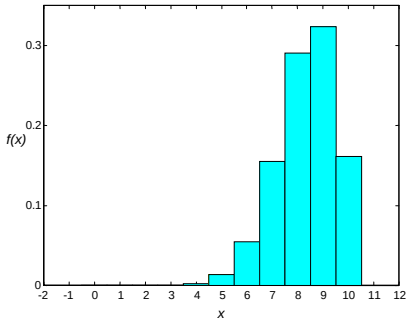


Figura 6

4.5 Distribuzione di Poisson

Vi sono fenomeni in cui determinati eventi, con riferimento a un certo intervallo di tempo o di spazio, accadono raramente: il numero di eventi che si verificano in quell'intervallo varia da 0 a *n*, e *n* non è determinabile a priori. Ad esempio, il numero di automobili che transitano in una strada poco frequentata in un intervallo di tempo di 5 minuti scelto a caso, può essere considerato un evento raro; analogamente sono eventi rari il numero di infortuni sul lavoro che accadono in una azienda in una settimana o il numero di errori di stampa presenti in una pagina di un libro.
Nello studio degli eventi rari, come quelli degli esempi citati, è fondamentale il riferimento a uno specifico intervallo di tempo o di spazio.
Per lo studio di eventi rari del tipo di quelli descritti si utilizza la **distribuzione di probabilità di Poisson**, così chiamata in onore del matematico francese Simeon Denis Poisson (1781-1840), che per primo ricavò la distribuzione; questa distribuzione è molto usata come modello di probabilità in

biologia e medicina. La distribuzione di Poisson è usata come modello nei casi in cui gli eventi o realizzazioni di un processo, distribuiti a caso nello spazio o nel tempo, sono dei conteggi, ovvero delle variabili discrete.
La distribuzione binomiale è basata su un insieme di ipotesi che definiscono le prove bernoulliane; lo stesso accade per la distribuzione di Poisson.
Le seguenti condizioni descrivono il così detto **processo di Poisson**:
1 – le realizzazioni degli eventi sono indipendenti: il verificarsi di un evento in un intervallo di tempo o di spazio non ha alcun effetto sulla probabilità di verificarsi dell'evento una seconda volta nello stesso, o in un altro, intervallo;
2 – la probabilità di una singola realizzazione dell'evento in un dato intervallo è proporzionale alla lunghezza dell'intervallo;
3 – in ogni parte arbitrariamente piccola dell'intervallo, la probabilità che l'evento si verifichi più di una volta è trascurabile.
Sia X la variabile aleatoria che indica il numero di volte in cui si verifica un evento raro in un dato intervallo di tempo o di spazio, ossia il numero di successi; la variabile X può assumere i valori $x = 0,1,2,\ldots$ Si dimostra il seguente risultato.

Teorema 2
La probabilità che la variabile aleatoria X assuma il valore x è data dalla **distribuzione di probabilità di Poisson**

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{per } x = 0, 1, 2, \ldots \tag{4.10}$$

dove il parametro $\lambda > 0$ indica il numero medio di realizzazioni dell'evento nell'intervallo assegnato.

Una variabile aleatoria che ammette questa distribuzione è detta **variabile aleatoria di Poisson** con parametro λ.
La distribuzione di Poisson viene anche indicata con il simbolo $f(x;\lambda)$; la corrispondente **funzione di distribuzione di Poisson** è data da

$$F(x) = P(X \le x) = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!}$$

e viene anche indicata con il simbolo $F(x;\lambda)$.
Si dimostra la seguente proprietà.

Proprietà 4
Il **valor medio** e la **varianza** della distribuzione di Poisson di parametro λ sono dati da

$$\mu = \lambda \qquad \sigma^2 = \lambda \tag{4.11}$$

Una importante differenza tra la distribuzione di Poisson e la binomiale riguarda i numeri di prove e di successi: per una distribuzione binomiale il numero n di prove è finito e il numero x di successi non può superare n; per una distribuzione di Poisson il numero di prove è essenzialmente infinito e il numero di successi può essere infinitamente grande, anche se la probabilità di avere x successi diventa molto piccola al crescere di x.
Per il calcolo della distribuzione di Poisson sono utili le relazioni elencate nella proprietà 2, valide anche per questa distribuzione discreta.
La distribuzione di Poisson ha molte applicazioni in vari ambiti diversi, perché può essere usata per approssimare una distribuzione binomiale di parametri n e p, quando il numero di prove n è grande e la probabilità di successo p è piccola, ossia si tratta di un evento raro.
Per dimostrare questo, indichiamo con X una variabile aleatoria avente distribuzione binomiale con parametri n e p, con n grande e p piccola, e sia $\lambda = np$; si ha

$$b(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x}=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}=$$

$$=\frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}=\frac{n(n-1)(n-2)...(n-x+1)}{x!\,n^x}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}=$$

$$=\frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)....\left(1-\frac{x-1}{n}\right)}{x!}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Per $n\to\infty$, si ha

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)....\left(1-\frac{x-1}{n}\right)=1$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-x}=e^{-\lambda}$$

quindi

$$\lim_{n\to\infty}b(x;n,p)=\frac{\lambda^x e^{-\lambda}}{x!}$$

ossia la distribuzione di Poisson è il limite per $n\to\infty$, e con $\lambda=np$, della distribuzione binomiale di parametri *n* e *p*.
Da questo segue che, quando il numero di prove *n* è grande e la probabilità di successo *p* è piccola, la distribuzione binomiale può essere approssimata con la distribuzione di Poisson avente media $\lambda=np$ (vedere § 4.9).

Esempio 23
Dalle statistiche degli ultimi 5 anni, un'azienda ha calcolato che ogni giorno sono assenti in media 1.8 operai. Calcolare la probabilità che in un giorno qualsiasi ci siano 3 operai assenti contemporaneamente.

Il numero medio di assenti giornalieri è piccolo, perciò si può usare la distribuzione di Poisson con parametro $\lambda = 1.8$; si trova

$$P(X=3)=\frac{e^{-1.8}(1.8)^3}{3!}=0.1607$$

Esempio 24
Ad un servizio di guardia medica arrivano in media 3.5 richieste ogni ora di interventi urgenti a domicilio.
a – Calcolare la probabilità che in una stessa ora arrivino 3, 4, 5 chiamate urgenti.
b – Calcolare la probabilità che in una stessa ora arrivi un numero di chiamate compreso fra 3 e 5.
c – Calcolare la probabilità che in una stessa ora arrivi un numero di chiamate maggiore di 4.

a – Le probabilità possono essere calcolate con la distribuzione di Poisson, con parametro $\lambda = 3.5$; si ha

$$P(X=3)=\frac{e^{-3.5}(3.5)^3}{3!}=0.2158$$

$$P(X=4)=\frac{e^{-3.5}(3.5)^4}{4!}=0.1888$$

$$P(X=5)=\frac{e^{-3.5}(3.5)^5}{5!}=0.1322$$

b – $$P(3 \le X \le 5) = P(X = 3) + P(X = 4) + P(X = 5) = \\ = 0.2158 + 0.1888 + 0.1322 = 0.5368$$

c – $$P(X > 3) = 1 - P(X \le 3) = 1 - \left[P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\right] = \\ = 1 - \left(e^{-3.5} + e^{-3.5} \cdot 3.5 + \frac{e^{-3.5}(3.5)^2}{2} + \frac{e^{-3.5}(3.5)^3}{3!} \right) = \\ = 1 - \left(0.03020 + 0.1057 + 0.1850 + 0.2158\right) = 0.4633$$

Esempio 25
Un libro di 500 pagine contiene 50 errori di stampa. Qual è la probabilità di trovare almeno 3 errori su una pagina aperta a caso?

Il numero medio di errori su una pagina è $\lambda = \frac{50}{500} = 0.1$; con la distribuzione di Poisson si ha

$$P(X \ge 3) = 1 - P(X \le 2) = 1 - \left[P(X = 0) + P(X = 1) + P(X = 2)\right] = \\ = 1 - \left(e^{-0.1} + 0.1 \cdot e^{-0.1} + \frac{0.1^2}{2} e^{-0.1} \right) = 1 - 0.99985 = 0.00015$$

4.6 Uso delle tavole della distribuzione di Poisson

Poiché la distribuzione di Poisson ha molte importanti applicazioni, sono disponibili delle tavole, riportate nell'Appendice A, che forniscono il valore della funzione di distribuzione

$$F(x;\lambda) = P(X \le x)$$

per vari valori di λ, variabili fra 0.02 e 25.
Per il calcolo della distribuzione di probabilità $f(x;\lambda)$ con l'uso delle tavole, è utile l'identità

$$f(x;\lambda) = F(x;\lambda) - F(x-1;\lambda)$$

(si ricordi la proprietà (4.8)).

Esempio 26
La variabile aleatoria X ha la distribuzione di probabilità di Poisson con valor medio $\lambda = 2$. Calcolare le probabilità

a – $P(4 < X < 7)$

b – $P(3 < X \le 7)$

c – $P(X > 3)$

d – $P(X = 5)$

Con l'uso delle tavole si ha

a – $P(4 < X < 7) = P(X \le 6) - P(X \le 4) = 0.9955 - 0.9473 = 0.0482$

b – $P(3 < X \le 7) = P(X \le 7) - P(X \le 3) = 0.9989 - 0.8571 = 0.1418$

c – $P(X > 3) = 1 - P(X \le 3) = 1 - 0.8571 = 0.1429$

d – $P(X = 5) = P(X \le 5) - P(X \le 4) = 0.9834 - 0.9473 = 0.0361$

Esempio 27
Data la variabile aleatoria X avente distribuzione di Poisson, trovare il valor medio λ, sapendo che

a – $P(X \le 5) = 0.9896$

b – $P(X > 4) = 0.0527$.

Leggendo le tavole in modo contrario si trova

a – $P(X \le 5) = 0.9896 \Rightarrow \lambda = 1.8$

b – $P(X > 4) = 0.0527 = 1 - P(X \le 4)$

$P(X \le 4) = 1 - 0.0527 = 0.9473 \Rightarrow \lambda = 2.$

4.7 Relazione di ricorrenza per la distribuzione di Poisson

In alcuni casi è richiesto di calcolare più valori della distribuzione di Poisson per lo stesso valor medio $\mu = \lambda$ non presente sulle tavole. Se λ è grande ($\lambda \ge 10$), la distribuzione di Poisson può essere approssimata dalla distribuzione normale, come si vedrà nel capitolo 5; altrimenti può essere utile la seguente relazione di ricorrenza, simile a quella valida per la distribuzione binomiale.

Proprietà 5 – Relazione di ricorrenza per la distribuzione di Poisson

$$P(X = x+1) = \frac{\lambda}{x+1} P(X = x) \qquad (4.12)$$

Con questa relazione, partendo da $P(X = 0) = e^{-\lambda}$, si possono calcolare successivamente le probabilità $P(X = 1)$, $P(X = 2)$,….

Esempio 28

La variabile aleatoria X ha la distribuzione di probabilità di Poisson con valor medio $\lambda = 3.5$. Calcolare $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, $P(X = 3)$, $P(X = 4)$, $P(X = 5)$ ...

Usando la relazione di ricorrenza si ha

$$P(X = 0) = e^{-3.5} = 0.0302$$

$$P(X = 1) = 3.5 \cdot P(X = 0) = 3.5 \cdot 0.0302 = 0.1057$$

$$P(X = 2) = \frac{3.5}{2} P(X = 1) = \frac{3.5}{2} \cdot 0.1057 = 0.1850$$

$$P(X = 3) = \frac{3.5}{3} P(X = 2) = \frac{3.5}{3} \cdot 0.1850 = 0.2158$$

$$P(X = 4) = \frac{3.5}{4} P(X = 3) = \frac{3.5}{4} \cdot 0.2158 = 0.1888$$

$$P(X = 5) = \frac{3.5}{5} P(X = 4) = \frac{3.5}{5} \cdot 0.1888 = 0.1322$$

.....

4.8 Rappresentazione grafica della distribuzione di Poisson

Anche la distribuzione di Poisson viene rappresentata graficamente con un istogramma o con un diagramma a barre. Al crescere di λ il grafico presenta un aspetto maggiormente simmetrico, come si può osservare dai grafici della figura 7, pag. seguente, dove sono rappresentate alcune distribuzioni di Poisson per valori crescenti di λ; si noti che i diagrammi sono troncati dopo un opportuno valore di x perché, anche se la variabile X può assumere valori maggiori, le corrispondenti probabilità sono molto basse.

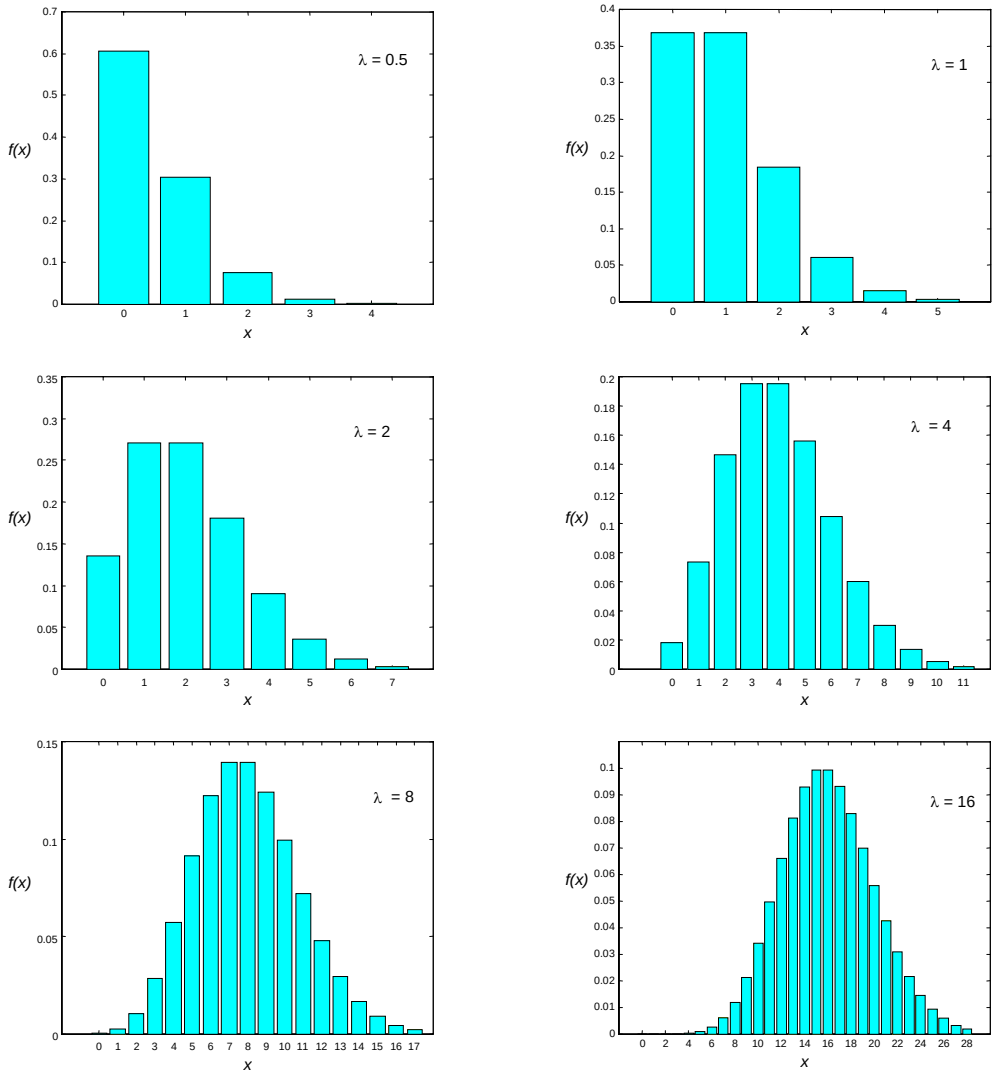


Figura 7

Esempio 29

La probabilità che un oggetto prodotto da una macchina sia difettoso è p = 0.15; calcolare le probabilità che in un campione di 10 oggetti scelti a caso, ci siano 0, 1, 2, ...,10 oggetti difettosi usando la distribuzione binomiale e la distribuzione di Poisson, e confrontare su un grafico i risultati ottenuti.

Con l'uso delle tavole si ottengono i seguenti valori delle probabilità.

a – Distribuzione binomiale

$$n = 10 \qquad p(\text{difettoso}) = 0.15$$

$$P(X = 0) = 0.1969$$

$$P(X = 1) = 0.5443 - 0.1969 = 0.3474$$

$$P(X = 2) = 0.8202 - 0.5443 = 0.2759$$

$$P(X = 3) = 0.9500 - 0.8202 = 0.1298$$

.....

b – Distribuzione di Poisson

$$n = 10 \qquad p(\text{difettoso}) = 0.15 \qquad \lambda = np = 1.5$$

$$P(X = 0) = 0.2231$$

$$P(X = 1) = 0.5578 - 0.2231 = 0.3347$$
$$P(X = 2) = 0.8088 - 0.5578 = 0.2510$$
$$P(X = 3) = 0.9344 - 0.8088 = 0.1256$$
.....

I risultati sono posti a confronto nella figura 8

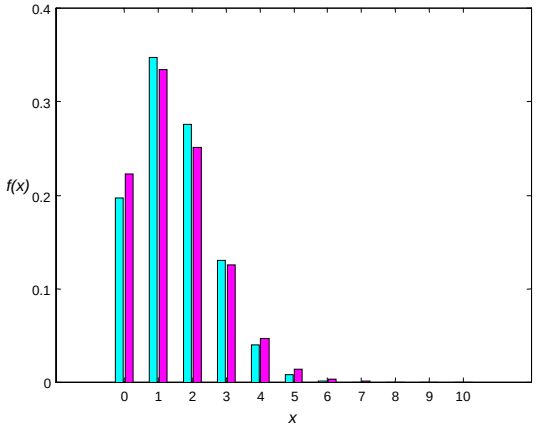


Figura 8

Sia dal confronto dei risultati numerici che dai grafici si osserva che la distribuzione di Poisson approssima in modo non troppo preciso i valori trovati con la binomiale; ciò è dovuto al fatto che i valori di *n* e *p* non soddisfano la regola pratica suggerita per usare tale approssimazione con risultati soddisfacenti.

4.9 Approssimazione della distribuzione binomiale con la distribuzione di Poisson

Come già detto (§ 4.5) , quando il numero di prove *n* è grande e la probabilità di successo *p* è piccola, la distribuzione binomiale può essere approssimata con la distribuzione di Poisson avente media $\lambda = np$.

Una **regola pratica** accettabile è di usare questa approssimazione se $n \geq 50$ e $p \leq 0.1$. La regola comunque non è rigida: si può dire che più è piccola la probabilità *p*, migliore è l'approssimazione, e analogamente più è grande *n*, migliore è l'approssimazione (vedere esempio 33).
Gli esempi che seguono illustrano l'uso della distribuzione di Poisson per approssimare la distribuzione binomiale.

Esempio 30
Se il 3% delle lampadine costruite da una fabbrica sono difettose, trovare la probabilità che in un campione di 100 lampadine 2 siano difettose usando
a – la distribuzione binomiale;
b – la distribuzione di Poisson.

a – Sostituendo $n = 100$, $x = 2$ e $p = 0.03$ nella formula della distribuzione binomiale si ottiene

$$P(X = 2) = \binom{100}{2}(0.03)^2(0.97)^{98} = 0.22515$$

b – Sostituendo $x = 2$ e $\lambda = np = 100 \cdot 0.03 = 3$ nella formula della distribuzione di Poisson, si ottiene

$$P(X = 2) = \frac{e^{-3} \cdot 3^2}{2!} = 0.22404$$

Esempio 31
Se la probabilità che una persona sia allergica a un dato farmaco è $p = 0.001$, determinare le probabilità che su 2000 persone
a – meno di 2 siano allergiche;
a – 3 siano allergiche;
b – più di 2 siano allergiche.
La variabile X = numero delle persone allergiche è una variabile aleatoria con distribuzione binomiale, ma, poiché un caso di allergia è un evento raro, si può supporre che X segua la distribuzione di Poisson.
Si ha

a – $n = 2000 \qquad p = 0.001 \qquad \lambda = np = 2000 \cdot 0.001 = 2$

$$P(X < 2) = P(X = 0) + P(X = 1) = e^{-2} + \frac{e^{-2} 2^1}{1!} = 3e^{-2} = 0.4060$$

b – $$P(X = 3) = \frac{e^{-2} 2^3}{3!} = 0.1804$$

c – $$P(X > 2) = 1 - P(X \le 2) = 1 - \left[P(X = 0) + P(X = 1) + P(X = 2)\right] =$$
$$= 1 - \left(e^{-2} + \frac{e^{-2} 2^1}{1!} + \frac{e^{-2} 2^2}{2!} \right) = 1 - 5e^{-2} = 0.3233$$

Il calcolo della probabilità con la distribuzione binomiale è molto più laborioso; infatti con la distribuzione binomiale al punto c si dovrebbe calcolare la quantità seguente

c – $$P(X > 2) = 1 - P(X \le 2) = 1 - \left[P(X = 0) + P(X = 1) + P(X = 2)\right] =$$
$$= 1 - \left[\binom{2000}{0} (0.001)^0 (0.999)^{2000} + \binom{2000}{1} (0.001)^1 (0.999)^{1999} + \right.$$
$$\left. + \binom{2000}{2} (0.001)^2 (0.999)^{1998} \right]$$

Esempio 32
Un allevatore di galline per la produzione di uova ha acquistato 900 pulcini. Il venditore dichiara che, essendo stati selezionati accuratamente, solo un pulcino su 150 potrà risultare un maschio. Calcolare la probabilità che l'allevatore, quando i pulcini saranno adulti, si ritrovi
a – 7 galli e 893 galline;
b – meno di 4 galli;
c – più di 4 galli;
d – tutte galline.

Con le tavole della distribuzione di Poisson si trova

a – $$n = 900 \qquad p(\text{maschio}) = \frac{1}{150} \qquad \lambda = np = \frac{900}{150} = 6$$
$$P(X = 7) = P(X \le 7) - P(X \le 6) = 0.7440 - 0.6063 = 0.1377$$

b – $$P(X < 4) = P(X \le 3) = 0.1512$$

c – $$P(X > 4) = 1 - P(X \le 4) = 1 - 0.2851 = 0.7149$$

d – $$P(X = 0) = e^{-6} = 0.0025$$

In quest'ultimo caso la probabilità è molto bassa, perciò o l'allevatore è stato molto fortunato, oppure il venditore ha fatto un'affermazione falsa.

Esempio 33

Sia data la variabile aleatoria X avente distribuzione binomiale con parametri n e p; usare la distribuzione di Poisson per approssimare le probabilità nei seguenti casi

a – dati $n = 40$ e $p = 0.1$, calcolare $P(X \le 3)$ e $P(X \ge 3)$;

b – dati $n = 100$ e $p = 0.02$, calcolare $P(X \ge 2)$ e $P(X < 4)$;

c – dati $n = 55$ e $p = \frac{1}{11}$, calcolare $P(3 \le X \le 6)$.

Con l'uso delle tavole della distribuzione di Poisson si ottiene

a – $n = 40 \qquad p = 0.1 \qquad \lambda = np = 4$

$$P(X \le 3) = 0.4335$$

$$P(X \ge 3) = 1 - P(X < 3) = 1 - P(X \le 2) = 1 - 0.2381 = 0.7619$$

b – $n = 100 \qquad p = 0.02 \qquad \lambda = np = 2$

$$P(X \ge 2) = 1 - P(X < 2) = 1 - P(X \le 1) = 1 - 0.4060 = 0.5940$$

$$P(X < 4) = P(X \le 3) = 0.8571$$

c – $n = 55 \qquad p = \frac{1}{11} \qquad \lambda = np = 5$

$$P(3 \le X \le 6) = P(X \le 6) - P(X \le 2) = 0.7622 - 0.1247 = 0.6375$$

Con la distribuzione binomiale, effettuando i calcoli con un software statistico, si ottengono i valori

a – $P(X \le 3) = 0.4231 \qquad P(X \ge 3) = 0.7772$

b – $P(X \ge 2) = 0.5967 \qquad P(X < 4) = 0.8590$

c – $P(3 \le X \le 6) = 0.6565$

I valori nei casi a e c sono un po' meno accurati, rispetto al caso b: si ricordi la regola pratica per l'uso dell'approssimazione ($n \ge 50$ e $p \le 0.1$).

Esempio 34

Una compagnia di assicurazioni ha 3840 assicurati; se la probabilità che ognuno degli assicurati denunci almeno un incidente all'anno è $p = \frac{1}{1200}$, trovare le probabilità che 0, 1, 2, 3, 4, ... assicurati denuncino almeno un incidente all'anno.

La distribuzione binomiale non può essere usata per evidenti motivi pratici; si può usare invece la distribuzione di Poisson.

Si ha

$$n = 3840 \qquad p = \frac{1}{1200} \qquad \lambda = 3840 \cdot \frac{1}{1200} = 3.2$$

e con le tavole della distribuzione di Poisson si trova

$$f(0; 3.2) = F(0; 3.2) = 0.0408$$

$$f(1; 3.2) = F(1; 3.2) - F(0; 3.2) = 0.1712 - 0.0408 = 0.1304$$

$$f(2; 3.2) = F(2; 3.2) - F(1; 3.2) = 0.3799 - 0.1712 = 0.2087$$

.....

5. Distribuzioni di probabilità continue

Fra le densità di probabilità continue, la più importante è la **densità di probabilità normale**, di solito detta semplicemente **distribuzione normale** o anche **distribuzione di Gauss**, in onore del matematico Carl Friedrich Gauss (1777-1855), che diede importanti contributi allo studio di questa distribuzione.
La distribuzione è anche nota come **legge degli errori**, in quanto essa descrive in particolare la distribuzione degli errori casuali relativi a successive misure di una quantità fisica (vedere § 5.3).
La distribuzione normale è importante in statistica per tre motivi fondamentali:
1 – diversi fenomeni continui seguono, almeno approssimativamente, una distribuzione normale;
2 – la distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete;
3 – la distribuzione normale è alla base dell'inferenza statistica, in virtù del teorema del limite centrale, che sarà discusso nel capitolo 6.

5.1 Distribuzione normale o di Gauss

Definizione 1
La **densità di probabilità normale**, o **distribuzione normale** o **di Gauss**, è definita dalla funzione

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty \tag{5.1}$$

di parametri μ e σ, con $\sigma > 0$.

Si dimostra che μ e σ sono rispettivamente il valor medio e lo scarto quadratico medio della variabile aleatoria X distribuita secondo la distribuzione normale.
Le caratteristiche più importanti della distribuzione normale sono le seguenti.
La funzione $f(x)$ è definita su tutto l'asse reale e assume valori sempre positivi; è simmetrica rispetto alla retta $x = \mu$, cioè rispetto al valor medio della distribuzione. La moda e la mediana coincidono con il valor medio.
Il valore massimo della funzione viene assunto nel punto di ascissa μ ed è $y_{max} = \frac{1}{\sigma\sqrt{2\pi}}$; questo valore è perciò inversamente proporzionale a σ.
Lo scarto quadratico medio σ è uguale alla distanza dei punti di flesso da μ, ossia i punti di flesso hanno ascissa rispettivamente $\mu - \sigma$ e $\mu + \sigma$.
La distribuzione normale ha una forma a campana, il grafico di $f(x)$ è del tipo illustrato nella figura 1.

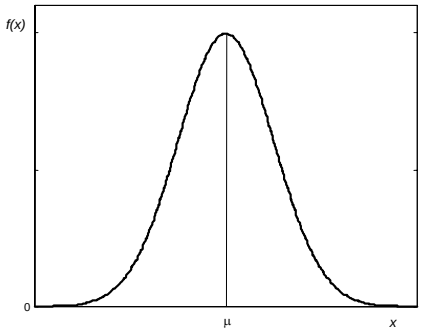


Figura 1

Poiché la curva rappresenta l'andamento della funzione di densità di una variabile aleatoria, il valore di tutta l'area sottesa da tale curva è uguale a 1.

La distribuzione normale è completamente individuata dai parametri μ e σ, ossia in corrispondenza di ogni valore di μ e σ rimane specificata una diversa curva normale appartenente alla famiglia rappresentata dall'equazione (5.1).
Nella figura 2 si riportano i grafici della distribuzione normale per un dato valore di μ e per diversi valori di σ: a parità di valor medio le variazioni della forma caratteristica a campana della curva dipendono essenzialmente dal valore dello scarto quadratico medio, che dà informazioni su come i valori sono più o meno concentrati intorno alla media: infatti facendo variare σ si ottengono curve più o meno appiattite (vedere anche l'esempio 5 e la figura 14).
Nella figura 3 si riportano invece i grafici della distribuzione normale per un dato valore di σ e per diversi valori di μ: in questo caso le variazioni del valore di μ comportano solo una traslazione della curva.

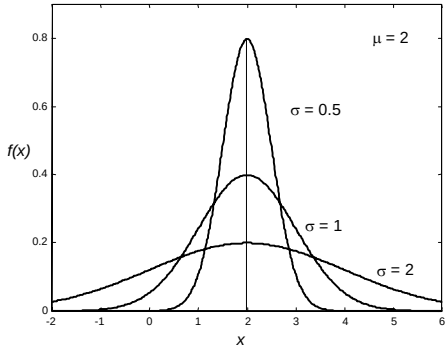


Figura 2

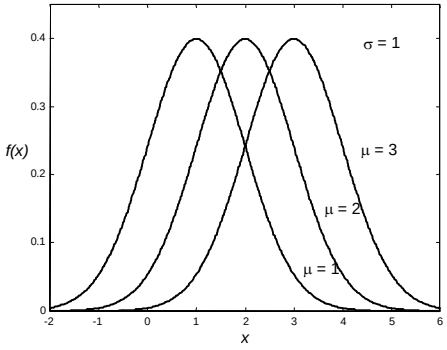


Figura 3

La **funzione di distribuzione** o **funzione di ripartizione normale** è data da

$$F(x) = P(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \qquad -\infty < x < \infty \qquad (5.2)$$

Nella figura 4 si riporta il grafico della funzione di distribuzione $F(x)$ per $\mu = 2$ e $\sigma = 1$

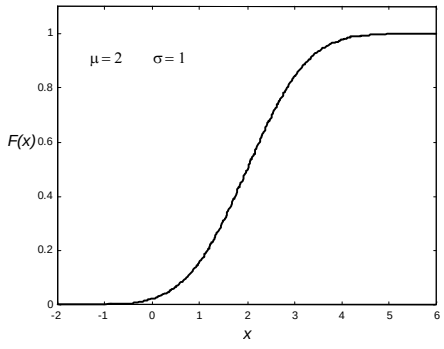


Figura 4

5.2 Distribuzione normale standardizzata

Come già osservato, la distribuzione normale è una famiglia di distribuzioni in cui ogni membro è distinto dall'altro in base ai valori di μ e σ. La curva più importante della famiglia è la **distribuzione normale standardizzata**. Per ricavare questa distribuzione, data la variabile aleatoria X distribuita normalmente con media μ e varianza σ^2, si passa alla nuova variabile aleatoria Z, detta **variabile standardizzata**, ponendo

$$Z = \frac{X - \mu}{\sigma}.$$

La trasformazione operata fa in modo che la media di Z sia 0 e la varianza 1.

La **distribuzione di probabilità della variabile normale standardizzata** Z è data da

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad -\infty < z < \infty \tag{5.3}$$

La **funzione di distribuzione** o **di ripartizione della variabile normale standardizzata** Z è data da

$$F(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt \qquad -\infty < z < \infty \tag{5.4}$$

I grafici della distribuzione normale standardizzata $f(z)$ e della relativa funzione di distribuzione $F(z)$ sono riportati nelle figure 5 e 6.

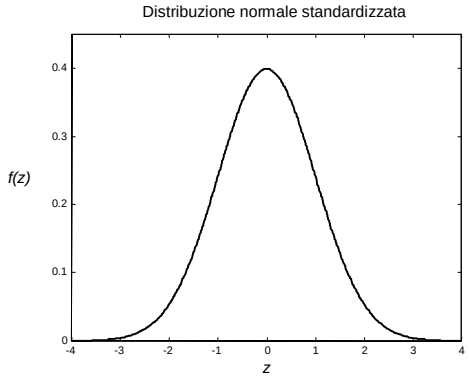


Figura 5

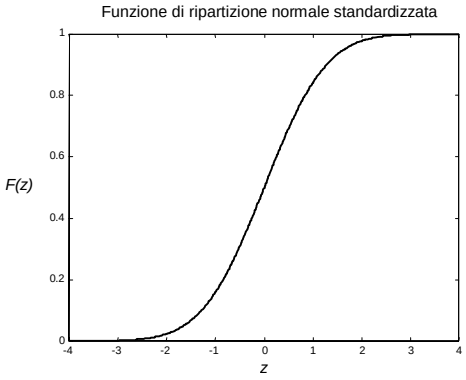


Figura 6

Nei grafici della figura 7, riproducenti la distribuzione normale standardizzata, indichiamo le aree comprese[1] rispettivamente tra -1 e 1, tra -2 e 2 e tra -3 e 3, pari al 68.27%, al 95.44% e al 99.73% dell'area totale, che è 1. Questo significa che

$$P(-1 < Z < 1) = 0.6827 \cong 68.3\%$$

$$P(-2 < Z < 2) = 0.9544 \cong 95.4\%$$

$$P(-3 < Z < 3) = 0.9973 \cong 99.7\%$$

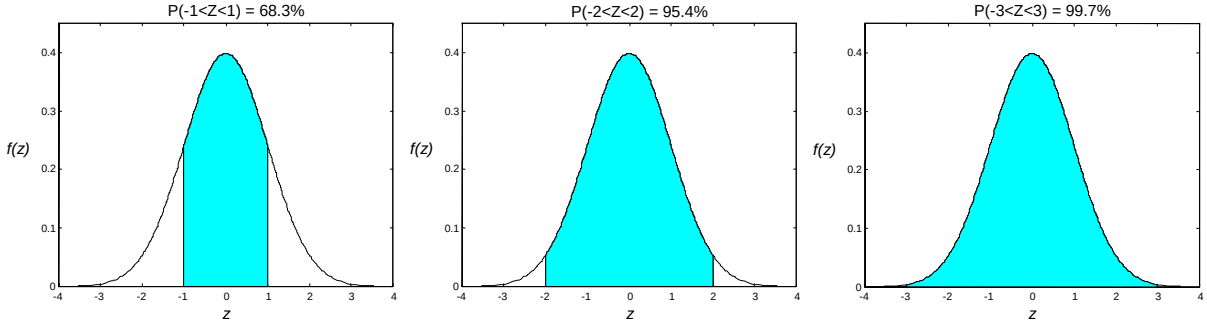


Figura 7

[1] Si ricordi l'osservazione 2, pag. 102: per le variabili aleatorie continue usare il segno < o il segno ≤ è indifferente.

Tenendo conto che per la variabile normale standardizzata lo scarto quadratico medio è uguale a 1, dal primo grafico della figura 7 si deduce sostanzialmente che una variabile aleatoria distribuita normalmente ha probabilità del 68.3% di discostarsi dalla media per meno di σ; analogamente dal secondo e dal terzo grafico si deduce che una variabile aleatoria normale ha probabilità del 95.4% di discostarsi dalla media per meno di 2σ e del 99.7% per meno di 3σ, cioè è quasi impossibile che si discosti dalla media per più di 3σ

$$P(\mu - \sigma < X < \mu + \sigma) \cong 68.3\%$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) \cong 95.4\%$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) \cong 99.7\%$$

5.3 Alcune applicazioni della distribuzione normale

Dopo aver introdotto da un punto di vista matematico la distribuzione normale e le sue proprietà elementari, illustriamo alcuni esempi nei quali la distribuzione normale viene utilizzata come modello probabilistico.

1 – Curva degli errori casuali nella misurazione di una grandezza fisica.

La misura, affetta da errore, di una qualunque grandezza fisica può essere vista come la somma del valore esatto della grandezza (che sarà un numero, costante) e dell'errore di misurazione, che è una variabile aleatoria, in quanto misure diverse forniscono in generale valori diversi.

La variabile aleatoria X = "errore di misurazione" ha come tipica densità di probabilità una curva a campana: l'errore può essere per eccesso o per difetto, perciò X può assumere valori positivi o negativi, in modo simmetrico; l'errore sarà in genere abbastanza piccolo, quindi la curva sarà rapidamente decrescente. Il fatto che, tra le infinite curve con questa proprietà, la normale rappresenti bene questo tipo di errori fu messo in evidenza da Gauss.

Se gli errori hanno media nulla, si dice che c'è solo **errore casuale**. Più grande è σ, maggiore sarà l'**inaccuratezza** della misura. Se poi il valor medio μ non è nullo, si dice che siamo anche in presenza di un **errore sistematico** μ che si somma all'errore casuale. Più grande è $|\mu|$, maggiore è l'**imprecisione** della misura. Si osservi che l'errore sistematico è una costante, mentre l'errore casuale è una variabile aleatoria.

2 – Distribuzione di una caratteristica quantitativa di una popolazione, che presenta oscillazioni casuali attorno a una media.

Molte grandezze antropometriche, come la statura, il peso, ecc., all'interno di una popolazione omogenea (ad esempio adulti, maschi, femmine, …) sono rappresentabili da una distribuzione gaussiana. Il valor medio μ della distribuzione è il valor medio della grandezza nella popolazione in esame; la varianza σ^2 è ragionevolmente piccola, se la popolazione è stata scelta in modo omogeneo. Anche altre misure di tipo fisiologico e biologico hanno un comportamento del tipo qui descritto.

3 – Dimensione effettiva di oggetti prodotti in serie, che si cerca di produrre in modo identico.

Ad esempio una ditta produce confezioni di biscotti che devono avere il peso di 250 g; il peso effettivo può essere rappresentato da una variabile aleatoria normale di valor medio μ = 250g e varianza più piccola possibile.

I tre tipi di esempi discussi sono simili, ma non uguali. Nel primo caso la variabilità è nelle misure che si fanno di una grandezza fissata una volta per tutte, ad esempio la massa di un oggetto che viene pesato tante volte; nel secondo caso la variabilità è tra individui diversi presenti in natura, ad esempio il peso di persone diverse; nel terzo caso la variabilità è tra oggetti diversi che vengono prodotti con l'intento di ottenerli uguali (per esempio il peso delle scatole di biscotti).

In tutti i casi si interpreta la variabilità della grandezza, vedendo il valore della variabile aleatoria X come il risultato di vari piccoli contributi; ad esempio l'errore nel misurare una lunghezza è dovuto al concorso di varie cause: inaccuratezza di chi esegue la misura, piccole variazioni della lunghezza dell'oggetto o dello strumento di misura, dovute a variazioni di temperatura, e così via.

5.4 Uso delle tavole della distribuzione normale

Poiché la distribuzione di probabilità $f(x)$ di una variabile aleatoria *X* distribuita normalmente non può essere integrata in forma chiusa fra gli estremi *a* e *b* di un intervallo, per il calcolo di $f(x)$ e di $F(x)$ si usano delle tavole. Tuttavia, poiché la (5.1) individua una famiglia di distribuzioni, ed esistono infinite combinazioni dei parametri μ e σ che individuano una curva della famiglia, non è possibile predisporre un numero infinito, o almeno molto elevato, di tavole. Si ricorre perciò alla variabile aleatoria standardizzata: è sempre possibile trasformare una distribuzione normale di parametri μ e σ nella corrispondente distribuzione standardizzata per mezzo del cambiamento di variabile

$$Z = \frac{X - \mu}{\sigma} \tag{5.5}$$

La tavola 3 riportata nell'Appendice A fornisce il valore della funzione di distribuzione della variabile aleatoria standardizzata *Z*

$$F(z) = P(Z \le z)$$

ossia il valore dell'area sottesa dalla curva normale standardizzata $f(z)$, a sinistra di un valore *z* assegnato; l'area è rappresentata nella figura 8

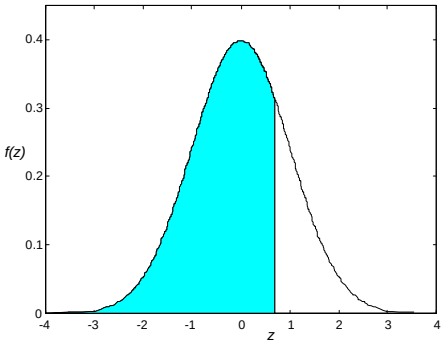


Figura 8

Valgono alcune proprietà utili per l'uso delle tavole[2].

Proprietà 1

$$P(-\infty < Z < \infty) = 1 \tag{5.6}$$

$$P(-\infty < Z < 0) = P(0 < Z < \infty) = F(0) = \frac{1}{2} \tag{5.7}$$

$$P(Z \le -z) = F(-z) = 1 - F(z), \quad z > 0 \tag{5.8}$$

$$P(z_1 \le Z \le z_2) = F(z_2) - F(z_1) \tag{5.9}$$

$$P(-z_1 \le Z \le 0) = P(0 \le Z \le z_1) \tag{5.10}$$

Esempio 1

Calcolare, usando la tavola della distribuzione normale standardizzata, la probabilità che una variabile aleatoria *Z* avente la distribuzione normale standardizzata assuma valori tali che[3]

[2] Si ricordi l'osservazione 2, pag. 102: per le variabili aleatorie continue usare il segno < o il segno ≤ è indifferente; questa proprietà sarà ripetutamente applicata negli esercizi che seguono.

[3] Nel calcolo di probabilità del tipo proposto in questo esempio (e in numerosi altri esempi di tipo analogo in queste lezioni), può essere molto utile tracciare un grafico qualitativo dell'area da calcolare, come nelle figure della pagina seguente. Spesso il grafico può suggerisce la lettura corretta delle tavole, e può mettere in risalto eventuali errori: se ad esempio l'area da calcolare è una gran parte dell'area sottesa dalla curva $f(x)$, ci si attende che la probabilità sia prossima 1, e così via. In particolare in questi tipi di calcolo è frequente commettere errori di segno: ottenere come risultato una probabilità negativa o maggiore di 1 indica senza alcun dubbio un qualche errore.

a – $0.87 \le Z \le 1.28$ (figura 9)
b – $-0.34 \le Z \le 0.62$ (figura 10)
c – $Z \ge 0.85$ (figura 11)
d – $Z \ge -0.65$ (figura 12)

a – $P(0.87 \le Z \le 1.28) = P(Z \le 1.28) - P(Z \le 0.87) =$
$= F(1.28) - F(0.87) = 0.8997 - 0.8078 = 0.0919$

b – $P(-0.34 \le Z \le 0.62) = F(0.62) - F(-0.34) =$
$= 0.7324 - \left[1 - F(0.34)\right] = 0.7324 - 1 + 0.6331 = 0.3655$

c – $P(Z \ge 0.85) = 1 - P(Z \le 0.85) = 1 - F(0.85) = 1 - 0.8023 = 0.1977$

d – $P(Z \ge -0.65) = P(Z \le 0.65) = F(0.65) = 0.7422$

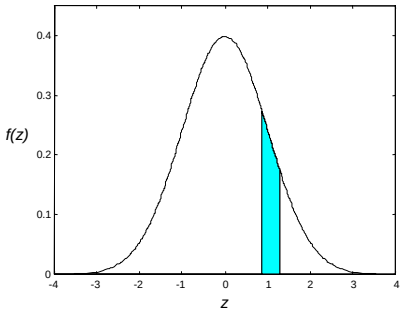


Figura 9

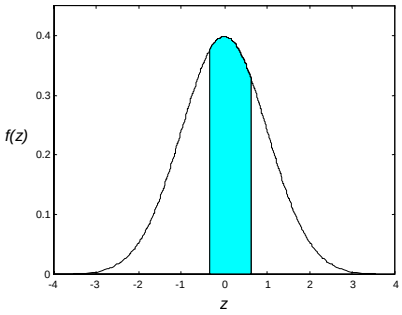


Figura 10

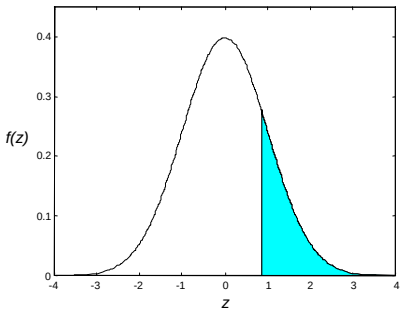


Figura 11

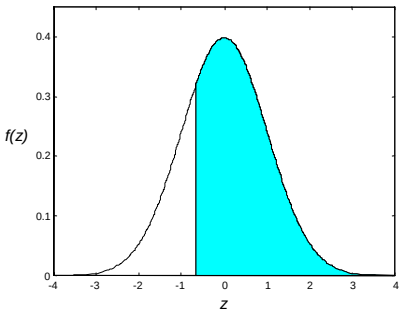


Figura 12

Esempio 2
Calcolare, usando la tavola della distribuzione normale standardizzata, la probabilità che una variabile aleatoria Z avente la distribuzione normale standardizzata assuma valori tali che
a – $1 < Z < 2$
b – $-1 < Z < 2$
c – $|Z| > 1.2$

a – $P(1 < Z < 2) = F(2) - F(1) = 0.9772 - 0.8413 = 0.1359$

b – $P(-1 < Z < 2) = F(2) - F(-1) = 0.9772 - \left[1 - F(1)\right] = 0.9772 - 1 + 0.8413 = 0.8185$

c – $P(|Z| > 1.2) = P(Z > 1.2) + P(Z < -1.2) = 1 - F(1.2) + F(-1.2) =$
$= 1 - F(1.2) + 1 - F(1.2) = 2 - 2 \cdot F(1.2) = 2 - 2 \cdot 0.8849 = 0.2302$

Se la variabile aleatoria non è standardizzata, prima di poter usare le tavole si deve ricorrere al cambiamento di variabile (5.5) per standardizzarla.

Esempio 3
Sia X una variabile aleatoria avente distribuzione normale, con $\mu = 4.35$ e $\sigma = 0.59$; trovare la probabilità $P(4 \le X \le 5)$ (figura 13).

Con il cambiamento di variabile $Z = \dfrac{X - \mu}{\sigma}$ si passa alla variabile standardizzata

$$X = 4 \;\Rightarrow\; Z = \frac{4 - 4.35}{0.59} = -0.5932$$

$$X = 5 \;\Rightarrow\; Z = \frac{5 - 4.35}{0.59} = 1.1017$$

$$P(4 \le X \le 5) = P(-0.59 \le Z \le 1.10) = F(1.10) - F(-0.59) =$$
$$= 0.8643 - 1 + F(0.59) = 0.5867$$

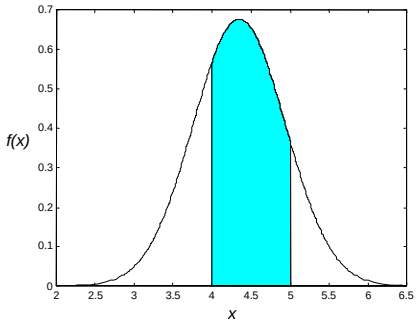


Figura 13

Esempio 4
L'altezza di un gruppo di ragazzi è distribuita normalmente con media $\mu = 174$ cm e scarto quadratico medio $\sigma = 15$ cm. Calcolare la probabilità che un ragazzo scelto a caso abbia una statura superiore a 190 cm.

Con il cambiamento di variabile $Z = \dfrac{X - \mu}{\sigma}$ si passa alla variabile standardizzata Z

$$\mu = 174 \qquad \sigma = 15$$

$$X = 190 \;\Rightarrow\; Z = \frac{190 - 174}{15} \cong 1.07$$

$$P(Z > 1.07) = 1 - P(Z < 1.07) = 1 - F(1.07) =$$
$$= 1 - 0.8577 = 0.1423 = 14.23\%$$

Esempio 5
Il diametro effettivo delle sfere di acciaio prodotte da una ditta può essere considerato una variabile aleatoria normale di media $\mu = 5.1$ cm e scarto quadratico medio $\sigma = 0.1$ cm.
a – Calcolare la probabilità che il diametro di una sfera scelta a caso sia compreso tra 5.0 e 5.2 cm.
b – Calcolare la stessa probabilità, supponendo che lo scarto quadratico medio sia $\sigma = 0.5$ cm.

Con il cambiamento di variabile $Z = \dfrac{X - \mu}{\sigma}$ si passa alla variabile standardizzata Z.

a – $\mu = 5.1 \qquad \sigma = 0.1$

$$X = 5.0 \;\Rightarrow\; Z = \frac{5.0 - 5.1}{0.1} = -1$$

$$X = 5.2 \;\Rightarrow\; Z = \frac{5.2 - 5.1}{0.1} = 1$$

$$P(5.0 \le X \le 5.2) = P(-1 \le Z \le 1) = 2\left[P(\mathrm{Z} \le 1) - 0.5\right] =$$

$$= 2(0.8413 - 0.5) = 0.6826 \cong 68\%$$

b – $\mu = 5.1 \qquad \sigma = 0.5$

$$X = 5.0 \;\Rightarrow\; Z = \frac{5.0 - 5.1}{0.5} = -0.2$$

$$X = 5.2 \;\Rightarrow\; Z = \frac{5.2 - 5.1}{0.5} = 0.2$$

$$P(5.0 \le X \le 5.2) = P(-0.2 \le Z \le 0.2) = 2\left[P(\mathrm{Z} \le 0.2) - 0.5\right] =$$

$$= 2(0.5793 - 0.5) = 0.1586 \cong 16\%$$

Si può osservare (figura 14) che aumentando la varianza, diminuisce la probabilità che i valori della variabile aleatoria avente distribuzione normale con media μ = 5.1 appartengano all'intervallo (5.0,5.2).

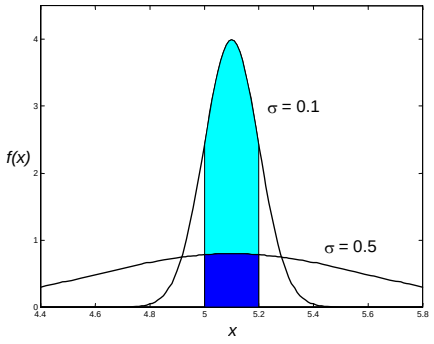


Figura 14

Esempio 6

La quantità di radiazioni cosmiche a cui è esposta una persona che attraversa in aereo gli Stati Uniti è una variabile aleatoria avente la distribuzione normale con media μ = 4.35 mrem e deviazione standard σ = 0.59 mrem. Trovare la probabilità che la quantità di radiazioni cosmiche a cui la persona sarà esposta sia

a – tra 4.00 e 5.00 mrem;

b – più di 5.50 mrem.

Con il cambiamento di variabile $Z = \dfrac{X - \mu}{\sigma}$ si passa alla variabile standardizzata

a – $$X = 4.00 \;\Rightarrow\; Z = \frac{4.00 - 4.35}{0.59} = -0.59$$

$$X = 5.00 \;\Rightarrow\; Z = \frac{5.00 - 4.35}{0.59} = 1.10$$

$$P(4.00 < X < 5.00) = P(-0.59 < Z < 1.10) = F(1.10) - F(-0.59) =$$

$$= F(1.10) - \left[1 - F(0.59)\right] = 0.8643 - 1 + 0.7224 = 0.5867$$

b – $$X = 5.50 \;\Rightarrow\; Z = \frac{5.50 - 4.35}{0.59} = 1.95$$

$$P(X > 5.50) = P(Z > 1.95) = 1 - F(1.95) = 1 - 0.9744 = 0.0256$$

Esempio 7
Il peso di certe confezioni alimentari prodotte in modo automatico è una variabile aleatoria normale X con media $\mu = 250$ g e deviazione standard $\sigma = 3$ g. Calcolare la probabilità che una confezione
a – pesi meno di 245 g;
b – pesi più di 250 g;
c – abbia un peso tra 247 g e 253 g.

Con il cambiamento di variabile $Z = \dfrac{X - \mu}{\sigma}$ si passa alla variabile standardizzata

a – $X = 245 \Rightarrow Z = \dfrac{245 - 250}{3} = -1.67$

$$P(X < 245) = P(Z < -1.67) = 1 - F(1.67) = 1 - 0.9525 = 0.0475$$

b – $\mu = 250 \Rightarrow P(X > 250) = 0.5$

c – $X = 247 \Rightarrow Z = \dfrac{247 - 250}{3} = -1$

$$X = 253 \Rightarrow Z = \frac{253 - 250}{3} = 1$$

$$P(247 < X < 253) = P(-1 < Z < 1) = F(1) - F(-1) = 2F(1) - 1 =$$
$$= 2 \cdot 0.8413 - 1 = 0.6826$$

Esempio 8
Il punteggio ottenuto in un test sul quoziente di intelligenza è una variabile aleatoria X avente distribuzione normale con media $\mu = 100$ e deviazione standard $\sigma = 15$.
Trovare la probabilità che il punteggio ottenuto da un candidato sia
a – minore di 118;
b – maggiore di 112;
c – compreso fra 100 e 112.

Con il cambiamento di variabile $Z = \dfrac{X - \mu}{\sigma}$ si passa alla variabile standardizzata

a – $X = 118 \Rightarrow Z = \dfrac{118 - 100}{15} = 1.2$

$$P(X < 118) = P(Z < 1.2) = F(1.2) = 0.8849$$

b – $X = 112 \Rightarrow Z = \dfrac{112 - 100}{15} = 0.8$

$$P(X > 112) = P(Z > 0.8) = 1 - F(0.8) = 1 - 0.7881 = 0.2119$$

c – $P(100 < X < 112) = P(0 < Z < 0.8) = 0.7881 - 0.5 = 0.2881$

Esempio 9
La lunghezza di una sbarretta costruita da una macchina automatica è una variabile aleatoria X distribuita normalmente, con media $\mu = 10$ cm e varianza $\sigma^2 = 0.005$.
Determinare la probabilità di scartare una sbarretta, se le dimensioni accettabili delle sbarrette sono 10 ± 0.05 cm.

Calcoliamo la probabilità che la sbarretta abbia dimensioni accettabili

$$P(9.95 \le X \le 10.05)$$

$\mu = 10 \qquad \sigma = \sqrt{0.005} \cong 0.0707$

Con il cambiamento di variabile $Z = \dfrac{X - \mu}{\sigma}$ si passa alla variabile standardizzata Z

$$X = 9.95 \;\Rightarrow\; Z = \frac{9.95-10}{0.0707} \cong -0.71$$

$$X = 10.05 \;\Rightarrow\; Z = \frac{10.05-10}{0.0707} \cong 0.71$$

$$P(9.95 \le X \le 10.05) = P(-0.71 \le Z \le 0.71) = F(0.71) - F(-0.71)$$
$$= 2\left[F(0.71) - 0.5\right] = 0.5222 \cong 52\%$$

Pertanto la probabilità di scartare una sbarretta è

$$P(|Z| > 0.71) = 1 - 0.5222 = 0.4778 \cong 48\%.$$

Trattando le variabili aleatorie continue, in particolare le variabili con distribuzione normale, capita spesso di dover risolvere il problema inverso a quello, già esaminato, del calcolo della probabilità $P(X \le x)$, ovvero: assegnato un valore $\alpha \in (0,1)$ determinare un numero reale x_α tale che $P(X > x_\alpha) = \alpha$; in altre parole x_α è il valore per cui l'area sottesa dalla distribuzione $f(x)$ a destra di x_α è uguale a α.

Se la funzione di ripartizione di X è strettamente crescente, allora x_α è determinato in modo unico; questo è il caso che si verifica con le più note distribuzioni continue.

Per la distribuzione normale standardizzata, oltre alla tavola 3, che riporta la funzione di ripartizione $F(z)$, nell'Appendice A è riportata la tavola 4, in cui compaiono i valori di z_α per i quali $P(Z > z_\alpha) = \alpha \cdot 100\%$, per alcuni valori notevoli di α; z_α è, come già osservato, il valore per il quale l'area sottesa dalla distribuzione $f(z)$ a destra di z_α è uguale a α. La tavola prende anche il nome di **tavola dei percentili della distribuzione normale standardizzata**.

Da questa tabella si legge ad esempio che il valore di z_α per il quale il 30% dei valori di *Z* cade a destra di z_α è $z_\alpha = 0.524$ (figura 15).

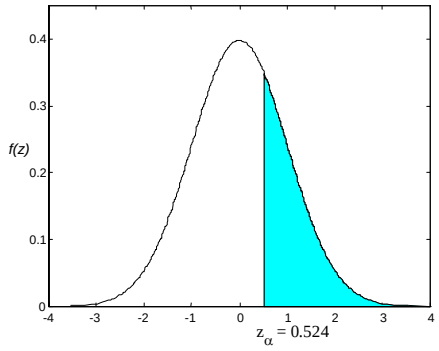


Figura 15

Gli esempi seguenti illustrano il modo di risolvere questo tipo di problema con l'utilizzo delle tavole 3 e 4.

Esempio 10

La variabile aleatoria *Z* ha la distribuzione normale standardizzata. Determinare il valore di z_α per cui

a – $P(Z < z_\alpha) = 0.9953$

b – $P(Z > z_\alpha) = 0.2743$

c – $P(0 \le Z \le z_\alpha) = 0.3770$

d – $P(|Z| < z_\alpha) = 0.5762$

e – $P(z_\alpha < Z < 1.6) = 0.7865$

a – Dalla tavola 3 si legge che $P(Z < 2.6) = 0.9953$, quindi

$$z_\alpha = 2.6$$

b – Si ha

$$P(Z < z_\alpha) = 1 - P(Z > z_\alpha) = 1 - 0.2743 = 0.7257$$

Leggendo la tavola 3 si trova che $P(Z < 0.6) = 0.7257$, quindi

$$z_\alpha = 0.6$$

c – Si ha

$$P(0 \le Z \le z_\alpha) = P(Z < z_\alpha) - 0.5 = 0.3770$$

$$P(Z < z_\alpha) = 0.8770$$

Leggendo la tavola 3 si trova che $P(Z < 1.16) = 0.8770$, quindi

$$z_\alpha = 1.16$$

d – Si ha

$$P(|Z| < z_\alpha) = P(-z_\alpha < Z < z_\alpha) = 2 \cdot P(0 < Z < z_\alpha) =$$

$$= 2 \cdot [P(Z < z_\alpha) - 0.5] = 2 \cdot P(Z < z_\alpha) - 1 = 0.5762$$

Pertanto

$$P(Z < z_\alpha) = \frac{1 + 0.5762}{2} = 0.7881$$

Leggendo la tavola 3 si trova che $F(0.8) = 0.7881$, quindi

$$z_\alpha = 0.8$$

e – $$P(z_\alpha < Z < 1.6) = P(Z < 1.6) - P(Z < z_\alpha) = 0.9452 - P(Z < z_\alpha) = 0.7865$$

$$P(Z < z_\alpha) = 0.9452 - 0.7865 = 0.1587$$

Dato che $P(Z < z_\alpha) = 0.1587 < 0.5$, segue che z_α è a sinistra dell'origine; cerchiamo allora il punto z_α^* , simmetrico di z_α rispetto all'origine

$$P(Z < z_\alpha) = P(Z > z_\alpha^*) = 0.1587$$

$$P(Z > z_\alpha^*) = 1 - P(Z < z_\alpha^*) = 0.1587$$

$$P(Z < z_\alpha^*) = 1 - 0.1587 = 0.8413$$

Leggendo la tavola 3 si trova che $F(1) = 0.8413$, quindi

$$z_\alpha^* = 1 \quad \text{e} \quad z_\alpha = -1.$$

Esempio 11

La variabile aleatoria Z ha la distribuzione normale standardizzata. Trovare il valore z_α tale che

a – $P(Z \ge z_\alpha) = 0.01 = 1\%$;

b – $P(Z \ge z_\alpha) = 0.05 = 5\%$;

c – $P(-z_\alpha < Z < z_\alpha) = 0.6$.

a – Dalla tavola 4 si ricava che il valore z_α per il quale $P(Z \ge z_\alpha) = 0.01 = 1\%$ è $z_\alpha = 2.326$.

b – Dalla tavola 4 si ricava che il valore z_α per il quale $P(Z \ge z_\alpha) = 0.05 = 5\%$ è $z_\alpha = 1.645$.

c – $$P(-z_\alpha < Z < z_\alpha) = 0.6 \Rightarrow P(0 < Z < z_\alpha) = 0.3 = 30\%$$

$$P(Z \ge z_\alpha) = 50\% - 30\% = 20\%$$

Dalla tavola 4 si ricava che $z_\alpha = 0.842$.

Esempio 12

La variabile aleatoria X ha la distribuzione normale con valor medio $\mu = 19$ e varianza $\sigma^2 = 49$; determinare il valore x_α tale che

a – $P(X > x_\alpha) = 0.20 = 20\%$;

b – $P(X < x_\alpha) = 0.90 = 90\%$.

a – Passando alla variabile normale standardizzata si ha

$$P(X > x_\alpha) = P\left(Z > z_\alpha = \frac{x_\alpha - 19}{7} \right) = 0.20 = 20\%$$

Sulla tavola 4 si trova

$$z_\alpha = \frac{x_\alpha - 19}{7} = 0.842$$

$$x_\alpha = 19 + 7 \cdot 0.842 = 24.894 \cong 24.9$$

b – La condizione richiesta significa che il 90% dell'area sottesa dalla curva normale è a destra di x_α, quindi il 10% è a sinistra. Passando alla variabile normale standardizzata si ha

$$P(X < x_\alpha) = P\left(Z < z_\alpha = \frac{x_\alpha - 19}{7} \right) = 0.90 = 90\%$$

$$P\left(Z > z_\alpha = \frac{x_\alpha - 19}{7} \right) = 0.10 = 10\%$$

Sulla tavola 4 si trova

$$z_\alpha = \frac{x_\alpha - 19}{7} = 1.282$$

$$x_\alpha = 19 + 7 \cdot 1.282 = 27.97$$

Esempio 13

Una macchina viene usata per tagliare assi di legno; la lunghezza media è di 2m, ma il 10% degli assi tagliati hanno una lunghezza inferiore a 1.95m.
Assumendo che le lunghezze degli assi tagliati abbiano una distribuzione normale, determinare la percentuale di assi più lunghi di 2.10m.

Sia X la variabile aleatoria che misura la lunghezza; X è distribuita normalmente con media $\mu = 2$; inoltre si sa che

$$P(X < 1.95) = 10\% .$$

Si deve calcolare $P(X > 2.10)$ e per far questo occorre prima determinare lo scarto quadratico medio σ. Passando alla variabile aleatoria standardizzata si ha

$$P(X < 1.95) = P\left(Z < \frac{1.95 - 2.00}{\sigma} \right) = 10\%$$

$$P\left(Z < \frac{1.95 - 2.00}{\sigma} \right) = P\left(Z > -\frac{1.95 - 2.00}{\sigma} \right) = 10\%$$

Sulla tavola 4 si trova che

$$-\frac{1.95 - 2.00}{\sigma} = 1.282$$

$$0.05 = 1.282 \cdot \sigma \qquad \Rightarrow \qquad \sigma = \frac{0.05}{1.282} = 0.039$$

Calcoliamo ora $P(X > 2.10)$. Passando alla variabile aleatoria standardizzata si ha

$$X = 2.10 \;\Rightarrow\; Z = \frac{2.10 - 2.00}{0.039} \cong 2.56$$

$$P(X > 2.10) = P(Z > 2.56) = 1 - P(Z < 2.56) = 1 - 0.9948 = 0.0052 .$$

In altre parole la percentuale di assi più lunghi di 2.10m è circa dello 0.5%.

Esempio 14

La variabile aleatoria X ha distribuzione normale con media μ e varianza σ^2. E' noto che il 10% dei valori di X è maggiore di 17.24 e che il 25% dei valori è minore di 14.37. Trovare il valor medio e la varianza.

Sono note le probabilità

$$P(X > 17.24) = 10\% \qquad P(X < 14.37) = 25\% .$$

Standardizzando la variabile e usando la tabella 4 si trova

$$P(X > 17.24) = P\left(Z > \frac{17.24 - \mu}{\sigma} \right) = 10\%$$

$$\frac{17.24 - \mu}{\sigma} = 1.282 .$$

$$P(X < 14.37) = P\left(Z < \frac{14.37 - \mu}{\sigma} \right) = 25\%$$

$$P\left(Z > -\frac{14.37 - \mu}{\sigma} \right) = 25\%$$

$$-\frac{14.37 - \mu}{\sigma} = 0.674 .$$

Risolvendo il sistema seguente si determinano i valori di μ e σ

$$\begin{cases} \dfrac{17.24 - \mu}{\sigma} = 1.282 \\ \dfrac{14.37 - \mu}{\sigma} = -0.674 \end{cases}$$

$$\mu = 15.4 \qquad \sigma = 1.47$$

Esempio 15

La variabile aleatoria X ha distribuzione normale con media μ e varianza σ^2. E' noto che

$$P(X > 9) = 0.9192 \qquad P(X < 11) = 0.7580$$

Calcolare $P(X > 10)$.

Calcoliamo dapprima i valori di μ e σ.

$$P(X > 9) = 1 - P(X < 9) = 0.9192$$

$$P(X < 9) = 1 - 0.9192 = 0.0808$$

Standardizzando la variabile si ha

$$P\left(Z < \frac{9 - \mu}{\sigma} \right) = 0.0808 < 0.5 \quad \Rightarrow \quad \frac{9 - \mu}{\sigma} < 0$$

$$P\left(Z < \frac{9 - \mu}{\sigma} \right) = P\left(Z > -\frac{9 - \mu}{\sigma} \right) = 1 - P\left(Z < -\frac{9 - \mu}{\sigma} \right) = 0.0808$$

$$P\left(Z < -\frac{9 - \mu}{\sigma} \right) = 1 - 0.0808 = 0.9192$$

$$P(X < 11) = P\left(Z < \frac{11 - \mu}{\sigma} \right) = 0.7580$$

Usando la tavola 3 si ha

$$\begin{cases} -\dfrac{9 - \mu}{\sigma} = 1.4 \\ \dfrac{11 - \mu}{\sigma} = 0.7 \end{cases}$$

$$\mu = \frac{31}{3} \qquad \sigma = \frac{20}{21} .$$

Usando la tavola 3 si calcola $P(X > 10)$

$$P(X > 10) = P\left(Z > \frac{10 - \frac{31}{3}}{\frac{20}{21}} \right) = P\left(Z > -\frac{7}{20} \right) = P(Z < 0.35) = 0.6368$$

5.5 Relazione tra la distribuzione binomiale e la distribuzione normale

Sia X la variabile aleatoria che fornisce il numero di successi in n prove bernoulliane e p la probabilità di successo; quando il numero n delle prove è grande, il calcolo con la distribuzione binomiale è molto lungo. In tal caso è possibile utilizzare la distribuzione normale per approssimare la distribuzione binomiale.
Si può dimostrare che, quando n è grande e p è vicino a 0.5, la distribuzione binomiale della variabile aleatoria X può essere approssimata da una distribuzione normale con variabile aleatoria standardizzata

$$Z = \frac{X - np}{\sqrt{np(1-p)}}. \qquad (5.11)$$

L'approssimazione migliora al crescere di n e per $n \to \infty$ le due distribuzioni coincidono; se ricordiamo che per una variabile aleatoria binomiale X, la media e la varianza sono rispettivamente

$$\mu = np \qquad \sigma^2 = np(1-p)$$

allora la (5.11) non è altro che la formula per la standardizzazione della variabile X.
Di conseguenza la distribuzione della variabile aleatoria binomiale X di parametri n e p viene approssimata con la distribuzione normale di media $\mu = np$ e varianza $\sigma^2 = np(1-p)$.
Come **regola pratica** si usa la distribuzione normale per approssimare la binomiale se si verificano entrambe le condizioni $np \geq 5$ e $n(1-p) \geq 5$.
La regola suggerita è soddisfatta se n è abbastanza grande e l'approssimazione è tanto più precisa quanto più p è prossima a 0.5.
Si ricordi che se n è grande e p è piccolo, la binomiale può essere approssimata dalla distribuzione di Poisson con parametro $\lambda = np$; se invece p è prossimo a 1, si può contare il numero di insuccessi, anziché quello dei successi: in questo modo la probabilità di insuccesso $1-p$ è piccola e si può ancora usare la distribuzione di Poisson.
Nella figura 16, per illustrare l'approssimazione fra la distribuzione binomiale e la normale, sono riportati il grafico della distribuzione binomiale per $n = 20$ e $p = 0.5$ e il grafico della distribuzione normale avente valor medio $\mu = np = 10$ e varianza $\sigma^2 = np(1-p) = 5$.
Nella figura 17 si illustra un caso in cui l'approssimazione della binomiale con la normale non è altrettanto buona

$$n = 20 \qquad p = 0.1 \qquad \Rightarrow \qquad np = 2 \qquad n(1-p) = 18$$

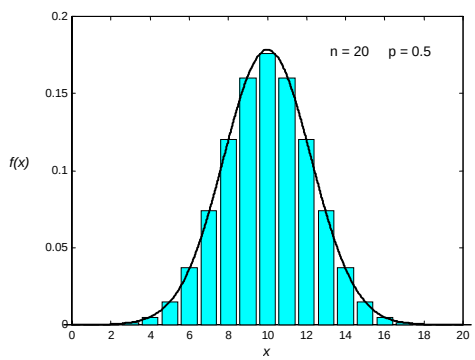


Figura 16

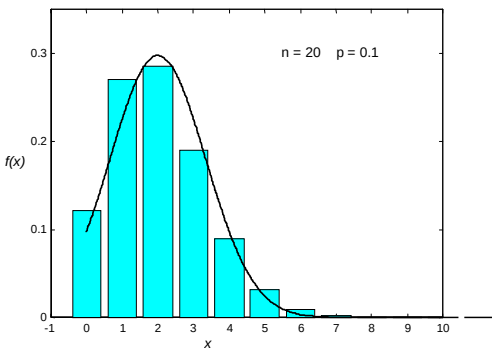


Figura 17

L'approssimazione migliora nei casi seguenti, in cui, malgrado sia $p = 0.1$, tuttavia

$$n = 50 \qquad p = 0.1 \qquad \Rightarrow \qquad np = 5 \quad n(1-p) = 45 \qquad \text{(figura 18)}$$

e

$$n = 100 \qquad p = 0.1 \qquad \Rightarrow \qquad np = 10 \quad n(1-p) = 90 \qquad \text{(figura 19)}$$

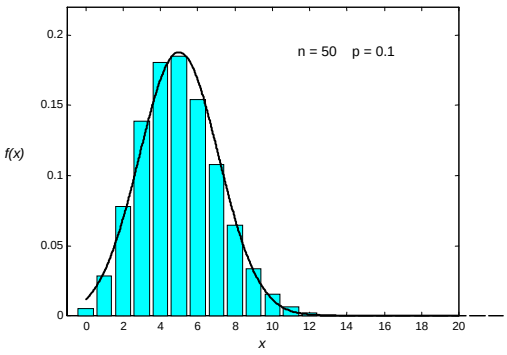


Figura 18

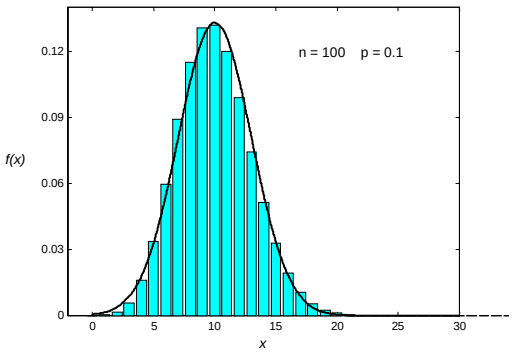


Figura 19

Per poter usare correttamente la distribuzione normale, che è continua, per approssimare la distribuzione di una variabile aleatoria discreta occorre effettuare la **correzione di continuità**[4]: questo avviene rappresentando ogni valore intero x assunto dalla variabile aleatoria discreta con l'intervallo di estremi $x-\frac{1}{2}$ e $x+\frac{1}{2}$. Quindi, se X è una variabile aleatoria con distribuzione binomiale di parametri n e p, la probabilità $P(a \le X \le b)$ che X assuma valori compresi fra a e b, viene approssimata con il valore della probabilità che la variabile aleatoria normale con media $\mu = np$ e varianza $\sigma^2 = np(1-p)$ assuma valori compresi tra $a-\frac{1}{2}$ e $b+\frac{1}{2}$, ossia con il valore dell'area sottesa dalla curva normale tra $a-\frac{1}{2}$ e $b+\frac{1}{2}$.

Nel caso particolare in cui $a = b$, la probabilità binomiale $P(X = a)$ viene approssimata con il valore della probabilità $P\left(a-\frac{1}{2} \le X \le a+\frac{1}{2}\right)$ calcolata con la distribuzione normale.

Esempio 16

Trovare la probabilità che in 100 lanci di una moneta, testa si presenti 40 volte, usando la distribuzione normale per approssimare la distribuzione binomiale.

Per calcolare la probabilità $P(X = 40)$ usando la distribuzione normale, occorre effettuare la correzione di continuità e calcolare la probabilità

$$P\left(40-\frac{1}{2} \le X \le 40+\frac{1}{2}\right) = P(39.5 \le X \le 40.5)$$

Standardizzando la variabile con la (5.5) si ha

$$\mu = np = 100 \cdot \frac{1}{2} = 50 \qquad \sigma^2 = np(1-p) = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25$$

$$X = 39.5 \quad \Rightarrow \quad Z = \frac{39.5-50}{\sqrt{25}} = -2.1$$

$$X = 40.5 \quad \Rightarrow \quad Z = \frac{40.5-50}{\sqrt{25}} = -1.9$$

[4] Vedere anche l'osservazione a pag. 161 e l'esempio 21

Usando le tavole della distribuzione normale si trova

$$P(-2.1 < Z < -1.9) = P(1.9 < Z < 2.1) =$$
$$= P(Z < 2.1) - P(Z < 1.9) = 0.9821 - 0.9713 = 0.0108$$
$$P(X = 40) \cong 0.0108$$

Questa approssimazione è molto buona, perché il valore di *n* è sufficientemente grande e il valore di *p* è 0.5.

Esempio 17

Trovare la probabilità che, in 10 lanci di una moneta, testa si presenti un numero di volte compreso fra 3 e 6, usando

a – la distribuzione binomiale;

b – la distribuzione normale per approssimare la distribuzione binomiale.

a – Sia X la variabile aleatoria binomiale.

Si deve calcolare la probabilità $P(3 \le X \le 6)$. Con le tavole della distribuzione binomiale si ha

$$n = 10 \qquad p = \frac{1}{2}$$
$$P(3 \le X \le 6) = P(X \le 6) - P(X \le 2) = 0.8281 - 0.0547 = 0.7734$$

b – Se si considera la variabile X come continua, si deve fare la correzione di continuità e calcolare la probabilità $P(2.5 \le X \le 6.5)$; standardizzando la variabile con la (5.11) si ha

$$\mu = np = 10 \cdot \frac{1}{2} = 5 \qquad \sigma^2 = np(1-p) = 10 \cdot \frac{1}{2} \cdot \frac{1}{2} = 2.5$$
$$X = 2.5 \quad \Rightarrow \quad Z = \frac{2.5 - 5}{\sqrt{2.5}} = -1.58$$
$$X = 6.5 \quad \Rightarrow \quad Z = \frac{6.5 - 5}{\sqrt{2.5}} = 0.95$$

Usando le tavole della distribuzione normale si trova

$$P(-1.58 < Z < 0.95) = P(Z < 0.95) - P(Z < -1.58) =$$
$$= 0.8289 - [1 - P(Z < 1.58)] = 0.8289 - 1 + 0.9429 = 0.7718$$
$$P(3 \le X \le 6) \cong 0.7718$$

Il valore ottenuto con la distribuzione normale approssima sufficientemente bene il valore esatto trovato con la binomiale, anche se *n* non è molto grande, perché $p = 0.5$.

Nella figura 20 l'area ombreggiata rappresenta il valore trovato con la binomiale; il valore calcolato con la normale è uguale all'area sottesa dalla normale fra 2.5 e 6.5.

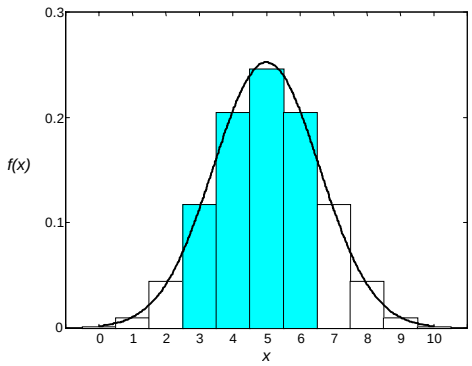


Figura 20

Esempio 18
Si effettuano 500 lanci di una moneta; calcolare la probabilità che il numero di teste non differisca da 250
a – per più di 10;
b – per più di 30.
Usare l'approssimazione della distribuzione binomiale con la normale.

a – In questo caso si cerca la probabilità che il numero di teste sia compreso fra 240 e 260, ossia, con la correzione di continuità, la probabilità

$$P(239.5 < X < 260.5)$$

Effettuando il passaggio alla variabile standardizzata si ha

$$n = 500 \qquad p = \frac{1}{2} \qquad \mu = np = 250$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{500\frac{1}{2}\frac{1}{2}} = 11.18$$

$$X = 239.5 \quad \Rightarrow \quad Z = \frac{239.5 - 250}{11.18} = -0.94$$

$$X = 260.5 \quad \Rightarrow \quad Z = \frac{260.5 - 250}{11.18} = 0.94$$

Usando le tavole della distribuzione normale si trova

$$P(-0.94 < Z < 0.94) = P(Z < 0.94) - \left[1 - P(Z < 0.94)\right] =$$
$$= 2 \cdot 0.8264 - 1 = 0.6528 \cong 65.3\%$$
$$P(240 \le X \le 260) \cong 0.6528$$

b – In questo caso si cerca la probabilità che il numero di teste sia compreso fra 220 e 280, ossia, con la correzione di continuità, la probabilità

$$P(219.5 < X < 280.5)\,.$$

Effettuando il passaggio alla variabile standardizzata si ha

$$\mu = np = 250 \qquad \sigma = 11.18$$

$$X = 219.5 \quad \Rightarrow \quad Z = \frac{219.5 - 250}{11.18} = -2.73$$

$$X = 280.5 \quad \Rightarrow \quad Z = \frac{280.5 - 250}{11.18} = 2.73$$

Usando le tavole della distribuzione normale si trova

$$P(-2.73 < Z < 2.73) = 2 \cdot 0.9968 - 1 = 0.9936 \cong 99.4\%$$
$$P(220 \le X \le 280) \cong 0.9936$$

Esempio 19
Un dado viene lanciato 120 volte. Calcolare la probabilità che il numero 3 si presenti al più 15 volte.

La faccia con il numero 3 ha la probabilità $p = \frac{1}{6}$ di presentarsi. La probabilità che il numero 3 si presenti un numero di volte compreso fra 0 e 15, con la distribuzione binomiale è

$$n = 120 \qquad p = \frac{1}{6}$$
$$P(0 \le X \le 15) = P(X = 0) + P(X = 1) + ... + P(X = 15)$$

$$P(0 \le X \le 15) = \binom{120}{0}\left(\frac{1}{6}\right)^{0}\left(\frac{5}{6}\right)^{120} + \binom{120}{1}\left(\frac{1}{6}\right)^{1}\left(\frac{5}{6}\right)^{119} + \ldots$$
$$+ \binom{120}{14}\left(\frac{1}{6}\right)^{14}\left(\frac{5}{6}\right)^{106} + \binom{120}{15}\left(\frac{1}{6}\right)^{15}\left(\frac{5}{6}\right)^{105}$$

Il lavoro necessario per il calcolo dei 16 addendi presenti nella somma è eccessivo ed è preferibile usare l'approssimazione con la normale; si ottiene una buona approssimazione, dato che

$$np = 120 \cdot \frac{1}{6} = 20 \qquad \text{e} \qquad n(1-p) = 120 \cdot \frac{5}{6} = 100 .$$

Effettuando la correzione di continuità e standardizzando la variabile si trova

$$\mu = np = 120 \cdot \frac{1}{6} = 20 \qquad \sigma = \sqrt{np(1-p)} = \sqrt{120 \cdot \frac{1}{6} \cdot \frac{5}{6}} = 4.08$$

$$X = -0.5 \quad \Rightarrow \quad Z = \frac{-0.5 - 20}{4.08} = -5.02$$

$$X = 15.5 \quad \Rightarrow \quad Z = \frac{15.5 - 20}{4.08} = -1.10$$

$$P(-5.02 < Z < -1.10) = P(1.10 < Z < 5.02) =$$
$$= 0.9999997 - 0.8643 = 0.1357$$

$$P(0 \le X \le 15) \cong 0.1357$$

Effettuando con un software statistico il calcolo della probabilità con la distribuzione binomiale si trova il valore

$$P(0 \le X \le 15) = 0.1335 .$$

L'approssimazione ottenuta con la normale è buona, anche se la probabilità di successo $p = \frac{1}{6}$ non è vicina a 0.5; ciò è dovuto al valore elevato del numero di prove (si veda anche l'esempio seguente).

Esempio 20

Il 20% dei chip di memoria prodotti da un'azienda di componenti elettronici è difettoso; calcolare la probabilità che in un campione di 100 chip scelto a caso per un controllo

a – al più 15 siano difettosi;

b – esattamente 15 siano difettosi.

a – Si deve calcolare la probabilità $P(X \le 15)$.

Usando l'approssimazione con la normale ed effettuando la correzione di continuità, si ha

$$np = 100 \cdot 0.2 = 20 \qquad n(1-p) = 100 \cdot 0.8 = 80$$

$$\mu = np = 100 \cdot 0.2 = 20 \qquad \sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot 0.2 \cdot 0.8} = 4$$

$$X = 15.5 \quad \Rightarrow \quad Z = \frac{15.5 - 20}{4} = -1.13$$

$$P(Z < -1.13) = 1 - P(Z < 1.13) = 1 - 0.8708 = 0.1292$$

$$P(X \le 15) \cong 0.1292$$

b – Si deve calcolare $P(X = 15)$.

Usando l'approssimazione con la normale ed effettuando la correzione di continuità, si ha

$$X = 14.5 \quad \Rightarrow \quad Z = \frac{14.5 - 20}{4} = -1.38$$

$$X = 15.5 \quad \Rightarrow \quad Z = \frac{15.5 - 20}{4} = -1.13$$

$$P(-1.38 < Z < -1.13) = P(Z < 1.38) - P(Z < 1.13) =$$
$$= 0.9162 - 0.8708 = 0.0454$$
$$P(X = 15) \cong 0.0454$$

Per confronto si può effettuare con un software statistico il calcolo delle probabilità con la distribuzione binomiale e si trovano i valori

$$P(X \le 15) = 0.1285$$
$$P(X = 15) = 0.0481$$

Osservazione

Per poter applicare la distribuzione normale ad un caso di dati discreti è necessario trattare i dati come se fossero continui e quindi occorre effettuare la correzione di continuità (anche se non si tratta di approssimare una distribuzione discreta). Si consideri a questo proposito il seguente esempio.

Esempio 21

I voti di un questionario vanno da 1 a 10, a seconda del numero di risposte a 10 domande. Il voto medio è $\mu = 6.7$ e lo scarto quadratico medio è $\sigma = 1.2$. Supponendo che i voti siano distribuiti normalmente determinare

a – la percentuale di studenti che ha ottenuto il voto 6;

b – il voto minimo del miglior 10% del gruppo di studenti;

c – il voto massimo del peggior 10% del gruppo di studenti.

a – Effettuando la correzione di continuità, calcoliamo con la distribuzione normale la probabilità $P(5.5 < X < 6.5)$.

Standardizzando la variabile con la (5.5) si ha

$$\mu = 6.7 \qquad \sigma = 1.2$$
$$X = 5.5 \quad \Rightarrow \quad Z = \frac{5.5 - 6.7}{1.2} = -1.0$$
$$X = 6.5 \quad \Rightarrow \quad Z = \frac{6.5 - 6.7}{1.2} \cong -0.17$$

Usando le tavole della distribuzione normale si trova

$$P(5.5 < X < 6.5) = P(-1.0 < Z < -0.17) = P(0.17 < Z < 1.0) =$$
$$= F(1.0) - F(0.17) = 0.8413 - 0.5675 = 0.2738 \cong 27.4\%$$

b – Sia x_1 il voto minimo richiesto e z_1 il voto corrispondente in unità standardizzate.

Dalla figura 21 (pagina seguente), si vede che l'area a destra di z_1 è il 10% dell'area totale. Dalle tavole dei quantili per la distribuzione normale si ricava

$$z_1 = 1.282$$

Dalla relazione (5.5) si ottiene

$$z_1 = \frac{x_1 - 6.7}{1.2} = 1.282 \qquad \Rightarrow \qquad x_1 = 1.2 \cdot 1.282 + 6.7 \cong 8.24$$

Il voto minimo del miglior 10% degli studenti è 8 (l'intero più prossimo a x_1)

c – Il punto z_2 è il simmetrico di z_1 rispetto all'origine, ossia $z_1 = -1.282$; quindi

$$z_2 = \frac{x_2 - 6.7}{1.2} = -1.282 \qquad \Rightarrow \qquad x_2 = -1.2 \cdot 1.282 + 6.7 \cong 5.16$$

Il voto massimo del peggior 10% degli studenti è perciò 5 (l'intero più prossimo a x_2).

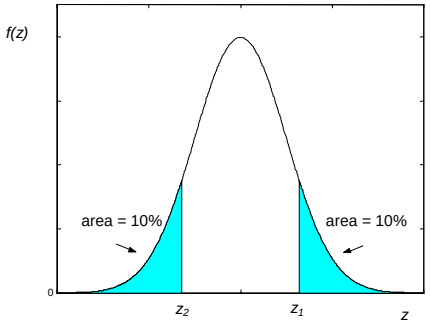


Figura 21

5.6 Relazione tra la distribuzione normale e la distribuzione di Poisson

Ricordiamo che la distribuzione di Poisson è stata ottenuta come il limite per $n \to \infty$ di una distribuzione binomiale; questo fatto suggerisce che esista anche una relazione fra la distribuzione normale e la distribuzione di Poisson.
Si dimostra che se X è una variabile aleatoria avente la distribuzione di Poisson, con media $\mu = \lambda$ e varianza $\sigma^2 = \lambda$, allora al crescere λ la distribuzione della variabile X può essere approssimata da una distribuzione normale con variabile aleatoria standardizzata

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \tag{5.12}$$

Come per la binomiale, anche per la distribuzione di Poisson, trattandosi di una distribuzione discreta, occorre fare la correzione di continuità.
L'approssimazione è sufficientemente buona per $\lambda \geq 10$.
Nella figura 22, per illustrare l'approssimazione fra la distribuzione di Poisson e la normale, sono riportati il grafico della distribuzione di Poisson per $\lambda = 10$ e il grafico della distribuzione normale avente valor medio $\mu = \lambda = 10$ e scarto quadratico medio $\sigma = \sqrt{\lambda} = \sqrt{10}$.

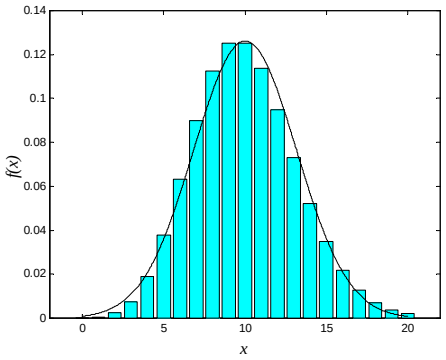


Figura 22

Esempio 22
La variabile aleatoria X ha distribuzione di Poisson con media $\lambda = 50$. Calcolare la probabilità $P(X < 40)$ usando l'approssimazione con la normale.

Si deve calcolare

$$P(X < 40) = P(X \leq 39)$$

Usando la distribuzione normale con la correzione di continuità si trova

$$X = 39.5 \quad \Rightarrow \quad Z = \frac{39.5 - 50}{\sqrt{50}} \cong -1.48$$

$$P(Z < -1.48) = 1 - P(Z > 1.48) = 1 - 0.9306 = 0.0694$$

$$P(X < 40) \cong 0.0694$$

Effettuando con un software statistico il calcolo delle probabilità con la distribuzione di Poisson, si trova il valore

$$P(X \le 39) = 0.0646 .$$

Esempio 23
Il numero di incidenti d'auto che si verificano in un giorno ad un incrocio è una variabile aleatoria con distribuzione di Poisson e media 1.4; calcolare la probabilità che accadano più di 50 incidenti in un periodo di 4 settimane.

Il numero di incidenti che si verificano in 28 giorni è una variabile X con media $\lambda = 1.4 \cdot 28 = 39.2$. Si ha

$$P(X > 50) = 1 - P(X \le 50)$$
$$Z = \frac{50.5 - 39.2}{\sqrt{39.2}} \cong 1.80$$
$$P(X > 50) \cong 1 - P(Z < 1.80) = 1 - 0.9645 = 0.0355$$

5.7 Distribuzione uniforme

La distribuzione studiata nell'esempio 9, pag. 96 fornisce un esempio di una distribuzione discreta, detta **distribuzione uniforme discreta**.
La distribuzione uniforme che viene introdotta con la definizione seguente è l'analoga nel caso continuo della distribuzione uniforme discreta.

Definizione 2
Dati due numeri reali a e b , con $a < b$, si dice che la variabile aleatoria X ha **distribuzione uniforme** con parametri a e b, se la sua densità di probabilità è

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{altrimenti} \end{cases} \tag{5.13}$$

La **funzione di distribuzione uniforme** ha la seguente espressione

$$F(x) = P(X \le x) = \begin{cases} 0 & x \le a \\ \dfrac{x-a}{b-a} & a < x < b \\ 1 & x \ge b \end{cases} \tag{5.14}$$

Come esempio, si riportano nella figura 23 i grafici di $f(x)$ e $F(x)$ nel caso $a = 2$, $b = 4$.

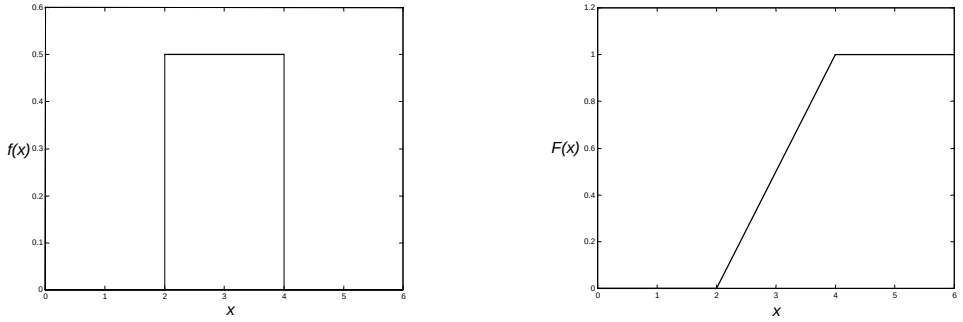


Figura 23

Proprietà 2
Il **valor medio** e la **varianza** della distribuzione uniforme continua sono dati da

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12} \tag{5.15}$$

Infatti si ha

$$\mu = \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}$$

$$\sigma^2 = \int_a^b \frac{x^2}{b-a} dx - \mu^2 = \frac{x^3}{3(b-a)}\Big|_a^b - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$$

Esempio 24
Una variabile aleatoria X è distribuita uniformemente nell'intervallo (0,100).
a – Calcolare la probabilità $P(20 < X < 60)$;
b – calcolare la media μ e la varianza σ^2 e trovare la probabilità $P(|X - \mu| < \sigma)$.

La variabile X ha la distribuzione uniforme (figura 24)

$$f(x) = \begin{cases} \dfrac{1}{100} & 0 < x < 100 \\ 0 & \text{altrimenti} \end{cases}$$

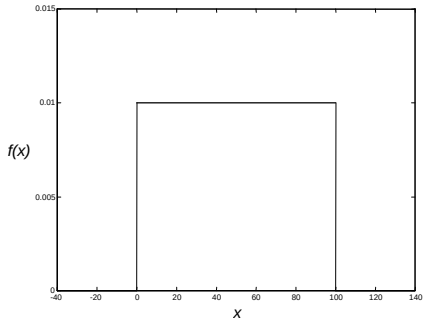


Figura 24

La funzione di distribuzione è (figura 25)

$$F(x) = \begin{cases} 0 & x \le 0 \\ \dfrac{x}{100} & 0 < x < 100 \\ 1 & x \ge 100 \end{cases}$$

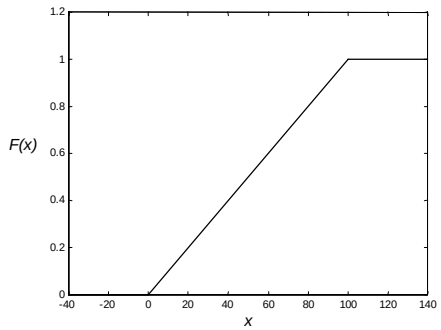


Figura 25

a – $P(20 < X < 60) = F(60) - F(20) = \frac{60}{100} - \frac{20}{100} = 0.4$

b – $\mu = 50$

$$\sigma^2 = \frac{100^2}{12} \qquad \sigma = \frac{100}{\sqrt{12}} = \frac{50}{\sqrt{3}}$$

$$P\left(|X - 50| < \frac{50}{\sqrt{3}}\right) = P\left(50 - \frac{50}{\sqrt{3}} < X < 50 + \frac{50}{\sqrt{3}}\right) =$$

$$= F\left(50 - \frac{50}{\sqrt{3}}\right) - F\left(50 + \frac{50}{\sqrt{3}}\right) = \frac{50 + \frac{50}{\sqrt{3}}}{100} - \frac{50 - \frac{50}{\sqrt{3}}}{100} = \frac{1}{\sqrt{3}} \cong 0.577$$

Esempio 25

In certi esperimenti l'errore commesso nella determinazione della solubilità di una sostanza è una variabile aleatoria X avente distribuzione uniforme con $a = -0.025$ e $b = 0.025$.
Trovare la probabilità che l'errore
a – sia compreso fra 0.010 e 0.015;
b – sia compreso fra -0.012 e 0.012.

a – La variabile X ha la seguente distribuzione uniforme (figura 26)

$$f(x) = \begin{cases} \frac{1}{0.05} = 20 & -0.025 < x < 0.025 \\ 0 & \text{altrimenti} \end{cases}$$

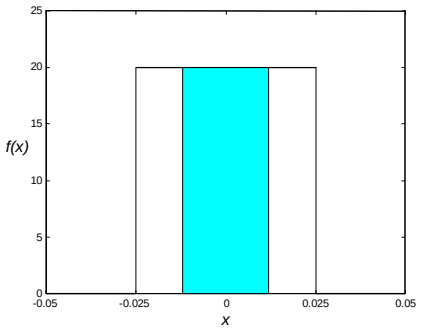


Figura 26

La funzione di distribuzione è la seguente

$$F(x) = \begin{cases} 0 & x \le 0.025 \\ \frac{x + 0.025}{0.05} & 0.025 < x < 0.025 \\ 1 & x \ge 0.025 \end{cases}$$

$$P(0.010 < X < 0.015) = F(0.015) - F(0.010) =$$

$$= \frac{0.015 + 0.025}{0.05} - \frac{0.010 + 0.025}{0.05} = 0.1$$

$$P(-0.12 < X < 0.012) = F(0.012) - F(-0.012) =$$

$$= \frac{0.012 + 0.025}{0.05} - \frac{-0.012 + 0.025}{0.05} = 0.48$$

Questi risultati possono anche essere ottenuti per via geometrica; ad esempio la probabilità $P(-0.012 < X < 0.012)$ può essere ottenuta calcolando l'area del rettangolo ombreggiato nella figura 26.

Esempio 26

La variabile aleatoria X è distribuita uniformemente nell'intervallo (a,b); sapendo che $P(X<3)=\frac{1}{4}$ e $P(X<7)=\frac{3}{4}$, calcolare a e b.

La distribuzione uniforme della variabile X è la seguente (figura 27)

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a<x<b\\ 0 & \text{altrimenti}\end{cases}$$

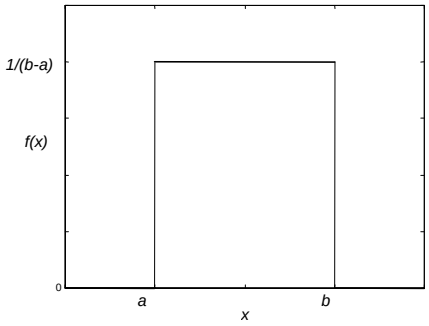


Figura 27

Dai valori delle probabilità assegnate si deduce subito che deve essere $a < 3$ e $b > 7$.

La probabilità $P(X < 3)$ è uguale all'area del rettangolo di base $3-a$ e altezza $\frac{1}{b-a}$; analogamente la probabilità $P(X < 7)$ è uguale all'area del rettangolo di base $7-a$ e altezza $\frac{1}{b-a}$; si ottiene il sistema

$$\begin{cases}(3-a)\dfrac{1}{b-a}=\dfrac{1}{4}\\ (7-a)\dfrac{1}{b-a}=\dfrac{3}{4}\end{cases}$$

Risolvendo il sistema si ricava

$$a=1 \qquad b=9$$

$$f(x)=\begin{cases}\dfrac{1}{8} & 1<x<9\\ 0 & \text{altrimenti}\end{cases}$$

6. Teoria elementare dei campioni

6.1 Popolazioni e campioni

Come già detto, l'uso del termine popolazione in statistica deriva dai tempi in cui la statistica veniva usata per fenomeni demografici o economici.
Per **popolazione** si intende oggi un insieme o collezione di oggetti, numeri, misure o osservazioni, che sono oggetto di studio. Per **campione** si intende invece una parte della popolazione, che viene selezionata per l'analisi.
Si supponga ad esempio che il preside della facoltà voglia condurre un sondaggio per conoscere il parere degli studenti sull'organizzazione dei corsi: la popolazione è composta in questo caso da tutti gli studenti iscritti, mentre il campione consiste dei soli studenti selezionati per partecipare al sondaggio.
Lo scopo del sondaggio è descrivere alcune caratteristiche dell'intera popolazione e questo viene fatto utilizzando le informazioni che si ottengono sulla base del campione di studenti.
Una popolazione può essere **finita** o **infinita**; ad esempio la popolazione costituita da tutti i bulloni prodotti in una fabbrica in un dato giorno è finita; la popolazione costituita da tutte le possibili uscite T o C in successivi lanci di una moneta è infinita.
Le popolazioni sono spesso descritte dalle **distribuzioni** dei loro valori ed è comune riferirsi alle popolazioni in termini delle loro distribuzioni.
Per popolazioni finite si fa riferimento alla distribuzione effettiva dei valori, detta distribuzione di frequenza; per popolazioni infinite alla corrispondente distribuzione di probabilità o densità di probabilità. Ad esempio un campione costituito da un certo numero di lanci di una moneta proviene da una popolazione binomiale; un campione di misure di dati proviene invece da una popolazione normale. Quindi per popolazione $f(x)$ si intende una popolazione i cui elementi hanno una distribuzione o densità di probabilità $f(x)$.
Uno degli aspetti principali della **statistica inferenziale** consiste nel trarre delle conclusioni sui parametri di una popolazione utilizzando i corrispondenti valori campionari.
La necessità di ricorrere ai metodi della statistica inferenziale deriva dalla necessità del campionamento: se la popolazione è infinita, è impossibile osservarne tutti i valori, ma anche quando è finita, questo può essere non pratico o antieconomico.
Le ragioni per cui la ricerca viene effettuata per campione, piuttosto che attraverso una rilevazione totale, sono principalmente le seguenti:
1 – l'estrazione di un campione richiede meno tempo rispetto all'esame dell'intera popolazione;
2 – un campione è meno costoso;
3 – un campione è più pratico da gestire;
4 – a volte l'esame dell'intera popolazione è impossibile: ad esempio è letale estrarre tutto il sangue di un paziente per effettuare il conteggio dei globuli rossi!
5 – qualche volta è disponibile solo un piccolo campione di dati, e non per motivi economici. Si pensi ad esempio ad un antropologo che vuole provare una certa teoria riguardante una popolazione oggi quasi estinta ed ha a disposizione solo gli ultimi sopravvissuti, 1000 persone che vivono in una certa isola: la dimensione del campione è fissata dalla natura e non dalle risorse finanziarie.
Si usa perciò un **campione**, e si traggono da esso, ossia si inferiscono, risultati riguardanti l'intera popolazione. La **teoria dei campioni** è lo studio delle relazioni esistenti tra una popolazione ed i campioni estratti da essa.
Tale teoria si applica ad esempio per ottenere la **stima dei parametri** ignoti di una popolazione, come la media μ o la varianza σ^2, quando si conoscono i valori corrispondenti del campione, media $\bar{x}$ e varianza s^2, detti **statistiche**; o anche per stabilire se ad esempio le differenze osservate tra due campioni possono essere dovute al caso o se sono significative: le risposte a questo tipo di quesito implicano l'uso dei **test di ipotesi**.

Il **calcolo delle probabilità** è l'anello di congiunzione, perché permette di determinare con quale probabilità i risultati provenienti dal campione riflettono i risultati ottenibili dall'intera popolazione.

6.2 Campionamento

Affinché le conclusioni della teoria dei campioni siano valide, i campioni devono essere scelti in modo da essere rappresentativi della popolazione.
Nel caso dei sondaggi elettorali ad esempio, la proporzione campionaria dei voti per un dato partito può essere scarsamente rappresentativa della proporzione della popolazione per uno o entrambi dei seguenti motivi:
1 – per quanto il comportamento sia stato corretto e la procedura di campionamento adeguata, è possibile essere stati così sfortunati da estrarre un campione a maggioranza favorevole a un certo partito da una popolazione favorevole invece ad un altro;
2 – il campionamento può essere stato condotto in modo scorretto o errato.
Ad esempio nel campionare una popolazione di votanti è un errore ricavare i loro nomi da un elenco telefonico, perché verrebbero ad essere mal rappresentati i votanti che non dispongono del telefono o che per motivi personali non vogliono comparire nell'elenco.
Ci sono fondamentalmente due tipi di campioni:
– **campioni non probabilistici**;
– **campioni probabilistici**.
Un **campione non probabilistico** è un campione in cui gli individui vengono scelti senza tenere conto della probabilità di ciascun individuo di appartenere al campione.
Un **campione probabilistico** è un campione in cui gli individui vengono scelti tenendo conto della probabilità nota di ciascun individuo di essere scelto per far parte del campione.
Siccome nei campioni non probabilistici gli individui sono scelti senza conoscere la loro probabilità di selezione (e in alcuni casi si autoselezionano), la teoria sviluppata per il campionamento probabilistico non può essere applicata.
Ad esempio molte aziende conducono sondaggi dando ai visitatori del loro sito Web la possibilità di compilare dei questionari e inviarli elettronicamente.
Le risposte a questi sondaggi possono fornire molti dati velocemente, ma il campione si compone di utilizzatori di Internet che si autoselezionano.
I campioni non probabilistici possono avere alcuni vantaggi: comodità, velocità di estrazione, costi bassi; d'altro lato hanno degli svantaggi: mancanza di accuratezza dovuta alla selezione distorta, impossibilità di generalizzare i risultati. Spesso gli svantaggi compensano ampiamente i vantaggi. Di solito i campioni non probabilistici si usano per ottenere indicazioni grezze e a basso costo, o per piccoli studi pilota, che saranno successivamente seguiti da indagini più rigorose.
Il campionamento probabilistico deve essere usato ogni qual volta sia possibile, perché è il solo metodo che consente di ottenere inferenze corrette sulla base di un campione.
I tipi di campionamento probabilistico più usati sono:
– **campionamento casuale semplice**;
– **campionamento sistematico**;
– **campionamento stratificato**;
– **campionamento a grappolo**.
Questi tipi differiscono fra loro per il costo, l'accuratezza e la complessità.

Campionamento casuale semplice
E' la più semplice tecnica di selezione di un campione; il procedimento è sostanzialmente simile allo schema di estrazione da un'urna.
Un **campione casuale semplice** è un campione in cui ogni individuo della popolazione ha la stessa probabilità di essere scelto; inoltre campioni della stessa dimensione hanno tutti la stessa probabilità di essere selezionati.

Nel campionamento casuale semplice si indica con ***n*** la **dimensione del campione**, ossia il numero di elementi del campione, e con ***N*** la **dimensione della popolazione**, ossia il numero di elementi della popolazione.

La probabilità che ogni individuo della popolazione ha di essere scelto alla prima estrazione è $\frac{1}{N}$.

La selezione del campione può essere fatta in due modi:
– **con reimmissione**;
– **senza reimmissione**.
Nel **campionamento con reimmissione** ciascun elemento della popolazione è disponibile ad ogni estrazione, quindi ad ogni estrazione ogni individuo ha sempre probabilità $\frac{1}{N}$ di essere estratto.
In questo modo un individuo può essere nuovamente estratto in una successiva estrazione.

Esempio 1
In un'urna si introducono N biglietti con il nome di N persone diverse. Alla prima estrazione si estrae il nome Paolo Rossi; la probabilità di essere estratto è $\frac{1}{N}$. Se si rimette il biglietto nell'urna, alla seconda estrazione Paolo Rossi ha la stessa probabilità $\frac{1}{N}$ degli altri di essere estratto. Il processo si ripete alle successive estrazioni.

In genere però si preferisce avere campioni composti di individui diversi, per non ripetere misure o prove sullo stesso individuo.
Nel **campionamento senza reimmissione** un individuo, una volta selezionato, non viene rimesso nella popolazione e non può più essere scelto di nuovo.
Nell'esempio precedente la probabilità che Paolo Rossi venga estratto alla prima estrazione è ancora $\frac{1}{N}$; alla successiva estrazione, poiché Paolo Rossi (già estratto) non è più presente nella popolazione, la probabilità che un altro individuo venga scelto è $\frac{1}{N-1}$. Allo stesso modo si prosegue per le successive estrazioni.
In questo modo però gli individui non hanno tutti la stessa probabilità di essere estratti, perché si altera la composizione dell'urna dopo ogni estrazione.
Generalmente nelle applicazioni il campione è estratto senza reimmissione, per i motivi già detti poco sopra. Questo ha delle conseguenze, che saranno esaminate nel seguito.
Una popolazione finita nella quale si compie un campionamento con reimmissione può essere considerata infinita, poiché si può estrarre un numero qualsiasi di campioni senza esaurire la popolazione.
Indipendentemente dal fatto che il campione sia estratto con o senza reimmissione, la tecnica di estrazione dall'urna non è concretamente praticabile.
Di solito per scegliere il campione si usa una tecnica basata sulle **tavole dei numeri casuali**, di cui si riproduce un esempio nella pagina seguente. Queste tavole si compongono di una serie di cifre da 0 a 9, generate casualmente simulando l'estrazione a sorte da un'urna, in modo che ogni cifra abbia la stessa probabilità di essere estratta, ed elencate nell'ordine secondo cui sono state generate.
Poiché il sistema decimale ha 10 cifre (le cifre 0,1,2,3,…,9), queste hanno tutte la stessa probabilità $\frac{1}{10}$ di essere generate casualmente.
Le cifre sono riunite in gruppi di cinque per facilitare la lettura; poiché tutte le cifre o successioni di cifre nella tavola sono casuali, si può leggere sia in senso orizzontale che verticale, dall'alto o dal basso, specificando però prima di iniziare a usare la tavola il criterio scelto; bisogna inoltre scegliere un punto di partenza nella tavola dei numeri casuali (ad esempio puntando a caso con una

matita a occhi chiusi o scegliendo a caso una riga e una colonna della tabella e una delle cinque cifre della casella).
Per usare la tavola si assegna ad ogni elemento della popolazione un codice numerico, ad esempio si fa una lista numerata; si può ottenere un campione leggendo la tavola dei numeri casuali e selezionando gli individui della lista il cui codice coincide con il numero casuale.

TAVOLA DI NUMERI CASUALI

	COLONNA							
RIGA	01	02	03	04	05	06	07	08
01	49280	88924	35779	00283	81163	07275	89863	02348
02	61870	41657	07468	08612	98083	97349	20775	45091
03	43898	65923	25078	86129	78496	97653	91550	08078
04	62993	93912	30454	84598	56095	20664	12872	64647
05	33850	58555	51438	85507	71865	79488	76783	31708
06	9734**0**	**03364**	**88472**	**04334**	**63919**	36394	11095	92470
07	70543	29776	10087	10072	55980	64688	68239	20461
08	89382	93809	00796	95945	34101	81277	66090	88872
09	37818	72142	67140	50785	22380	16703	53362	44940
10	60430	22834	14130	96593	23298	56203	92671	15925
11	82975	66158	84731	19436	55790	69229	28661	13675
12	39087	71938	40355	54324	08401	26299	49420	59208
13	55700	24586	93247	32596	11865	63397	44251	43189
14	14756	23997	78643	75912	83832	32768	18928	57070
15	32166	53251	70654	92827	63491	04233	33825	69662
16	23236	73751	31888	81718	06546	83246	47651	04877
17	45794	26926	15130	82455	78305	55058	52551	47182
18	09893	20505	14225	68514	46427	56788	96297	78822
19	54382	74598	91499	14523	68479	27686	46162	83554
20	94750	89923	37089	20048	80336	94598	26940	36858
….	….	….	….	….	….	….	….	….

Esempio 2
Da una popolazione di 800 persone si vuole formare un campione di 30 persone.
Si scrive un elenco nominativo delle 800 persone e ad ogni persona si assegna un codice. Dato che la dimensione della popolazione (N = 800) è un numero di tre cifre, ciascun codice deve avere tre cifre, perciò al primo individuo dell'elenco si assegna il codice 001, al secondo il codice 002, ecc.., all'ultimo il codice 800.
Scelto un punto di partenza nella tavola (ad esempio riga 06, colonna 05), si leggono da sinistra verso destra le sequenze di tre cifre senza saltarne nessuna:

003 364 884 720 433 463 ecc.

Il codice 884 non c'è nella lista, perciò si scarta. Le prime persone scelte sono quelle con i codici 003, 364, 720, 433, 463,....
Si continua fino ad ottenere un campione di 30 persone. Se un codice di tre cifre si ripete, la persona corrispondente viene di nuovo inclusa nel campione, se il campionamento avviene con reimmissione; se invece si campiona senza reimmissione, si continua nella scelta di un ulteriore individuo, fino a raggiungere la dimensione richiesta di 30 persone.

In pratica solo raramente si ha l'opportunità di numerare ogni individuo della popolazione, in modo da scegliere un campione casuale con la tecnica sopra descritta, e si deve spesso assumere che il campione scelto abbia tutte le proprietà di un campione casuale, senza che sia stato costruito formalmente in questo modo. Nell'esempio la popolazione è finita e non troppo grande e può essere abbastanza facilmente numerata con il codice, ma in molti casi è infinita o troppo grande e non può essere numerata.

La maggior parte dei computer dispongono di un generatore di numeri casuali, da usare in alternativa alle tavole. In realtà i numeri "casuali" generati dalla maggior parte dei computer sono "pseudocasuali", perché sono il risultato di una qualche formula deterministica. Tuttavia soddisfano la maggior parte degli scopi pratici.

Campionamento sistematico
Un altro tipo di campionamento è il campionamento sistematico. In questo caso si procede nel modo seguente. Data la popolazione di N individui e fissata la dimensione n del campione, si calcola il quoziente intero $R = \frac{N}{n}$.
Si sceglie un numero k a caso (ad esempio da un'urna) compreso fra 1 e R; si includono nel campione gli individui della lista che occupano i posti k, $k + R$, $k + 2R$, ...

Esempio 3
Da una popolazione di 1000 individui si vuole formare un campione di 50 individui; in questo caso

$$N = 1000 \qquad n = 50 \qquad R = \frac{1000}{50} = 20$$

Si sceglie un numero k a caso fra 1 e 20, sia ad esempio $k = 15$.
Il campione sarà formato dagli elementi della lista che portano il numero 15, 35, 55, 75, ...

Se l'elenco di tutti gli individui della popolazione è fatto in modo casuale, anche il campione sarà casuale. Se invece l'elenco non è casuale rispetto alla variabile che si vuole studiare, il campione estratto può essere distorto.
Il campionamento sistematico è più facile da eseguire, ma il suo uso acritico può portare con facilità a campioni affetti da errori sistematici; questo rischio non c'è con il campionamento casuale semplice.
In generale i risultati di un campionamento sistematico dipendono in larga misura dalle caratteristiche dell'indagine che si vuole fare e dalla popolazione da cui si campiona.

Esempio 4
Si vuole effettuare un'indagine sulle abitudini alimentari di una popolazione di 100.000 ragazzi di 10 anni, scegliendone un campione di 3000: si possono prendere i nati in un dato giorno del mese di un anno fissato.
Se però si volesse usare lo stesso campione per studiare il quoziente di intelligenza, questo campione sarebbe distorto, perché il quoziente di intelligenza, come il campione, è influenzato dall'età.

Esempio 5
Si pensa di studiare l'inquinamento atmosferico in una città per un certo periodo di tempo (un anno). Se si effettuano i prelievi alle ore 8, 12, 16, 20 di ogni giorno si potrà avere un'idea distorta dell'inquinamento giornaliero (sovrastima), in quanto queste sono le ore di maggior traffico, attività industriale, commerciale, quindi si avrebbero livelli sistematicamente maggiori di quelli risultanti da prelievi in altre ore della giornata.

La scelta casuale, semplice o sistematica, presenta inconvenienti quando l'indagine è di vasta mole; in questi casi si usano altri tipi di campionamento.

Campionamento stratificato
Un altro tipo di campionamento è il campionamento stratificato. E' una delle tecniche di campionamento più famose e usate; consiste nel dividere gli N individui della popolazione in sottopopolazioni, o **strati**, sulla base di una caratteristica comune; nell'estrarre poi un campione casuale semplice da ogni strato in modo indipendente, e nel riunire insieme i risultati dei singoli campionamenti per formare un unico campione dell'ampiezza richiesta.

Questo metodo è più efficace perché assicura che gli individui della popolazione siano rappresentati adeguatamente nel campione; questo garantisce una maggior precisione nelle stime dei parametri della popolazione.
Il ricorso alla stratificazione presuppone che si abbiano delle conoscenze sulla popolazione, in modo da poterla suddividere in strati, ad esempio classi di età, classi di reddito, ecc.
La stratificazione consente di aumentare la precisione delle stime, senza comportare un aumento del numero totale di elementi del campione.
Infatti la bontà dei risultati di un'indagine campionaria dipende essenzialmente da due fattori:
– dimensione del campione;
– variabilità del fenomeno in esame.
Quindi per aumentare la precisione dei risultati si può agire aumentando la dimensione del campione, con conseguente aumento dei costi; se si pone il vincolo sul numero di elementi del campione, l'unica possibilità per aumentare la significatività dei risultati della rilevazione è utilizzare un campionamento stratificato.

Esempio 6
Studio dell'incidenza di una data patologia, che è influenzata dall'età, in un gruppo di *N* individui.
Con un campionamento semplice può accadere che il campione sia composto prevalentemente da giovani o da anziani.
Se anziché applicare il campionamento casuale semplice all'intera popolazione, si procede prima a una stratificazione degli individui secondo tre grandi classi di età (giovani, adulti, anziani) e poi si attua un campionamento semplice nell'ambito di ciascuna classe, si ha la certezza che tutte e tre le categorie entrino a far parte del campione in modo equilibrato.
L'ampiezza del campione in ogni strato (non tutti gli strati hanno la stessa numerosità) può essere stabilità in vari modi diversi.

Campionamento a grappolo
Nel campionamento a grappolo, gli *N* individui nella popolazione sono suddivisi in molti gruppi, detti **grappoli** (sottopopolazioni), in modo tale che ogni grappolo sia rappresentativo dell'intera popolazione.
Si estrae poi un campione casuale di grappoli e tutti gli individui di ciascuno dei grappoli selezionati sono inclusi nel campione.
I grappoli possono essere definiti sulla base di raggruppamenti naturali, come quelli determinati dalle regioni, dalle città, dalle circoscrizioni elettorali, dai quartieri urbani, dagli edifici o dalle famiglie.
Il campionamento a grappolo può essere meno costoso del campionamento casuale semplice, soprattutto quando la popolazione sottostante è disseminata su una vasta area geografica.
Comunque, il campionamento a grappolo tende a essere meno efficiente sia del campionamento casuale semplice, che del campionamento stratificato, e si rende necessaria una dimensione complessiva del campione più grande per ottenere risultati precisi come quelli che si ottengono con altri procedimenti.

Studi clinici sperimentali e randomizzazione
Il campionamento casuale è il metodo che consente di ottenere campioni rappresentativi da una popolazione. Lo stesso principio di deve applicare anche in esperimenti in cui il caso deve operare a un livello un po' diverso.
Se si devono sperimentare diverse dosi di un farmaco su una popolazione di cavie, o nuove medicine su una popolazione di malati, la selezione casuale non si applica alla scelta delle cavie o dei malati da mettere in esperimento, ma si usa più oltre, nella loro distribuzione fra i gruppi sperimentali.
Infatti ad esempio potremo preferire cavie non scelte a caso, ma tutte di ugual peso, per ottenere maggior omogeneità nel confronto degli effetti di dosi diverse di un farmaco.
Potremmo essere interessati ad eseguire il confronto fra un nuovo farmaco e un farmaco già noto, limitando l'esame ai malati di una forma particolare di una data malattia.

Occorre in ambedue gli esempi che i gruppi assegnati ai diversi trattamenti sperimentali (le diverse dosi del farmaco nel primo esempio, il nuovo e il vecchio tipo di terapia nel secondo esempio) siano confrontabili, cioè diversi solo per errore di campionamento casuale; bisogna cioè essere sicuri che l'assegnazione delle cavie o dei malati a ciascun gruppo avvenga con un criterio realmente casuale.
Il caso si deve applicare non alla scelta iniziale delle cavie o dei malati, ma alla distribuzione delle cavie scelte per l'esperimento nei vari gruppi, o all'assegnazione dei malati in esame alle due terapie: si parla di **randomizzazione**.
La randomizzazione è l'unico metodo che ci permette di prevedere l'entità degli errori dovuti al caso.
Se non si usa questo sistema di assegnazione non si può essere sicuri che al caso, che non si può eliminare, non si aggiunga qualche altro fattore che può falsare le conclusioni o alterare i risultati.
Nell'esempio delle cavie, si abbiano 30 cavie da distribuire in tre gruppi da 10; se non si procede alla randomizzazione e si prelevano le prime 10 da assegnare al primo gruppo, le seconde 10 da assegnare al secondo gruppo e le ultime da assegnare all'ultimo gruppo, il risultato sarà che le cavie più lente o meno agili o in peggior stato di salute siano più facilmente collocate nel primo gruppo, quelle più sveglie o più sane nell'ultimo gruppo: in tal modo un'eventuale azione del farmaco può essere confusa con effetti dovuti allo stato di salute, il che può alterare i risultati sperimentali.
In generale è possibile individuare due grandi settori in cui si possono svolgere analisi di tipo medico e di ricerca: un primo settore di indagine, osservazione e intervento sulla popolazione, un secondo settore di tipo clinico-sperimentale.
Ciascuno dei due settori presenta una complessità di approccio tale da richiedere necessariamente metodologie sofisticate per la rilevazione, il controllo, e l'analisi dei modi di procedere e dei dati.
Le metodologie proprie dell'epidemiologia[1] e della statistica mettono in luce gli aspetti quantitativi e la necessità di una formalizzazione matematica.
Nell'ambito degli studi di tipo medico epidemiologico si possono distinguere **studi osservazionali** e **studi sperimentali**.
In uno studio osservazionale il ricercatore medico non può controllare direttamente le condizioni sotto cui avviene lo studio, limitandosi ad osservare ad esempio la presenza o assenza di una data malattia in un campione di soggetti, o l'incidenza di una data malattia, ossia il numero di nuovi casi, in una popolazione a rischio.
In un contesto sperimentale invece, è possibile specificare le condizioni sotto cui lo studio sarà condotto e pertanto è possibile assegnare i soggetti ai diversi gruppi di studio con la tecnica di randomizzazione: in questo risiede sostanzialmente la validità di questo tipo di studi sperimentali.
Tra gli studi sperimentali si possono distinguere sperimentazioni cliniche, sperimentazione umane e su animali.
Esempi di **sperimentazioni cliniche** sono:
– le sperimentazioni terapeutiche, nelle quali una procedura terapeutica, clinica o chirurgica vengono adottate nel tentativo di alleviare i sintomi e/o migliorare la guarigione o la sopravvivenza di coloro che presentano la malattia;
– le sperimentazioni di intervento, nelle quali il ricercatore interviene prima che la malattia si sia sviluppata su individui ritenuti a rischio, ad esempio la somministrazione di farmaci anti-ipertensivi per ridurre il rischio di sviluppare un ictus;
– le sperimentazioni preventive, nelle quali si tenta di determinare l'efficacia di una procedura preventiva, ad esempio un vaccino.
In uno studio clinico sperimentale, l'esperimento pianificato coinvolge dei pazienti ed è progettato per valutare l'efficacia di un dato trattamento: la somministrazione di un farmaco a una certa dose, un tipo di metodica chirurgica, una cura dietetica, un tipo di strumento per prelievi, ecc.
Oggetto di studio può essere anche una misura di sanità pubblica o una maniera di praticare la stessa terapia, ad esempio ambulatoriale o in ospedale, in anestesia generale o locale.

[1] L'epidemiologia ha come oggetto di studio la distribuzione delle malattie in una popolazione umana e i fattori che le influenzano.

La caratteristica del metodo sperimentale, che lo distingue dagli altri studi, è che, come già osservato, il ricercatore può intervenire determinando il trattamento da applicare a un gruppo di pazienti, ne osserva gli esiti e li confronta con quelli di un altro gruppo di pazienti che hanno ricevuto un altro trattamento, oppure niente (placebo[2]); questo non accade con gli studi osservazionali, nei quali il ricercatore si limita ad osservare le conseguenze di un trattamento terapeutico avvenuto prima e al di fuori e non determinato e pianificato nello studio.
In ogni caso si effettua un controllo fra due o più gruppi di pazienti; i pazienti, che devono essere rappresentativi di una popolazione che il ricercatore vuole studiare, vengono reclutati dallo sperimentatore in base a date caratteristiche, che dipendono dallo scopo e dagli obiettivi dello studio.
I pazienti che presentano le caratteristiche stabilite costituiscono un campione rappresentativo della popolazione e entrano a far parte della sperimentazione; l'assegnazione dei pazienti del campione ai diversi gruppi di trattamento è un problema cruciale.
La randomizzazione, ossia l'attribuzione casuale dei pazienti ai diversi gruppi, è lo strumento con il quale il ricercatore evita di introdurre distorsioni conscie e inconscie nel processo di suddivisione degli individui, rendendo confrontabili i gruppi fra loro.
Si osservi che la conoscenza da parte del paziente e/o del medico del trattamento impiegato, può condizionare, anche in modo inconsapevole, la valutazione dei risultati e il risultato stesso.
Quando l'assegnazione al trattamento non è nota al paziente, ma è nota al medico, si parla di cecità semplice (**single blind**); quando anche il medico ignora il trattamento assegnato (cosa che in qualche caso è impossibile, ad esempio nel caso di una tecnica chirurgica) si parla di doppia cecità (doppio cieco, **double blind**); può anche accadere che anche chi analizza i risultati non conosca né quali pazienti, né quali trattamenti sono interessati nello studio (triplo cieco).
La cecità elimina, nei casi in cui può essere adottata, le distorsioni presenti nella risposta del paziente, nella valutazione del medico e nelle cure accessorie.
Attualmente lo standard ritenuto migliore nella ricerca clinica è lo studio randomizzato con la tecnica del doppio cieco.

6.3 Distribuzioni di campionamento

Consideriamo tutti i possibili campioni casuali di ampiezza n che possono essere estratti da una data popolazione, con o senza reimmissione. Per ciascun campione si può calcolare una data statistica, come la media, la varianza o lo scarto quadratico medio, che potrà variare da campione a campione. In tal modo otteniamo una distribuzione della statistica, detta distribuzione di campionamento della statistica stessa.
Se ad esempio la statistica usata è la media, la distribuzione è detta **distribuzione della media campionaria**.

Definizione
Si definisce **distribuzione di campionamento** di una data statistica la distribuzione di tutti i possibili valori che possono essere assunti dalla statistica stessa, calcolati da campioni casuali della stessa dimensione estratti dalla stessa popolazione.

Le distribuzioni di campionamento permettono di risolvere problemi di tipo probabilistico su statistiche campionarie, ma soprattutto forniscono gli strumenti teorici per la trattazione dell'inferenza statistica; tali distribuzioni possono essere costruite quando si campiona da una popolazione finita e discreta, procedendo nel modo seguente:
1 – da una popolazione finita di dimensione N si estraggono tutti i possibili campioni casuali di ampiezza n;
2 – si calcola la statistica di interesse per ogni campione;

[2] In questo caso possono sorgere problemi etici.

3 – si costruisce una tabella contenente i vari valori distinti assunti dalla statistica e le corrispondenti frequenze. Il procedimento è illustrato dal seguente esempio 7.
La costruzione effettiva di una distribuzione di campionamento è un lavoro impegnativo se la popolazione è grande, ed è impossibile se la popolazione è infinita. Tali distribuzioni possono però essere derivate matematicamente, con procedimenti che non saranno trattati in modo dettagliato in queste lezioni.
Le caratteristiche importanti di una distribuzione di campionamento, a cui siamo interessati, sono la sua media, la sua varianza e la sua forma.

6.4 Distribuzione della media campionaria (varianza σ^2 nota)

Un'importante distribuzione di campionamento è quella della media campionaria; per studiare questa distribuzione si ragiona nel seguente modo.
Si estrae un primo campione casuale di *n* elementi da una data popolazione, e si indica con $\bar{x}_1$ la sua media; se si estrae un secondo campione di *n* elementi dalla stessa popolazione, si ottiene un altro valore per la media $\bar{x}_2$, di solito diverso dal precedente; se si estraggono successivamente altri campioni, i valori delle medie saranno in generale diversi fra loro.
I valori delle medie possono essere visti come i valori assunti da una variabile aleatoria $\overline{X}$, detta **media campionaria**, su tutti i possibili campioni di ampiezza *n* che possono essere estratti dalla popolazione. La differenza fra i valori delle medie è dovuta al caso, e questo fatto suggerisce di studiare la distribuzione di tali valori.
Illustriamo con un esempio la costruzione della distribuzione della media campionaria nel caso di una popolazione finita di dimensione piccola.

Esempio 7
Si consideri una popolazione finita, costituita da $N = 4$ elementi, e avente la seguente distribuzione uniforme discreta

x_i	1	2	3	4
$f(x_i)$	0.25	0.25	0.25	0.25

Tabella 1

La media μ e la varianza σ^2 di questa popolazione sono

$$\mu = \frac{1+2+3+4}{4} = 2.5$$

$$\sigma^2 = 1\cdot\frac{1}{4} + 4\cdot\frac{1}{4} + 9\cdot\frac{1}{4} + 16\cdot\frac{1}{4} - (2.5)^2 = 1.25$$

Consideriamo tutti i possibili campioni di dimensione $n = 2$ estraibili da questa popolazione; quando il campionamento avviene con reimmissione, i campioni di ampiezza 2 sono in numero di $4^2 = 16$; tali campioni sono elencati nella tabella 2, insieme con le corrispondenti medie.

Campioni	*Medie*	*Campioni*	*Medie*
(1,1)	1	(3,1)	2
(1,2)	1.5	(3,2)	2.5
(1,3)	2	(3,3)	3
(1,4)	2.5	(3,4)	3.5
(2,1)	1.5	(4,1)	2.5
(2,2)	2	(4,2)	3
(2,3)	2.5	(4,3)	3.5
(2,4)	3	(4,4)	4

Tabella 2

Nella tabella 3 è riportata la distribuzione della media campionaria, ottenuta elencando i diversi valori della media campionaria nella prima riga e le rispettive frequenze nella seconda riga.

$\bar{x}_i$	1	1.5	2	2.5	3	3,5	4
$f(\bar{x}_i)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Tabella 3

Nella figura 1 rappresentiamo la distribuzione della popolazione; nella figura 2 rappresentiamo invece la distribuzione della media campionaria.

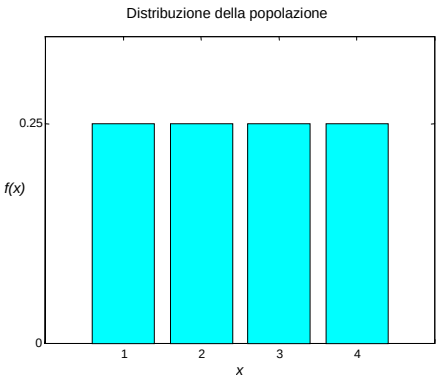


Figura 1

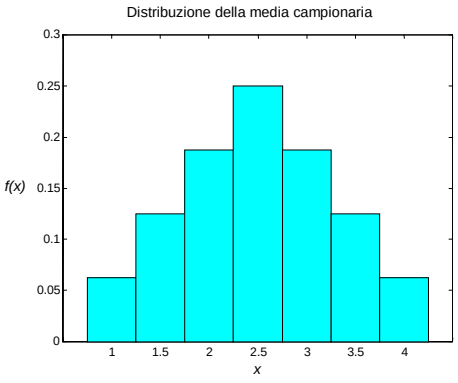


Figura 2

Gli istogrammi mostrano che la distribuzione della media campionaria ha una forma a campana, simile a una distribuzione normale, anche se la popolazione ha la distribuzione uniforme.
Calcoliamo la media della distribuzione della media campionaria

$$\mu_{\bar{X}} = 1\cdot\frac{1}{16} + 1.5\cdot\frac{2}{16} + 2\cdot\frac{3}{16} + 2.5\cdot\frac{4}{16} + 3\cdot\frac{3}{16} + 3.5\cdot\frac{2}{16} + 4\cdot\frac{1}{16} = 2.5$$

Questa media è uguale alla media della popolazione.
Calcoliamo infine la varianza della distribuzione della media campionaria

$$\sigma^2_{\bar{X}} = 1\cdot\frac{1}{16} + (1.5)^2\cdot\frac{2}{16} + (2)^2\cdot\frac{3}{16} + (2.5)^2\cdot\frac{4}{16} + (3)^2\cdot\frac{3}{16}$$
$$+ (3.5)^2\cdot\frac{2}{16} + (4)^2\cdot\frac{1}{16} - (2.5)^2 = 0.625$$

Questa varianza non è uguale alla varianza della popolazione, tuttavia si osserva che vale la relazione

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{1.25}{2} = 0.625$$

Se il campionamento viene fatto senza reimmissione, i campioni estraibili da questa popolazione finita costituita da 4 elementi sono soltanto 6, e sono elencati nella tabella 4; nella tabella 5 è riportata la corrispondente distribuzione della media campionaria.

Campioni	*Medie*
(1,2)	1.5
(1,3)	2
(1,4)	2.5
(2,3)	2.5
(2,4)	3
(3,4)	3.5

Tabella 4

$\bar{x}_i$	1.5	2	2.5	3	3,5
$f(\bar{x}_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tabella 5

In questo caso per la media e la varianza della distribuzione della media campionaria si ha

$$\mu_{\overline{X}} = 1.5 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 2.5 \cdot \frac{2}{6} + 3 \cdot \frac{1}{6} + 3.5 \cdot \frac{1}{6} = 2.5$$

$$\sigma^2_{\overline{X}} = (1.5)^2 \cdot \frac{1}{6} + (2)^2 \cdot \frac{1}{6} + (2.5)^2 \cdot \frac{2}{6} + (3)^2 \cdot \frac{1}{6} + (3.5)^2 \cdot \frac{1}{6} - (2.5)^2 = \frac{5}{12}$$

Osserviamo che la media della distribuzione della media campionaria è ancora uguale alla media della popolazione, mentre per la varianza si può verificare che vale la relazione

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{1.25}{2} \cdot \frac{4-2}{4-1} = \frac{1.25}{3} = \frac{5}{12}$$

Questi risultati sono validi per tutte le distribuzioni della media campionaria, ottenute con il campionamento con reimmissione o con il campionamento da popolazioni infinite, oppure ancora con il campionamento senza reimmissione da una popolazione finita.
Si possono infatti dimostrare due teoremi generali che esprimono le proprietà della **distribuzione della media campionaria**.
Il primo di essi, formalizzando quanto osservato nell'esempio precedente, fornisce delle espressioni per la media $\mu_{\overline{X}}$ e la varianza $\sigma^2_{\overline{X}}$ della distribuzione della media campionaria $\overline{X}$.
Il secondo teorema, di fondamentale importanza per l'inferenza statistica, consente di dimostrare che qualunque sia la distribuzione della popolazione da cui provengono i campioni, la distribuzione della media campionaria è legata alla distribuzione normale.

Teorema 1
Se si estraggono campioni casuali di ampiezza n da una popolazione avente media μ e varianza σ^2, allora la distribuzione della media campionaria $\overline{X}$ ha media

$$\mu_{\overline{X}} = \mu . \tag{6.1}$$

Per campioni estratti da popolazioni infinite, o se il campionamento è fatto con reimmissione, la varianza della distribuzione della media campionaria è

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} . \tag{6.2}$$

Per campioni estratti senza reimmissione da una popolazione finita di ampiezza N la varianza della distribuzione della media campionaria è

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} . \tag{6.3}$$

Lo scarto quadratico medio $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ è detto **errore standard della media**, o semplicemente **errore standard**, e rappresenta una misura quantitativa della variabilità delle medie dei campioni di ampiezza n estratti dalla popolazione avente varianza σ^2. L'errore standard decresce in proporzione alla radice quadrata di n: per esempio è necessario quadruplicare l'ampiezza del campione per dimezzare l'errore standard della distribuzione della media campionaria.
Il fattore $\frac{N-n}{N-1}$, detto **fattore correttivo per la popolazione finita**, ha un valore prossimo a 1 quando la dimensione del campione è piccola rispetto alla dimensione della popolazione; nella maggior parte delle applicazioni pratiche la correzione per popolazione finita non si usa, a meno che il campione non contenga più del 5% degli elementi della popolazione. In altre parole la correzione per popolazione finita può essere ignorata quando $\frac{n}{N} \le 0.05$ (vedere l'esempio 9).
Il teorema 1 fornisce informazioni solo parziali sulla distribuzione della media campionaria.
In generale è impossibile determinare tale distribuzione esattamente, senza conoscere l'effettiva distribuzione della popolazione; è però possibile trovare la distribuzione limite per $n \to \infty$ di una

variabile aleatoria i cui valori sono strettamente collegati ai valori di $\overline{X}$, supponendo solo che la popolazione abbia varianza σ^2 finita.
Questa variabile aleatoria è la **media campionaria standardizzata**

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Riferendoci a questa variabile, vale il teorema seguente.

Teorema 2 – Teorema del limite centrale
Sia data una popolazione avente media μ e varianza σ^2, e da essa si estraggano campioni casuali di ampiezza *n*; indicando con $\overline{X}$ la media campionaria, la variabile

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (6.4)$$

è una variabile aleatoria la cui distribuzione tende alla distribuzione normale standardizzata per $n \to \infty$.

Qualunque sia la distribuzione della popolazione, si può quindi affermare che la distribuzione della media campionaria $\overline{X}$ è approssimativamente normale con media μ e varianza $\frac{\sigma^2}{n}$, per *n* sufficientemente grande.
In pratica nella maggior parte dei casi la distribuzione normale è una buona approssimazione della distribuzione della media campionaria per $n \geq 30$, qualunque sia la distribuzione della popolazione.
Se il campione casuale proviene da una popolazione normale, la distribuzione della media campionaria è normale per ogni valore di *n* (anche minore di 30).
Riassumiamo i risultati fin qui ottenuti, riguardanti le caratteristiche della distribuzione della media campionaria, nel seguente schema.

Schema riassuntivo – Proprietà della distribuzione della media campionaria
1. Campionamento da una popolazione distribuita normalmente con media μ e varianza σ^2:
a – $\mu_{\overline{X}} = \mu$
b – $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$
c – la distribuzione della media campionaria $\overline{X}$ è normale.
2. Campionamento da una popolazione non distribuita normalmente con media μ e varianza σ^2:
a – $\mu_{\overline{X}} = \mu$
b – $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$ se $\frac{n}{N} \leq 0.05$
$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
c – la distribuzione della media campionaria è approssimativamente normale, per $n \geq 30$.

Come si vedrà nei capitoli successivi, le distribuzioni campionarie trovano la loro più importante applicazione nell'inferenza statistica.
La più semplice applicazione della distribuzione della media campionaria consiste nel calcolare la probabilità di ottenere un campione avente una certa media.

Esempio 8

La variabile aleatoria continua X ha media $\mu = 5$ e varianza $\sigma^2 = 25$. Si estrae un campione di 100 elementi da questa popolazione; determinare la probabilità che la media del campione sia maggiore di 5.4.

In base al teorema 1, la media campionaria $\overline{X}$ ha il valor medio e la varianza seguenti

$$\mu_{\overline{X}} = \mu = 5 \qquad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{25}{100} = \frac{1}{4}.$$

Applicando il teorema del limite centrale, si può affermare che la variabile $\overline{X}$ ha approssimativamente la distribuzione normale.

Per calcolare la probabilità che la media del campione sia maggiore di 5.4, occorre standardizzare la media campionaria con la formula

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 5}{\sqrt{\frac{1}{4}}}$$

$$\overline{X} = 5.4 \quad \Rightarrow \quad Z = \frac{5.4 - 5}{0.5} = 0.8$$

$$P(\overline{X} > 5.4) = P(Z > 0.8) = 1 - P(Z < 0.8) = 1 - 0.7881 = 0.2119$$

Esempio 9

I pesi di 20000 cuscinetti a sfere sono distribuiti normalmente con media $\mu = 22.4$g e scarto quadratico medio $\sigma = 0.048$g. Se da questa popolazione vengono estratti 300 campioni casuali di ampiezza 36, determinare la media e lo scarto quadratico medio della distribuzione della media campionaria nel caso che il campionamento venga fatto con reimmissione o senza reimmissione.

Determinare per quanti dei campioni casuali la media

a – è compresa fra 22.39 e 22.41;

b – è superiore a 22.42;

c – è inferiore a 22.37.

In base al teorema 1, se si effettua il campionamento con reimmissione si ottiene

$$\mu_{\overline{X}} = \mu = 22.4 \qquad \sigma_{\overline{X}} = \frac{0.048}{\sqrt{36}} = 0.008$$

Se invece si effettua il campionamento senza reimmissione, la popolazione è finita e si ottiene

$$\mu_{\overline{X}} = \mu = 22.4 \qquad \sigma_{\overline{X}} = \frac{0.048}{\sqrt{36}} \cdot \sqrt{\frac{20000 - 36}{20000 - 1}} = 0.007993$$

I due valori ottenuti per lo scarto quadratico medio sono circa uguali, dato che la popolazione è grande rispetto all'ampiezza del campione; poiché la popolazione è distribuita normalmente, la distribuzione della media campionaria è normale, con media $\mu_{\overline{X}} = 22.4$ e scarto quadratico medio $\sigma_{\overline{X}} = 0.008$.

Per risolvere i punti a, b e c, occorre standardizzare la media campionaria con la formula

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 22.4}{0.008}$$

a –
$$\overline{X} = 22.39 \quad \Rightarrow \quad Z = \frac{22.39 - 22.4}{0.008} = -1.25$$

$$\overline{X} = 22.41 \quad \Rightarrow \quad Z = \frac{22.41 - 22.4}{0.008} = 1.25$$

$$P\left(22.39 \le \overline{X} \le 22.41\right) = P(-1.25 \le Z \le 1.25) =$$

$$= 2P(Z \le 1.25) - 1 = 2 \cdot 0.8944 - 1 = 0.7888$$

Il numero di campioni atteso è $300 \cdot 0.7888 = 237$.

b – $$\overline{X} = 22.42 \quad \Rightarrow \quad Z = \frac{22.42 - 22.4}{0.008} = 2.5$$

$$P\left(\overline{X} > 22.42\right) = P(Z \geq 2.5) = 1 - P(Z \leq 2.5) = 1 - 0.9938 = 0.0062$$

Il numero di campioni atteso è $300 \cdot 0.0062 = 2$.

c – $$\overline{X} = 22.37 \quad \Rightarrow \quad Z = \frac{22.37 - 22.4}{0.008} = -3.75$$

$$P\left(\overline{X} < 22.37\right) = P(Z \leq -3.75) = 1 - P(Z \leq 3.75) = 1 - 0.9999 = 0.0001$$

Il numero di campioni atteso è $300 \cdot 0.0001 = 0.03$, cioè nessuno.

Esempio 10

Per un certo segmento ampio di popolazione e per un dato anno, il numero medio di giorni di assenza dal lavoro per malattia è 5.4 con una deviazione standard di 2.8 giorni. Calcolare la probabilità che un campione casuale di 49 persone estratto da questa popolazione abbia una media di assenze

a – maggiore di 6 giorni;

b – fra 4 e 6 giorni;

c – fra 4 giorni e mezzo e 5 giorni e mezzo.

La distribuzione della popolazione non è nota, ma, poiché abbiamo un campione più grande di 30, in base al teorema del limite centrale possiamo dire che la distribuzione della media campionaria è approssimativamente normale con media

$$\mu_{\overline{X}} = \mu = 5.4 \qquad \sigma_{\overline{X}} = \frac{2.8}{\sqrt{49}} = 0.4$$

Si standardizza la media campionaria con la formula

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 5.4}{0.4}$$

a – $$\overline{X} = 6 \quad \Rightarrow \quad Z = \frac{6 - 5.4}{0.4} = 1.5$$

La probabilità che un campione casuale di 49 persone abbia una media di assenze maggiore di 6 giorni è

$$P\left(\overline{X} > 6\right) = P(Z > 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668$$

b – $$\overline{X} = 4 \quad \Rightarrow \quad Z = \frac{4 - 5.4}{0.4} = -3.5$$

La probabilità che un campione casuale di 49 persone abbia una media di assenze compresa fra 4 e 6 giorni è

$$P\left(4 < \overline{X} < 6\right) = P(-3.5 \leq Z \leq 1.5) = P(Z \leq 1.5) - P(Z \leq -3.5) =$$
$$= 0.9332 - \left[1 - P(Z \leq 3.5)\right] = 0.9332 - 1 + 0.9998 = 0.933$$

c – $$\overline{X} = 4.5 \quad \Rightarrow \quad Z = \frac{4.5 - 5.4}{0.4} = -2.25$$

$$\overline{X} = 5.5 \quad \Rightarrow \quad Z = \frac{5.5 - 5.4}{0.4} = 0.25$$

La probabilità che un campione casuale di 49 persone abbia una media di assenze compresa fra 4 giorni e mezzo e 5 giorni e mezzo è

$$P\left(4.5 < \overline{X} < 5.5\right) = P(-2.25 \leq Z \leq 0.25) = P(Z \leq 0.25) - P(Z \leq -2.25) =$$
$$= 0.5987 - \left[1 - P(Z \leq 2.25)\right] = 0.5987 - 1 + 0.9878 = 0.5865$$

6.5 Distribuzione della media campionaria (varianza σ^2 incognita)

L'applicazione dei risultati del § 6.4 richiede la conoscenza della varianza σ^2 della popolazione.
Nel caso che il numero *n* degli elementi del campione sia grande (**grande campione**), se σ^2 non è nota, si sostituisce a σ^2 la varianza s^2 del campione.
Se invece l'ampiezza *n* del campione è piccola (**piccolo campione**), si hanno dei risultati solo se il campione proviene da una popolazione normale.
Si dimostra in questo caso il seguente teorema.

Teorema 3
Sia data una popolazione normale avente media μ e da essa si estraggano campioni casuali di ampiezza *n*; indicando con $\overline{X}$ la media campionaria e con *S* lo scarto quadratico medio campionario, la variabile

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{6.5}$$

è una variabile aleatoria avente la **distribuzione *t* di Student**[3] con grado di libertà $\nu = n - 1$.

Questo teorema da un lato è più generale del teorema del limite centrale, nel senso che non richiede la conoscenza di σ, ma d'altra parte richiede l'ipotesi più restrittiva di una popolazione normale.
La distribuzione *t* di Student non è un'unica distribuzione, ma una famiglia di distribuzioni dipendenti dal parametro ν, detto **grado di libertà**.
Nella figura 3 sono riportati il grafico della distribuzione *t* di Student per il grado di libertà $\nu = 4$, e il grafico della distribuzione normale standardizzata.

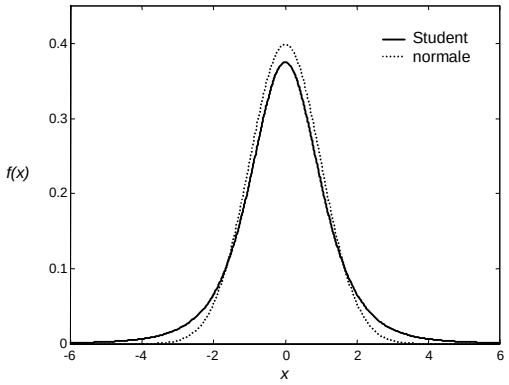


Figura 3

La forma della distribuzione *t* è simile alla normale: entrambe le distribuzioni sono a campana, simmetriche attorno alla media.
Come la distribuzione normale, la distribuzione *t* ha media $\mu = 0$; la sua varianza dipende dal grado di libertà ν; la varianza è maggiore di 1, e tende a 1 al crescere del grado di libertà.
Si può dimostrare che la distribuzione *t* con grado di libertà ν tende alla distribuzione normale standardizzata per $\nu \to \infty$.
Sono disponibili delle tavole, riportate nell'Appendice A, in cui sono tabulati alcuni valori scelti di t_α per vari valori di ν, dove t_α è tale che l'area alla destra di t_α è uguale ad α, come illustrato nella figura 4, pag. seguente.

[3] Lo studioso che studiò questa distribuzione è William S. Gosset (1876-1937), uno statistico impiegato presso le fabbriche di birra della Guinness in Irlanda. Egli affrontò il problema dello studio dei piccoli campioni per ragioni essenzialmente pratiche, il costo e il tempo necessari per studiare grandi campioni, e determinò la distribuzione *t*, rilevante per lo studio dei piccoli campioni. Poiché agli impiegati della Guinness non era concesso pubblicare lavori di ricerca, Gosset utilizzò lo pseudonimo di "Student"

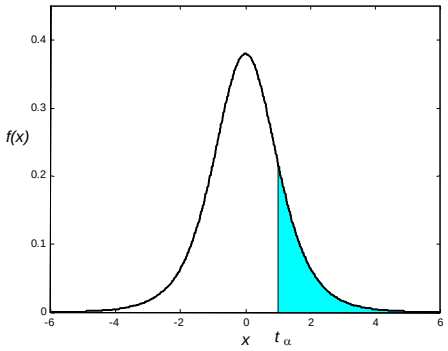


Figura 4

Non è necessario tabulare valori di t_α per $\alpha > 0.50$, perché la distribuzione è simmetrica.
I valori di t_α per $\nu > 29$ sono circa uguali ai corrispondenti valori tratti dalle tavole della distribuzione normale (vedere esempi 15 e 16): infatti la distribuzione normale è una buona approssimazione della distribuzione *t* per valori del grado di libertà $\nu > 29$.

Esempio 11
Data la distribuzione *t* con grado di libertà $\nu = 9$, trovare il valore di t_α tale che l'area a destra di t_α vale $\alpha = 0.05$ (figura 5).

Dalle tavole si deduce che $t_\alpha = 1.833$

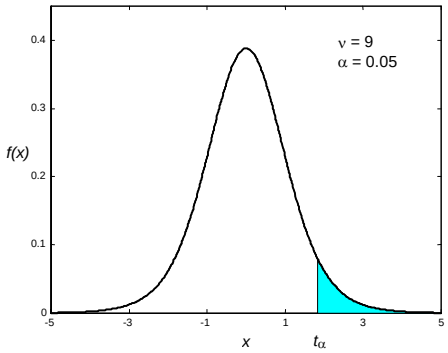


Figura 5

Esempio 12
Data la distribuzione *t* con grado di libertà $\nu = 9$, trovare il valore di t_α tale che la somma dell'area a destra di t_α e dell'area a sinistra di $-t_\alpha$ vale $\alpha = 0.05$ (figura 6).

Area totale delle due code = $\alpha = 0.05 \Rightarrow$ area a destra di t_α (una coda) = $\frac{\alpha}{2} = 0.025$.

Dalle tavole si deduce $t_\alpha = t_{0.025} = 2.262$

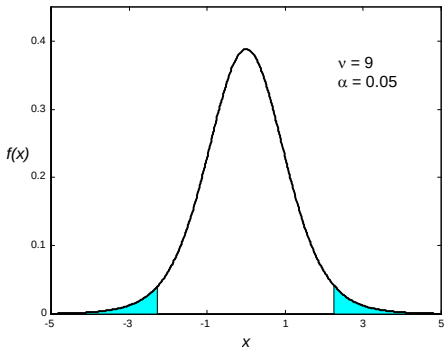


Figura 6

Esempio 13
Data la distribuzione *t* con grado di libertà $\nu = 10$, trovare il valore di t_α tale che l'area compresa fra $-t_\alpha$ e t_α vale $\alpha = 0.90$ (figura 7)

Area compresa fra $-t_\alpha$ e $t_\alpha = \alpha = 0.90 \Rightarrow$ Area totale delle due code $= 1 - \alpha = 0.1 \Rightarrow$
Area di una coda $= \dfrac{1-\alpha}{2} = 0.05$
Dalle tavole si deduce $t_\alpha = t_{0.05} = 1.812$

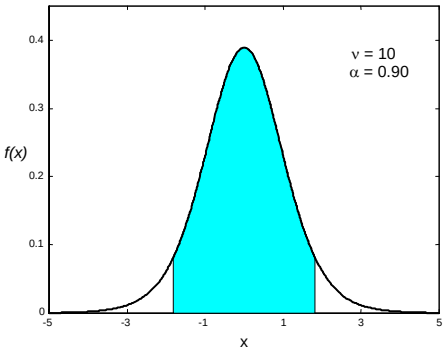


Figura 7

Esempio 14
Data la distribuzione *t* con grado di libertà $\nu = 9$, trovare il valore di t_α tale che l'area a destra di t_α vale $\alpha = 0.99$ (figura 8)

Area a destra di $t_\alpha = \alpha = 0.99 \Rightarrow$ Area di una coda $= 1 - \alpha = 0.01$
Dalle tavole si deduce

$$t_{0.01} = 2.821 \Rightarrow t_\alpha = -2.821$$

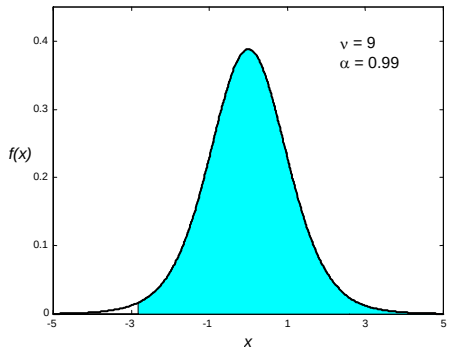


Figura 8

Esempio 15
Data la distribuzione *t* con grado di libertà $\nu > 29$, trovare il valore di t_α tale che l'area a destra di t_α vale $\alpha = 0.025$.
Verificare che si ottiene lo stesso valore con la tavola della distribuzione normale.

Dalla tavola della distribuzione *t* si ottiene

$$t_\alpha = 1.960$$

Dalla tavola dei quantili della distribuzione normale standardizzata si ottiene lo stesso valore

$$z_\alpha = 1.960$$

Esempio 16

Data la distribuzione *t*, trovare i valori di t_α tale che l'area a destra di t_α vale $\alpha = 0.05$ per i gradi di libertà $\nu = 16$, $\nu = 27$, $\nu = 200$.

Dalle tavole si trova

a – $\nu = 16$ $t_\alpha = 1.746$

b – $\nu = 27$ $t_\alpha = 1.703$

c – $\nu = 200$ $t_\alpha = 1.645$

Quest'ultimo valore è uguale al valore che si trova dalla tavola dei quantili della distribuzione normale standardizzata

$z_\alpha = 1.645$.

6.6 Distribuzione della varianza campionaria

Finora abbiamo esaminato la distribuzione della media campionaria; se nell'esempio 7 avessimo studiato la varianza campionaria, avremmo ottenuto la distribuzione di campionamento di questa statistica.

Studiamo la **distribuzione di campionamento della varianza campionaria** per campioni provenienti da una popolazione normale; otteniamo questa distribuzione estraendo tutti i possibili campioni casuali di ampiezza *n* da una popolazione avente distribuzione normale e determinando per ciascuno di essi la varianza campionaria s^2. Poiché s^2 non può essere negativa, ci si attende che la distribuzione della varianza campionaria non sia simmetrica, cioè non sia di tipo normale.

Vale il teorema

Teorema 4

Sia data una popolazione normale avente varianza σ^2 e da essa si estraggano campioni casuali di ampiezza *n*; indicando con S^2 la varianza campionaria, la variabile

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \tag{6.6}$$

è una variabile aleatoria avente la **distribuzione** χ^2 (**chi quadro**) con grado di libertà $\nu = n - 1$.

Il parametro ν è detto **grado di libertà**. Anche la distribuzione chi quadro non è un'unica distribuzione, ma una famiglia di distribuzioni dipendenti dal grado di libertà ν.

Si dimostra che la distribuzione χ^2 ha media $\mu = \nu$ e varianza $\sigma^2 = 2\nu$.

Nella figura 9 sono riportati i grafici della distribuzione χ^2 per valori di ν da 2 a 10.

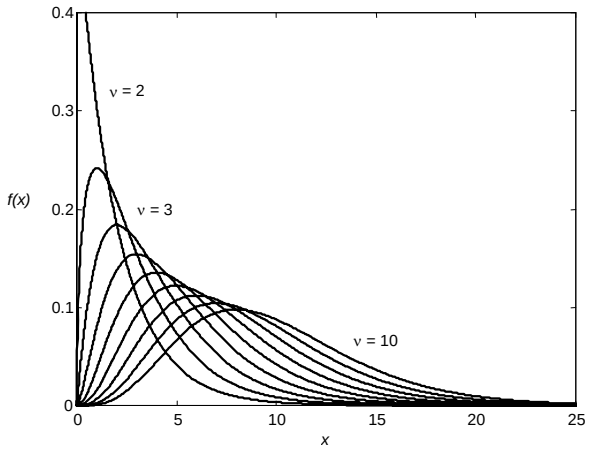


Figura 9

La distribuzione chi quadro è definita solo per valori positivi di x e in generale è asimmetrica; l'asimmetria diminuisce per valori elevati di ν.
Sono disponibili delle tavole, riportate nell'Appendice A, in cui sono tabulati alcuni valori scelti di χ^2_α per vari valori di ν, dove χ^2_α è tale che l'area alla destra di χ^2_α è uguale ad α (figura 10).

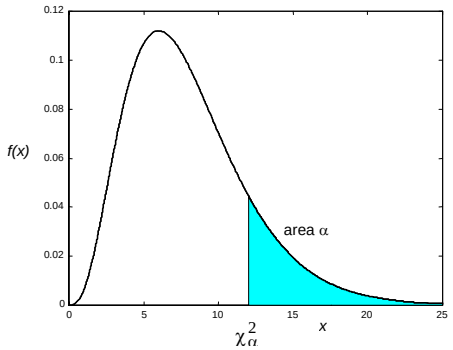


Figura 10

Esempio 17

Data la distribuzione χ^2 con grado di libertà $\nu = 5$, trovare il valore di χ^2_α tale che l'area a destra di χ^2_α vale $\alpha = 0.05$ (figura 11).

Dalle tavole, per $\nu = 5$ e $\alpha = 0.05$ si deduce $\chi^2_\alpha = 11.070$.

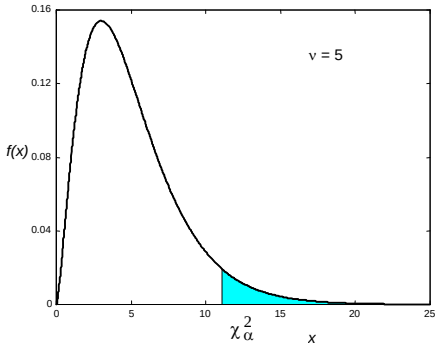


Figura 11

Esempio 18

Data la distribuzione χ^2 , trovare il valore χ^2_α tale che l'area a destra di χ^2_α vale $\alpha = 0.05$ per i gradi di libertà $\nu = 15$, $\nu = 25$ e $\nu = 30$.

Dalle tavole si deduce

a – $\nu = 15$ $\alpha = 0.05$ $\Rightarrow$ $\chi^2_\alpha = 24.996$

b – $\nu = 25$ $\alpha = 0.05$ $\Rightarrow$ $\chi^2_\alpha = 37.652$

c – $\nu = 30$ $\alpha = 0.05$ $\Rightarrow$ $\chi^2_\alpha = 43.773$

Esempio 19

Data la distribuzione χ^2 con grado di libertà $\nu = 5$, trovare il valore di χ^2_α tale che l'area a sinistra di χ^2_α vale $\alpha = 0.05$ (figura 12, pag. seguente).

Area a sinistra di $\chi^2_\alpha = \alpha = 0.05$ $\Rightarrow$ Area a destra di $\chi^2_\alpha = 1 - \alpha = 0.95$
Dalle tavole si deduce

$$\chi^2_\alpha = 1.145 .$$

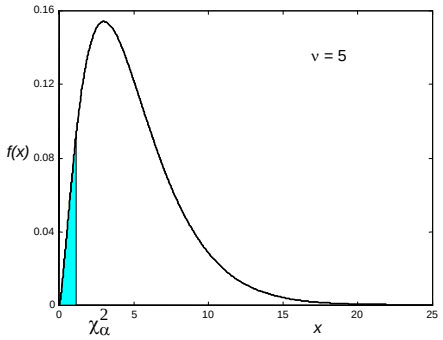


Figura 12

Esempio 20

Data la distribuzione χ^2 con grado di libertà $\nu = 5$, trovare i valori $\chi^2_{\alpha_1}$ e $\chi^2_{\alpha_2}$ tali che il totale dell'area a sinistra di $\chi^2_{\alpha_1}$ e dell'area a destra di $\chi^2_{\alpha_2}$ vale $\alpha = 0.05$ (figura 13).

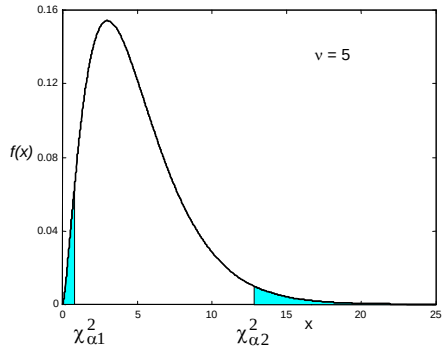


Figura 13

Poiché la distribuzione non è simmetrica, ci possono essere più valori $\chi^2_{\alpha_1}$ e $\chi^2_{\alpha_2}$ per i quali l'area totale è 0.05; ad esempio

area a sinistra = 0.04 e area a destra = 0.01
area a sinistra = 0.025 e area a destra = 0.025
.....

Di solito si scelgono le due code in modo che abbiano uguale area; in questo esempio entrambe hanno area uguale a 0.025.
Con le tavole si ricava

Area a destra di $\chi^2_{\alpha_2} = 0.025 \Rightarrow \chi^2_{\alpha_2} = 12.832$

Area a sinistra di $\chi^2_{\alpha_1} = 0.025 \Rightarrow$ Area a destra di $\chi^2_{\alpha_1} = 0.975 \Rightarrow \chi^2_{\alpha_1} = 0.831$

Un problema strettamente connesso a quello appena trattato dello studio della distribuzione di campionamento della varianza campionaria è quello di determinare la distribuzione del rapporto delle varianze di due campioni indipendenti.
Questo problema deve la sua importanza al fatto che capita spesso di dover confrontare due varianze, e in particolare in alcuni test di ipotesi si deve preliminarmente stabilire se due campioni provengono da popolazioni aventi la stessa varianza; se ciò accade, il loro rapporto sarà uguale a 1.
Di solito però non si conoscono le varianze delle due popolazioni, quindi qualunque confronto viene fatto sulla base delle varianze campionarie.

Per studiare il rapporto di due varianze si utilizza la distribuzione di campionamento della variabile

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

e si ricorre al seguente teorema.

Teorema 5

Siano date due popolazioni normali aventi varianze σ_1^2 e σ_2^2, e si estraggano da esse campioni casuali indipendenti di ampiezza rispettivamente n_1 e n_2; indicando con S_1^2 e S_2^2 le varianze campionarie, la variabile

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \tag{6.7}$$

è una variabile aleatoria avente la **distribuzione *F***, detta anche **distribuzione di Fisher**, di parametri $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$.

La distribuzione F dipende dai due parametri ν_1 e ν_2, detti **gradi di libertà del numeratore e del denominatore**.

La figura 14 mostra alcune distribuzioni F per differenti combinazioni dei gradi di libertà del numeratore e del denominatore.

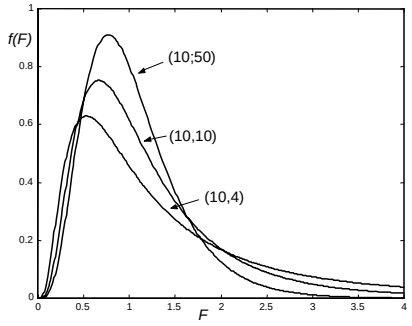


Figura 14

Sono disponibili delle tavole, riportate nell'Appendice A, in cui sono tabulati alcuni valori scelti di F_α, per varie combinazioni di valori di ν_1 e ν_2, dove F_α è tale che l'area alla destra di F_α è uguale ad α, come illustrato nella figura 15. La tavola 7 contiene i valori di F_α per alcuni valori scelti di α e per varie combinazioni di valori di ν_1 e ν_2.

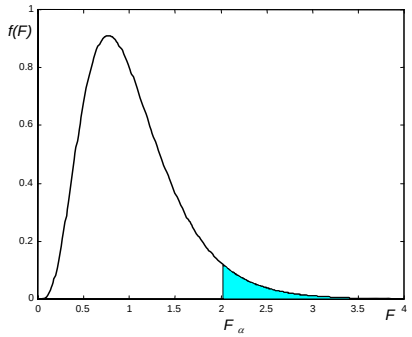


Figura 15

La tavola 7 può essere usata anche per trovare valori di F corrispondenti a code a sinistra di area fissata; a questo scopo si usa l'identità seguente, nella quale si scrive $F_\alpha(\nu_1, \nu_2)$ per indicare F_α con gradi di libertà ν_1 e ν_2

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)} \tag{6.8}$$

Esempio 21

Data la distribuzione F con gradi di libertà $\nu_1 = 15$, $\nu_2 = 25$, trovare il valore F_α tale che l'area a destra di F_α vale

a – $\alpha = 0.10$;
b – $\alpha = 0.05$;
c – $\alpha = 0.01$.

Dalle tavole, per i gradi di libertà $\nu_1 = 15$, $\nu_2 = 25$, si deduce

a – $F_{0.10}(15,25) = 1.77$

b – $F_{0.05}(15,25) = 2.09$

c – $F_{0.01}(15,25) = 2.85$

Esempio 22

Data la distribuzione F con gradi di libertà $\nu_1 = 10$, $\nu_2 = 20$, trovare il valore F_α tale che l'area a destra di F_α vale

a – $\alpha = 0.90$;
b – $\alpha = 0.95$;
c – $\alpha = 0.99$.

Dalle tavole, per i gradi di libertà $\nu_1 = 10$, $\nu_2 = 20$, facendo uso della (6.8) si deduce

a – $\alpha = 0.90 \quad \Rightarrow \quad 1-\alpha = 0.10$

$$F_{0.90}(10,20) = \frac{1}{F_{0.10}(20,10)} = \frac{1}{2.20} = 0.455$$

b – $\alpha = 0.95 \quad \Rightarrow \quad 1-\alpha = 0.05$

$$F_{0.95}(10,20) = \frac{1}{F_{0.05}(20,10)} = \frac{1}{2.77} = 0.361$$

c – $\alpha = 0.99 \quad \Rightarrow \quad 1-\alpha = 0.01$

$$F_{0.99}(10,20) = \frac{1}{F_{0.01}(20,10)} = \frac{1}{4.41} = 0.227$$

Esempio 23

Data la distribuzione F con gradi di libertà $\nu_1 = 15$, $\nu_2 = 10$, trovare i valori $F_{1-\frac{\alpha}{2}}$ e $F_{\frac{\alpha}{2}}$ tali che l'area compresa fra essi vale $\alpha = 0.90$.

Dato che la distribuzione F non è simmetrica, di solito si scelgono le due code in modo che abbiano uguale area; in questo esempio entrambe hanno area uguale a 0.05.

Dalle tavole, per $\frac{\alpha}{2} = 0.05$, $\nu_1 = 15$, $\nu_2 = 10$ si desume che

$$F_{0.05}(15,10) = 2.85$$

Con la (6.8), per $1-\frac{\alpha}{2} = 0.95$ si ha

$$F_{0.95}(15,10) = \frac{1}{F_{0.05}(10,15)} = \frac{1}{2.54} = 0.394\,.$$

Si ricordi che i teoremi 4 e 5 richiedono l'ipotesi che i campioni vengano estratti da una popolazione normale. Contrariamente a quanto accade con la distribuzione *t* (teorema 3), scostamenti anche modesti dalla distribuzione normale possono avere conseguenze serie sulle distribuzioni campionarie.

7. Stima dei parametri

7.1 Introduzione

Abbiamo visto come la teoria dei campioni possa essere usata per ottenere informazioni riguardanti campioni estratti casualmente da una popolazione.
Da un punto di vista applicativo è però spesso più importante trarre conclusioni sull'intera popolazione utilizzando i risultati ottenuti su campioni estratti da essa. Questi sono i problemi di cui si occupa l'**inferenza statistica**.
I metodi della statistica inferenziale riguardano essenzialmente due aree: la **stima dei parametri** e i **test di ipotesi**.
Il primo importante problema dell'inferenza statistica, di cui ci occupiamo in questo capitolo, è la stima dei parametri di una popolazione, media, varianza, scarto quadratico medio, per mezzo dei corrispondenti parametri campionari o statistiche del campione.
Il valore del parametro da stimare per la popolazione è incognito, e possiamo solo chiederci se, dopo ripetuti campionamenti, la distribuzione della statistica ha certe proprietà che possono garantirci che la statistica è vicina al valore incognito del parametro.
Ad esempio, sappiamo dal teorema 1, Cap. 6, che la distribuzione della media campionaria ha la stessa media della popolazione da cui è stato ottenuto il campione: ci aspettiamo perciò che, dopo più campionamenti, la media campionaria sia vicina alla media della popolazione.

7.2 Stime puntuali e stime per intervallo

Per i parametri di una popolazione è possibile calcolare due tipi di stima: una stima puntuale e una stima per intervallo.

Definizioni 1
Se la stima di un parametro della popolazione è data da un singolo numero, tale valore è detto **stima puntuale** del parametro.
Se invece la stima di un parametro della popolazione fornisce gli estremi di un intervallo fra i quali si può supporre, con un certo grado di fiducia, che il parametro sia compreso, tale stima è detta **stima per intervallo** del parametro.

I parametri che più frequentemente accade di dover stimare sono:
1 – la media μ di una popolazione;
2 – la varianza σ^2 di una popolazione;
3 – la proporzione p di individui di una popolazione che appartengono a una certa classe di interesse;
4 – la differenza fra le medie di due popolazioni $\mu_1 - \mu_2$;
5 – la differenza fra le proporzioni di due popolazioni $p_1 - p_2$.
Ragionevoli stime puntuali di questi parametri sono:
1 – per μ, la media campionaria $\bar{x}$;
2 – per σ^2, la varianza campionaria s^2;
3 – per p, la proporzione campionaria $\hat{p} = \frac{x}{n}$, dove x è il numero di individui in un campione di ampiezza n appartenenti alla classe di interesse;
4 – per $\mu_1 - \mu_2$, la differenza $\bar{x}_1 - \bar{x}_2$ fra le medie di due campioni indipendenti;
5 – per $p_1 - p_2$, la differenza $\hat{p}_1 - \hat{p}_2$ fra le proporzioni di due campioni indipendenti.

Si possono avere più stime puntuali per lo stesso parametro; per esempio se si vuole stimare la media di una popolazione, si potrebbe usare anche la mediana campionaria, o magari la media fra il più piccolo e il più grande fra i valori del campione[1].
Per decidere quale fra le possibili stime puntuali è preferibile usare, ci basiamo sulla verifica di alcune proprietà che gli stimatori devono possedere per essere giudicati i più adatti.
Una di queste è la proprietà della **correttezza** o **non distorsione**.

Definizione 2
Se la media di una distribuzione campionaria di una statistica è uguale al corrispondente parametro della popolazione, la statistica è detta **stimatore corretto** o **non distorto** del parametro.

I valori corrispondenti a tali statistiche sono detti **stime corrette**. In altre parole, una statistica è uno stimatore corretto se "in media" i suoi valori uguagliano il parametro che valuta.
Ad esempio la media della distribuzione campionaria della media è

$$\mu_{\bar{X}} = \mu$$

quindi la media campionaria $\bar{x}$ è una stima corretta della media μ di una popolazione.
Lo stimatore corretto di un parametro non è unico. Ad esempio anche la mediana campionaria è una stima corretta della media della popolazione.
Occorre quindi un'ulteriore proprietà, detta **efficienza**, per decidere quale tra più stime corrette sia la migliore per stimare un parametro.

Definizione 3
Se due statistiche sono entrambe stimatori corretti di un parametro, la statistica per cui la varianza della sua distribuzione campionaria è minore è detta **stimatore più efficiente**.

Si può dimostrare che, fra tutte le statistiche che stimano la media di una popolazione, la media campionaria è la più efficiente. Anche la varianza campionaria è una stima corretta ed efficiente della varianza di una popolazione.
In generale, si può affermare che le stime puntuali suggerite ai punti 1–5, pag. 183, sono stime corrette ed efficienti dei corrispondenti parametri della/delle popolazioni.

Esempio 1
Dato un campione di 5 misurazioni del diametro di una sferetta in cm

6.33 6.37 6.36 6.32 6.37

trovare stime corrette ed efficienti per la media e la varianza della popolazione.

La stima corretta ed efficiente per la media della popolazione è la media campionaria

$$\bar{x} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35\,\text{cm}$$

Anche per la varianza la stima corretta ed efficiente è la varianza campionaria

$$s^2 = \frac{1}{4}\sum_{i=1}^{5}\left(x_i - \bar{x}\right)^2 = \frac{(0.02)^2 + (0.02)^2 + (0.01)^2 + (0.03)^2 + (0.02)^2}{4} = 0.00055\,\text{cm}^2$$

Poiché non ci si può aspettare che una stima puntuale coincida esattamente con la quantità che essa deve stimare, è spesso preferibile usare una stima per intervallo, ossia un intervallo per il quale si può affermare con un certo grado di fiducia che conterrà il parametro della popolazione che si vuole stimare.
Tali stime per intervallo vengono comunemente chiamate **intervalli di confidenza**.

[1] Si vedano anche le osservazioni 1 e 2, pag. 199, 200.

7.3 Intervalli di confidenza per la media (varianza nota)

Come già detto, la media campionaria $\bar{x}$ è una buona stima, corretta ed efficiente, della media μ di una popolazione. Tuttavia, non c'è alcuna probabilità che la stima sia esattamente uguale a μ; ha quindi più significato stimare μ con un intervallo, che in qualche modo ci dà informazioni sulla probabile grandezza di μ.
Per ottenere una stima per intervallo, si utilizzano le proprietà delle distribuzioni campionarie. In questo caso, poiché si vuole stimare la media di una popolazione per mezzo della media di un campione, facciamo ricorso alla distribuzione della media campionaria.
Nel Cap. 6 si è visto come determinare, conoscendo la distribuzione della popolazione, la percentuale delle medie campionarie che cadono in un intervallo prefissato (vedere § 6.4, esempi 8, 9 e 10). Le conclusioni che si traggono sono basate su un ragionamento deduttivo.
Nell'inferenza statistica si fa invece un ragionamento induttivo: ci basiamo infatti sui risultati di un solo campione per trarre conclusioni sull'intera popolazione, e non viceversa. Questo comporta che non si giungerà sempre a delle conclusioni corrette partendo da un singolo campione.
Nel caso in cui si voglia stimare la media della popolazione, può accadere che per alcuni (si spera molti) campioni la stima per intervallo per la media μ sia corretta, ossia l'intervallo ottenuto comprenda effettivamente la media μ, e per altri campioni (si spera pochi) questo non accada.
Poiché nella pratica si estrae un solo campione, e ovviamente non conosciamo la media della popolazione, non possiamo essere certi che le conclusioni a cui si perviene siano corrette. Per risolvere questo problema, ogni stima per intervallo viene calcolata valutando anche la percentuale dei campioni che dà luogo a conclusioni corrette, ossia il **grado di fiducia**.

Si consideri una popolazione avente una distribuzione con varianza σ^2 nota e media incognita μ, e si estragga da questa popolazione un campione di ampiezza *n*.
In base al teorema del limite centrale possiamo affermare che, per grandi valori di *n*, la statistica

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{7.1}$$

ha approssimativamente la distribuzione normale standardizzata.

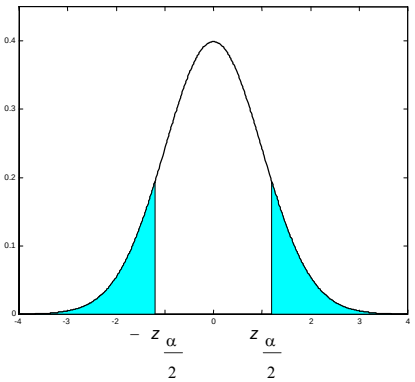


Figura 1

Se l'area sottesa dalla distribuzione normale a destra di $z_{\frac{\alpha}{2}}$ vale $\frac{\alpha}{2}$ (figura 1), allora l'area compresa fra $-z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$ vale $1-\alpha$, perciò

$$P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Di conseguenza, si può asserire, con probabilità uguale a $1-\alpha$, che è soddisfatta la disuguaglianza

$$-z_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}. \tag{7.2}$$

Dalla disuguaglianza (7.2), risolvendo rispetto a μ si ottiene

$$\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

Pertanto, una volta estratto il campione di ampiezza *n*, con *n* sufficientemente grande (***n* ≥ 30, grande campione**) e calcolato il valore $\overline{x}$ della media del campione, si ottiene la seguente stima per intervallo per la media μ, soddisfatta con probabilità $1 - \alpha$.

$$\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{7.3}$$

Si può quindi affermare con probabilità $1 - \alpha$ che l'intervallo $\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$ contiene la media μ della popolazione.

L'intervallo (7.3) è detto anche **intervallo di confidenza per la media μ**, per **grandi campioni**, con **grado di fiducia** $(1-\alpha)\cdot 100\%$.

La formula (7.3) vale per popolazioni anche non normali, purché il campione sia grande. Come già detto nel Cap. 6, nella pratica un campione viene ritenuto sufficientemente grande se $n \geq 30$ (vedere schema riassuntivo, pag. 178).

Se la popolazione da cui proviene il campione ha distribuzione normale, la (7.3) vale qualunque sia la dimensione del campione.

Poiché nelle applicazioni pratiche di solito lo scarto quadratico σ della popolazione non è noto, se il campione è grande, si può sostituire σ con lo scarto quadratico medio campionario *s*, commettendo un errore di approssimazione.

Il valore di $z_{\frac{\alpha}{2}}$ che compare nella (7.3) è detto **valore critico** della distribuzione; a ciascun grado di fiducia corrisponde un diverso valore critico.

I valori più comunemente usati per $1 - \alpha$ sono 0.90, 0.95 e 0.99 ; di solito si usa il termine **grado di fiducia** del 90%, del 95% o del 99% , anziché il termine probabilità uguale a 0.90, a 0.95, oppure a 0.99; i corrispondenti valori di $z_{\frac{\alpha}{2}}$ sono i seguenti

grado di fiducia del 90% $\quad z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

grado di fiducia del 95% $\quad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

grado di fiducia del 99% $\quad z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$

Questi valori possono essere letti sulla tabella dei percentili della distribuzione normale standardizzata.

Come già detto in precedenza, per trarre conclusioni sulla media della popolazione ci basiamo sui risultati di un singolo campione; questo ha come conseguenza che non si giungerà sempre a delle conclusioni corrette, ossia non è garantito che la media μ cadrà davvero nell'intervallo di confidenza ottenuto.

In generale quindi un intervallo di confidenza con grado di fiducia ad esempio del 95% va interpretato nel seguente modo: se si considerano tutti i possibili campioni di ampiezza *n*, e per

ciascuno di essi si calcola la media campionaria e il corrispondente intervallo di confidenza centrato su questa, il 95% degli intervalli così ottenuti contiene il corrispondente parametro della popolazione e solo il 5% non lo contiene.
Per quanto detto prima, non possiamo sapere se uno specifico intervallo contiene o meno il parametro della popolazione, tuttavia possiamo affermare che abbiamo un grado di fiducia ad esempio del 95% di aver scelto un campione a cui corrisponde una stima per intervallo comprendente il parametro della popolazione.

La lunghezza di un intervallo di confidenza con grado di fiducia $(1-\alpha)\cdot 100\%$ è

$$2z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

e dipende quindi da tre fattori
– *n*: al crescere dell'ampiezza del campione, la lunghezza dell'intervallo diminuisce, quindi la stima è più precisa;
– α: al crescere del grado di fiducia richiesto, la lunghezza dell'intervallo aumenta, quindi la stima è meno precisa;
– σ: al crescere della deviazione standard, che riflette la variabilità del campione, la lunghezza dell'intervallo aumenta.
Normalmente solo *n* e α possono essere controllati, mentre σ dipende dal tipo di dati studiati.
In definitiva, la precisione della stima e un elevato grado di fiducia sono due obiettivi tra loro in conflitto: se si vuole aumentare la precisione della stima, senza diminuire il grado di fiducia, si deve aumentare la dimensione del campione.
Nelle applicazioni pratiche può non essere facile trovare un buon compromesso tra grado di fiducia e ampiezza del campione: un maggior grado di fiducia comporta una perdita di precisione nella stima; un aumento delle dimensioni del campione può comportare problemi pratici o essere antieconomico. Solo l'esperienza e la conoscenza del problema trattato possono indicare la scelta più opportuna.

Esempio 2
Sia dato un campione di ampiezza $n = 100$ estratto da una popolazione avente scarto quadratico medio $\sigma = 5.1$; la media campionaria sia $\bar{x} = 21.6$. Costruire l'intervallo di confidenza al 95% per la media μ della popolazione.

Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Applicando la formula (7.3) si ottiene l'intervallo di confidenza

$$21.6 - 1.96 \cdot \frac{5.1}{\sqrt{100}} < \mu < 21.6 + 1.96 \cdot \frac{5.1}{\sqrt{100}}$$

$$20.6 < \mu < 22.6$$

Questo intervallo può anche non contenere μ, ma abbiamo un grado di fiducia del 95% che lo contenga. In altre parole, se applichiamo ripetutamente su tutti i campioni di ampiezza $n = 100$ estraibili dalla popolazione la formula (7.3) per calcolare l'intervallo di confidenza, il 95% degli intervalli di confidenza conterrà la media μ della popolazione.

Esempio 3
Costruire un intervallo di confidenza con grado di fiducia del 99% per la media della popolazione da cui è stato estratto il campione studiato nell'esempio 2, Cap. 1.

Per questo campione si è calcolato (esempio 34, Cap. 1)

$$\bar{x} = 18.8 \qquad s^2 = 31.96$$

Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$.

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$18.8 - 2.576 \cdot \frac{\sqrt{31.96}}{\sqrt{80}} < \mu < 18.8 + 2.576 \cdot \frac{\sqrt{31.96}}{\sqrt{80}}$$

$$17.1 < \mu < 20.5$$

Esempio 4

Le misure dei diametri di un campione casuale di 200 sferette da cuscinetto prodotte da una macchina in una settimana hanno una media campionaria $\bar{x} = 0.824$ cm e una deviazione standard campionaria $s = 0.042$ cm.

Determinare gli intervalli di confidenza per la media della popolazione con grado di fiducia del 95% e del 99%.

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$0.824 - 1.96 \cdot \frac{0.042}{\sqrt{200}} < \mu < 0.824 + 1.96 \cdot \frac{0.042}{\sqrt{200}}$$

$$0.818 < \mu < 0.830$$

b –Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$0.824 - 2.576 \cdot \frac{0.042}{\sqrt{200}} < \mu < 0.824 + 2.576 \cdot \frac{0.042}{\sqrt{200}}$$

$$0.816 < \mu < 0.832$$

Si osservi che aumentando il grado di fiducia l'ampiezza dell'intervallo aumenta, ossia a parità di numero di elementi del campione la stima è meno precisa.

Esempio 5

Si vuole stimare il numero medio di battiti cardiaci al minuto per una certa popolazione. Il numero medio di battiti al minuto per un campione di 49 soggetti è risultato uguale a 90. La popolazione è distribuita in modo normale con uno scarto quadratico medio $\sigma = 10$.

Trovare gli intervalli di confidenza per la media della popolazione con i gradi di fiducia del 90%, 95% e 99%.

a – Per il grado di fiducia del 90% il valore critico è $z_{\frac{\alpha}{2}} = 1.645$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$90 - 1.645 \cdot \frac{10}{\sqrt{49}} < \mu < 90 + 1.645 \cdot \frac{10}{\sqrt{49}}$$

$$87.65 < \mu < 92.35$$

b – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$90 - 1.96 \cdot \frac{10}{\sqrt{49}} < \mu < 90 + 1.96 \cdot \frac{10}{\sqrt{49}}$$

$$87.20 < \mu < 92.80$$

c – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$90 - 2.576 \cdot \frac{10}{\sqrt{49}} < \mu < 90 + 2.576 \cdot \frac{10}{\sqrt{49}}$$

$$86.32 < \mu < 93.68$$

Si osservi come, restando invariata l'ampiezza del campione, all'aumentare del grado di fiducia cresce l'ampiezza dell'intervallo di confidenza, ossia la stima diventa meno precisa.

Esempio 6

Sia dato un campione di 100 studenti tratto da una popolazione di studenti di sesso maschile iscritti ad un'università; la tabella 1 rappresenta la distribuzione di frequenza dei pesi in kg degli studenti. Trovare gli intervalli di confidenza al 95% e al 99% per il peso medio di tutti gli studenti.

Classi (peso)	*N° studenti (freq. ass.)*	*Valori centrali*
$60 \le x \le 62$	5	61
$63 \le x \le 65$	18	64
$66 \le x \le 68$	42	67
$69 \le x \le 71$	27	70
$72 \le x \le 74$	8	73

Tabella 1

Calcoliamo la media e la varianza campionarie usando i dati raggruppati

$$\bar{x} = \frac{5 \cdot 61 + 18 \cdot 64 + 42 \cdot 67 + 27 \cdot 70 + 8 \cdot 73}{100} = 67.45$$

$$s^2 = \frac{1}{99}\left[5 \cdot 61^2 + 18 \cdot 64^2 + 42 \cdot 67^2 + 27 \cdot 70^2 + 8 \cdot 73^2 - 100 \cdot 67.45^2\right] = 8.61$$

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$67.45 - 1.96 \cdot \frac{\sqrt{8.61}}{\sqrt{100}} < \mu < 67.45 + 1.96 \cdot \frac{\sqrt{8.61}}{\sqrt{100}}$$

$$66.87 < \mu < 68.02$$

b – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$67.45 - 2.576 \cdot \frac{\sqrt{8.61}}{\sqrt{100}} < \mu < 67.45 + 2.576 \cdot \frac{\sqrt{8.61}}{\sqrt{100}}$$

$$66.69 < \mu < 68.21$$

La disuguaglianza (7.2), valida con probabilità $1 - \alpha$, può anche essere usata per ricavare una formula che consente di determinare l'**ampiezza** *n* **del campione** necessaria per ottenere un errore prefissato.

La (7.2) equivale a

$$\frac{\left|\overline{X} - \mu\right|}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}$$

ossia

$$\left|\overline{X} - \mu\right| < z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Indicando il massimo dell'errore con

$$E = \max \left| \overline{X} - \mu \right|$$

la stima di E con probabilità $1 - \alpha$ è

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{7.4}$$

In altre parole, se si vuole stimare la media μ della popolazione con la media campionaria di un campione di ampiezza n ($n \geq 30$), si può affermare, con probabilità $1 - \alpha$, che l'errore $\left| \overline{X} - \mu \right|$ sarà al più uguale a $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

Dalla formula (7.4), risolvendo rispetto a n, si ricava l'ampiezza del campione necessaria per stimare la media con un errore prefissato E e con un dato grado di fiducia (si ricordi che n deve essere un intero)

$$n \geq \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2 \tag{7.5}$$

Esempio 7

Determinare l'ampiezza campionaria che consente di ottenere un intervallo di confidenza per la media μ di una popolazione con gradi di fiducia del 95% e del 99%, con un errore in valore assoluto non superiore a 5, supponendo che lo scarto quadratico medio sia $\sigma = 15$.

Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$. Con la formula (7.5) si ottiene

$$n \geq \left(\frac{1.96 \cdot 15}{5} \right)^2 = 34.6$$

Per ottenere la stima con la precisione fissata e con grado di fiducia del 95%, occorre scegliere un campione di ampiezza $n = 35$.

Per il grado di fiducia del 99% il valore critico è invece $z_{\frac{\alpha}{2}} = 2.576$. Con la formula (7.5) si ottiene

$$n \geq \left(\frac{2.576 \cdot 15}{5} \right)^2 = 59.7$$

Per ottenere la stima con la precisione fissata e con grado di fiducia del 99%, occorre scegliere un campione di ampiezza $n = 60$. Per avere un maggior grado di fiducia occorre quindi un campione di maggior ampiezza.

Esempio 8

Un medico misura i tempi di reazione dei suoi pazienti a un determinato stimolo. La stima dello scarto quadratico medio è $s = 0.05$ sec.

Calcolare quanto deve essere grande il campione di misurazioni affinché si possa asserire con grado di fiducia del 95% e del 99%, che l'errore nello stimare il tempo medio di reazione nella popolazione non è superiore a 0.01 sec.

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$.

Con la formula (7.5) si ottiene

$$n \geq \left(\frac{1.96 \cdot 0.05}{0.01} \right)^2 = 96.04$$

Quindi possiamo avere un grado di fiducia del 95% che l'errore nella stima del tempo medio sarà al più 0.01 sec, se prendiamo un campione di ampiezza $n = 97$.

b – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$. Con la formula (7.5) si ottiene

$$n \geq \left(\frac{2.576 \cdot 0.05}{0.01} \right)^2 = 165.9$$

Quindi il campione deve avere ampiezza $n = 166$.

Si osservi (esempi 7 e 8) che per avere un maggior grado di fiducia, a parità di errore, bisogna usare un campione di ampiezza più grande.

7.4 Intervalli di confidenza per la media (varianza incognita)

L'applicazione della (7.3) richiede la conoscenza di σ; se σ non è noto, si è già osservato che per grandi campioni può essere sostituito con lo scarto quadratico medio campionario *s*.

Per **piccoli campioni** (***n* < 30**), nell'ipotesi che la popolazione da cui si estrae il campione abbia distribuzione normale, ci si può servire del teorema 3, Cap. 6, in base al quale la statistica

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{7.6}$$

è una variabile aleatoria che ha la distribuzione *t* di Student con grado di libertà $\nu = n - 1$.

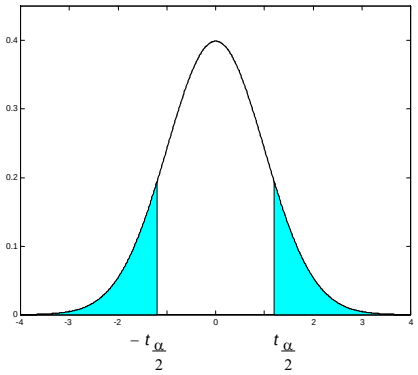


Figura 2

Procedendo come nel caso dei grandi campioni, se l'area sottesa dalla distribuzione *t* a destra di $t_{\frac{\alpha}{2}}$ vale $\frac{\alpha}{2}$ (figura 2), allora l'area compresa fra $-t_{\frac{\alpha}{2}}$ e $t_{\frac{\alpha}{2}}$ vale $1 - \alpha$, perciò

$$P\left(-t_{\frac{\alpha}{2}} < T < t_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

In altre parole si può asserire, con probabilità uguale a $1 - \alpha$, che è soddisfatta la disuguaglianza

$$-t_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{\frac{\alpha}{2}} \tag{7.7}$$

Pertanto, una volta estratto il campione di ampiezza *n*, con $n < 30$, e calcolati i valori della media $\overline{x}$

e dello scarto quadratico medio *s* del campione, si ottiene la stima per intervallo per la media μ, con probabilità $1 - \alpha$, o con grado di fiducia $(1-\alpha)\cdot 100\%$

$$\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad (7.8)$$

L'intervallo (7.8) è detto **intervallo di confidenza per la media μ**, per **piccoli campioni**, con grado di fiducia $(1-\alpha)\cdot 100\%$.

Si ricordi che il grado di libertà della distribuzione *t* è $\nu = n - 1$.

Il valore di $t_{\frac{\alpha}{2}}$ che compare nella (7.8) è detto **valore critico** della distribuzione; a ciascun grado di fiducia corrisponde un diverso valore critico, e diversamente dal caso dei grandi campioni, tale valore dipende anche dal grado di libertà della distribuzione *t*.

I valori più comunemente usati per $1 - \alpha$ sono 0.90, 0.95 e 0.99 ; i relativi gradi di fiducia sono il 90%, il 95% e il 99%; i corrispondenti valori di $t_{\frac{\alpha}{2}}$ sono

grado di fiducia del 90% $\quad t_{\frac{\alpha}{2}} = t_{0.05}$

grado di fiducia del 95% $\quad t_{\frac{\alpha}{2}} = t_{0.025}$

grado di fiducia del 99% $\quad t_{\frac{\alpha}{2}} = t_{0.005}$

Questi valori possono essere letti sulla tabella della distribuzione *t* in corrispondenza al grado di libertà $\nu = n - 1$.

Esempio 9

Sia dato un campione di 16 oggetti di cui si misura il peso, trovando un peso medio $\bar{x} = 3.42$ g e uno scarto quadratico medio $s = 0.68$ g.

Determinare un intervallo di confidenza con grado di fiducia del 99% per il peso medio della popolazione.

Poiché si tratta di misure, si può ragionevolmente ipotizzare che la popolazione da cui proviene il campione abbia distribuzione normale.

Il campione ha ampiezza $n = 16$, perciò il grado di libertà è

$$\nu = n - 1 = 15.$$

Dalle tavole della distribuzione *t* si ottiene

$$t_{0.005} = 2.947.$$

Con la formula (7.8) si ottiene l'intervallo di confidenza

$$3.42 - 2.947 \cdot \frac{0.68}{\sqrt{16}} < \mu < 3.42 + 2.947 \cdot \frac{0.68}{\sqrt{16}}$$

$$2.91 < \mu < 3.93$$

Esempio 10

Un campione di 10 misurazioni del diametro di una sferetta ha una media campionaria $\bar{x} = 4.38$ cm e una deviazione standard campionaria $s = 0.06$ cm. Determinare gli intervalli di confidenza con grado di fiducia del 90%, 95% e 99% per il diametro medio della popolazione.

Poiché si tratta di misure, si può ragionevolmente ipotizzare che la popolazione da cui proviene il campione abbia distribuzione normale.

Il campione ha ampiezza $n = 10$, perciò il grado di libertà è $\nu = n - 1 = 9$.

a – Per il grado di fiducia del 90% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.05} = 1.833$

Con la formula (7.8) si ottiene l'intervallo di confidenza

$$4.38 - 1.833 \cdot \frac{0.06}{\sqrt{10}} < \mu < 4.38 + 1.833 \cdot \frac{0.06}{\sqrt{10}}$$

$$4.34 < \mu < 4.42$$

b – Per il grado di fiducia del 95% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.025} = 2.262$

$$4.38 - 2.262 \cdot \frac{0.06}{\sqrt{10}} < \mu < 4.38 + 2.262 \cdot \frac{0.06}{\sqrt{10}}$$

$$4.33 < \mu < 4.43$$

c – Per il grado di fiducia del 99% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.005} = 3.250$

$$4.38 - 3.250 \cdot \frac{0.06}{\sqrt{10}} < \mu < 4.38 + 3.250 \cdot \frac{0.06}{\sqrt{10}}$$

$$4.31 < \mu < 4.45$$

Si osservi come, restando invariata l'ampiezza del campione, all'aumentare del grado di fiducia cresce l'ampiezza dell'intervallo di confidenza, ossia la stima è meno precisa.

Esempio 11

Le misure in kg del peso di un campione di 10 studenti maschi del primo anno di un'università sono

60	63	60	68	70	72	65	61	69	67

Trovare un intervallo di confidenza con grado di fiducia del 99% per il peso medio della popolazione universitaria maschile del primo anno di quella università.

Calcoliamo la media e la varianza campionaria

$$\bar{x} = \frac{60 + 63 + 60 + 68 + 70 + 72 + 65 + 61 + 69 + 67}{10} = 65.5$$

$$s^2 = \frac{1}{9} \cdot \left[60^2 + 63^2 + 60^2 + 68^2 + 70^2 + 72^2 + \right.$$

$$\left. + 65^2 + 61^2 + 69^2 + 67^2 - 10 \cdot 65.5^2 \right] = 18.94$$

Il campione ha ampiezza $n = 10$, perciò il grado di libertà è $\nu = n - 1 = 9$.

Per il grado di fiducia del 99% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.005} = 3.250$

$$65.5 - 3.250 \cdot \frac{\sqrt{18.94}}{\sqrt{10}} < \mu < 65.5 + 3.250 \cdot \frac{\sqrt{18.94}}{\sqrt{10}}$$

$$61.02 < \mu < 69.98$$

Osservazione 1

Come già osservato (Cap. 1, pag. 23), la media è sensibile ai valori estremi, ossia quelli che si discostano in modo quantitativamente apprezzabile dalla maggior parte dei dati dell'insieme. Questi valori sono a volte chiamati **outliers**.

Abbiamo anche osservato che la mediana, non essendo sensibile ai valori estremi, è da preferire alla media come misura di tendenza centrale, quando vi sono outliers.

Per lo stesso motivo si può usare preferibilmente la mediana campionaria come stimatore della mediana della popolazione per fare inferenza sulla tendenza centrale di una popolazione; la mediana campionaria, oltre a fornire una stima puntuale della mediana della popolazione, consente anche di costruire un intervallo di confidenza per la media.

Osservazione 2
Gli stimatori che non sono sensibili agli outliers sono chiamati **stimatori robusti**. Un altro stimatore robusto per la tendenza centrale è la **media trimmed**.
Dato un campione di n dati, la media trimmed al q% si calcola come segue:
1 – si ordinano i dati;
2 – si eliminano i q% dati più piccoli e i q% dati più grandi. I valori di solito usati sono q% = 10% o q% = 20%;
3 – si calcola la media dei rimanenti dati.
Generalmente il valore della media trimmed è compreso fra la media e la mediana.
Basandosi sulla media trimmed si può costruire un intervallo di confidenza per la media della popolazione.
L'effettiva costruzione degli intervalli di confidenza basati sull'uso della mediana e della media trimmed come stimatori non sarà trattata in queste lezioni.

Esempio 12
Calcolo e confronto di più stime della media di una popolazione, per dati con outliers.
Si estrae il seguente campione di n = 10 dati da una popolazione

12.8	9.4	8.7	11.6	13.1
9.8	14.1	8.5	12.1	10.3

Media

$$\bar{x} = \frac{12.8 + 9.4 + 8.7 + 11.6 + 13.1 + 9.8 + 14.1 + 8.5 + 12.1 + 10.3}{10} = 11.04$$

Dati ordinati

8.5 8.7 9.4 9.8 10.3 11.6 12.1 12.8 13.1 14.1

Mediana

$$M = \frac{10.3 + 11.6}{2} = 10.95$$

Media trimmed al 10%
Si scarta il 10% dei dati più piccoli e il 10% dei dati più grandi (ossia i dati 8.5 e 14.1) prima di calcolare la media

$$\bar{x}_{tr(10)} = \frac{8.7 + 9.4 + 9.8 + + 10.3 + 11.6 + 12.1 + 12.8 + 13.1}{8} = 10.98$$

Come si è detto, il valore della media trimmed è compreso fra la media e la mediana.

7.5 Intervalli di confidenza per la proporzione

Un caso particolarmente importante di stima della media per una popolazione non normale e per grandi campioni è quello di una popolazione bernoulliana. Si vuole stimare il valore del parametro p (probabilità di successo), che rappresenta la frequenza relativa o proporzione con cui una certa caratteristica si presenta negli individui di una data popolazione.
Esempi tipici di questa situazione sono i seguenti.
1 – Il sondaggio di opinione: si vuole stimare la proporzione p della popolazione complessiva che è d'accordo con una certa opinione, osservando il valore che questa proporzione ha su un campione di n individui.
2 – La produzione di un dato tipo di oggetto: il produttore vuole poter garantire che la proporzione di pezzi difettosi in una data produzione non superi un certo valore prefissato; occorre quindi determinare, esaminando un campione, un intervallo di confidenza per la proporzione p di pezzi difettosi in una produzione, ed eventualmente intervenire sulla produzione affinché la proporzione di pezzi difettosi non superi una certa soglia fissata.

3 – Lo studio della diffusione di una data malattia: si vuole stimare qual è la proporzione di pazienti di una certa popolazione che ha una data malattia, studiando il valore di questa proporzione su un campione di *n* persone appartenenti a quella popolazione.
Per stimare la proporzione di una popolazione procediamo nello stesso modo in cui abbiamo stimato la media di una popolazione.
Si estraggono campioni di ampiezza *n* dalla popolazione e si considera la proporzione campionaria $\hat{P} = \frac{X}{n}$, dove *X* è il numero di volte in cui la caratteristica osservata si presenta nel campione.
Questa proporzione campionaria è uno stimatore corretto della proporzione *p* della popolazione e viene usato come stima puntuale.
Nel § 5. 5 abbiamo visto che, quando si ha sia $np \geq 5$ che $n(1-p) \geq 5$, la distribuzione binomiale di parametri *n* e *p* può essere approssimata da una distribuzione normale avente media $\mu = np$ e varianza $\sigma^2 = np(1-p)$. In altri termini la statistica

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{7.9}$$

ha approssimativamente la distribuzione normale standardizzata per grandi valori di *n*.
Quindi quando *n* è grande si può costruire un intervallo di confidenza per il parametro *p*, usando l'approssimazione normale per la distribuzione binomiale. Possiamo affermare che

$$P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

ossia, con probabilità $1 - \alpha$, vale la disuguaglianza

$$-z_{\frac{\alpha}{2}} < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}} \tag{7.10}$$

Per ricavare l'intervallo di confidenza per *p* occorrerebbe risolvere la disuguaglianza (7.10) rispetto a *p*; questo non è difficile, ma il calcolo può essere notevolmente semplificato sostituendo nell'espressione $\sqrt{\frac{p(1-p)}{n}}$, che compare al denominatore, la quantità *p* con la proporzione campionaria $\hat{P} = \frac{X}{n}$ (facendo questa sostituzione si ottiene in effetti un intervallo di confidenza approssimato).
In questo modo, estraendo un campione di ampiezza *n* da una popolazione bernoulliana e indicando con $\hat{p}$ la proporzione del campione, si ottiene il seguente **intervallo di confidenza per la proporzione *p*** della popolazione bernoulliana, con grado di fiducia $(1-\alpha)\cdot 100\%$, valido per grandi campioni.

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{7.11}$$

Il valore critico $z_{\frac{\alpha}{2}}$ viene scelto con la stessa regola già indicata per l'intervallo di confidenza per la media, nel caso dei grandi campioni.
Osserviamo che per ottenere l'intervallo di confidenza (7.11) sono state fatte tre approssimazioni:
1 – l'approssimazione normale della binomiale;
2 – l'approssimazione di *p* con $\hat{p} = \frac{x}{n}$, nell'espressione $\sqrt{\frac{p(1-p)}{n}}$;

3 – non è stata fatta la correzione di continuità per l'approssimazione normale[2].
Questo implica che l'intervallo di confidenza trovato è un intervallo approssimato.
Per verificare le condizioni di applicabilità dell'approssimazione della binomiale con la normale, ossia $np \geq 5$ e $n(1-p) \geq 5$, possiamo solo verificare che sia $n\hat{p} \geq 5$ e $n(1-\hat{p}) \geq 5$; questa verifica si può fare solo dopo aver effettuato il campionamento: se le condizioni precedenti non sono soddisfatte, il risultato è privo di valore, e occorre ripetere il campionamento aumentando l'ampiezza *n* del campione.

Esempio 13
In un campione di 400 persone a cui è stato somministrato un dato vaccino, 136 di esse hanno avuto effetti collaterali di un certo rilievo. Determinare un intervallo di confidenza con grado di fiducia del 95% per la proporzione della popolazione che soffre di tali effetti collaterali.

Nel campione di $n = 400$ persone la proporzione campionaria è

$$\hat{p} = \frac{136}{400} = 0.34$$

Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$ e con la formula (7.11) si trova l'intervallo di confidenza

$$0.34 - 1.96 \cdot \sqrt{\frac{0.34 \cdot (1-0.34)}{400}} < p < 0.34 + 1.96 \cdot \sqrt{\frac{0.34 \cdot (1-0.34)}{400}}$$

$$0.29 < p < 0.39$$

Osserviamo che le condizioni per poter usare l'approssimazione della binomiale con la normale sono verificate, essendo

$$n\hat{p} = 400 \cdot 0.34 = 135 \qquad \text{e} \qquad n(1-\hat{p}) = 400 \cdot 0.66 = 264 .$$

Esempio 14
Un campione di 100 votanti scelto a caso fra tutti i votanti di una regione ha indicato che il 55% di essi è favorevole ad un certo candidato.
a – Determinare gli intervalli di confidenza con grado di fiducia del 95% e del 99% per la proporzione di tutti i votanti a favore del candidato.
b – Confrontare queste stime con la stima che si trova se si usa un campione di 2000 votanti, con la stessa percentuale campionaria di favorevoli.

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$; il risultato campionario indica che $\hat{p} = 0.55$ si e con la formula (7.11) si trova l'intervallo di confidenza

$$0.55 - 1.96 \cdot \sqrt{\frac{0.55 \cdot (1-0.55)}{100}} < p < 0.55 + 1.96 \cdot \sqrt{\frac{0.55 \cdot (1-0.55)}{100}}$$

$$0.45 < p < 0.65$$

Possiamo asserire con grado di fiducia del 95% che il candidato avrà a suo favore una percentuale di votanti compresa fra il 45% e il 65%.
Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$ e con la formula (7.11) si trova l'intervallo di confidenza

$$0.55 - 2.576 \cdot \sqrt{\frac{0.55 \cdot (1-0.55)}{100}} < p < 0.55 + 2.576 \cdot \sqrt{\frac{0.55 \cdot (1-0.55)}{100}}$$

$$0.42 < p < 0.68$$

[2] La correzione di continuità non comporta differenze rilevanti se *n* è grande.

Possiamo in questo caso asserire con grado di fiducia del 99% che il candidato avrà a suo favore una percentuale di votanti compresa fra il 42% e il 69%.
L'ampiezza degli intervalli di confidenza trovati è troppo grande, ossia la precisione delle stime è troppo bassa.
b – Se il campione è di 2000 votanti, con il grado di fiducia del 95% si trova il seguente intervallo di confidenza

$$0.55 - 1.96 \cdot \sqrt{\frac{0.55 \cdot (1-0.55)}{2000}} < p < 0.55 + 1.96 \cdot \sqrt{\frac{0.55 \cdot (1-0.55)}{2000}}$$

$$0.52 < p < 0.58$$

In questo caso, con un grado di fiducia del 95%, il candidato avrà a suo favore una percentuale di votanti compresa fra il 52% e il 58%, con una stima decisamente più precisa. La maggior precisione dipende dalla maggiore ampiezza del campione.

Con lo stesso procedimento già usato nel caso dell'intervallo di confidenza per la media di un grande campione, si può usare la disuguaglianza (7.10), valida con probabilità $1-\alpha$, per ricavare una formula che consente di determinare l'**ampiezza** n **del campione** necessaria per ottenere un errore prefissato.
La (7.10) equivale a

$$\frac{\left|\hat{P} - p\right|}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}$$

ossia

$$\left|\hat{P} - p\right| < z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Indicando con

$$E = \max \left|\hat{P} - p\right|$$

il massimo dell'errore che si commette approssimando la proporzione della popolazione p con la proporzione campionaria $\hat{P} = \frac{X}{n}$, la stima di E con probabilità $1 - \alpha$ è data da

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \tag{7.12}$$

In altre parole, se si vuole stimare la proporzione p della popolazione con la proporzione campionaria $\hat{p} = \frac{x}{n}$ di un campione di ampiezza n ($n \geq 30$), si può affermare, con probabilità $1-\alpha$, che l'errore $\left|\frac{X}{n} - p\right|$ sarà al più uguale a $z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$.
Dalla formula (7.12), risolvendo rispetto a n, si ricava l'ampiezza del campione necessaria per stimare la proporzione p con un errore prefissato E e con un dato grado di fiducia (si ricordi che n deve essere un intero)

$$n \geq p(1-p) \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \tag{7.13}$$

Questa formula non può essere usata se non si ha qualche informazione sul valore di *p*; se tali informazioni non sono disponibili, si può far uso del fatto che il valore massimo[3] che può assumere la quantità $p(1-p)$ è $\frac{1}{4}$, corrispondente a $p = \frac{1}{2}$.
In questo caso l'ampiezza necessaria per il campione è (si ricordi che *n* deve essere un intero)

$$n \geq \frac{1}{4}\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \tag{7.14}$$

Esempio 15
Problema del sondaggio di opinione. Supponiamo che si voglia stimare la proporzione di elettori che approva l'operato del capo del governo; su un campione di 150 persone intervistate, 90 si sono dichiarate favorevoli.
Determinare un intervallo di confidenza con grado di fiducia del 95% per la proporzione degli elettori favorevoli al capo del governo e valutare la precisione della stima.

La proporzione campionaria dei favorevoli è

$$\hat{p} = \frac{x}{n} = \frac{90}{150} = 0.6$$

L'intervallo di confidenza con grado di fiducia del 95% è il seguente

$$0.6 - 1.96 \cdot \sqrt{\frac{0.6 \cdot (1-0.6)}{150}} < p < 0.6 + 1.96 \cdot \sqrt{\frac{0.6 \cdot (1-0.6)}{150}}$$

$$0.52 < p < 0.68$$

La percentuale dei favorevoli, con un grado di fiducia del 95%, è compresa fra il 52% e il 68%: la stima è troppo imprecisa, l'ampiezza dell'intervallo è di 16 punti percentuali.
Può quindi essere utile determinare l'ampiezza del campione necessaria per ottenere una stima con precisione fissata. Stabiliamo ad esempio che si vuole una stima con una precisione dell'1% (corrispondente a un'ampiezza dell'intervallo non superiore a 2 punti percentuali), ossia fissiamo

$$E = 0.01.$$

Dato che non abbiamo informazioni circa la percentuale dei favorevoli nel nuovo campione, dobbiamo usare la formula (7.14) e in tal caso, per il grado di fiducia del 95% , si ottiene

$$n \geq \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604$$

Esempio 16
Supponiamo di voler stimare la proporzione di pezzi difettosi in un lotto di oggetti di un dato tipo con un errore $E = 0.04$ e un grado di fiducia del 95%; calcolare l'ampiezza necessaria per il campione, nel caso che
a – non si abbia alcuna informazione su quale possa essere la proporzione effettiva della popolazione;
b – si sappia che la proporzione della popolazione non supera il 12%.

[3] Per verificare questo fatto, ricordando che *p* è una probabilità e può quindi assumere solo valori compresi fra 0 e 1, basta cercare il massimo della funzione $f(p) = p(1-p)$ nell'intervallo (0, 1); tale massimo è $\frac{1}{4}$ e viene assunto per $p = \frac{1}{2}$.

a – Se non si ha alcuna informazione su *p*, si usa la formula (7.14), e con grado di fiducia del 95% si ricava

$$n \geq \frac{1}{4}\left(\frac{1.96}{0.04}\right)^2 = 600.3$$

Occorre quindi un campione di ampiezza $n = 601$.
b – Se sappiamo che $p \leq 0.12$, con la formula (7.13) e con grado di fiducia del 95% si ottiene

$$n \geq 0.12(1-0.12)\left(\frac{1.96}{0.04}\right)^2 = 253.5 .$$

Occorre in questo caso un campione di ampiezza $n = 254$.
Questo esempio illustra come il fatto di avere qualche informazione sul possibile valore della proporzione può sensibilmente ridurre la dimensione del campione.

7.6 Intervalli di confidenza per la differenza fra due medie (varianze note)

Molto spesso in una ricerca si è interessati a due popolazioni; in particolare si vuole studiare la differenza fra le medie di due popolazioni: in una indagine, per esempio, si può cercare di stabilire se le medie di due popolazioni sono diverse oppure si vuole stimare la grandezza della differenza fra le medie di due popolazioni.
In ricerche di questo genere è necessario conoscere le proprietà della distribuzione di campionamento della differenza fra due medie.

Date due distribuzioni aventi medie rispettivamente μ_1 e μ_2 e varianze σ_1^2 e σ_2^2, ricordiamo che vale la seguente proprietà[4].

Proprietà 1
Se le distribuzioni di due variabili aleatorie indipendenti hanno le medie μ_1 e μ_2 e le varianze σ_1^2 e σ_2^2, allora la distribuzione della loro differenza ha la media $\mu_1 - \mu_2$ e la varianza $\sigma_1^2 + \sigma_2^2$.

Date due popolazioni aventi distribuzioni normale, si estraggano da esse campioni di ampiezza rispettivamente n_1 e n_2; indicando con $\overline{X}_1$ e $\overline{X}_2$ le due medie campionarie, in base allo schema (Cap. 6, pag. 178) che riassume le proprietà della distribuzione della media campionaria, possiamo affermare che $\overline{X}_1$ e $\overline{X}_2$ hanno entrambe distribuzione normale con medie rispettivamente μ_1 e μ_2 e varianze $\frac{\sigma_1^2}{n_1}$ e $\frac{\sigma_2^2}{n_2}$; lo stesso risultato vale, almeno approssimativamente, per grandi campioni estratti da popolazioni non aventi la distribuzione normale. In entrambi i casi, la differenza $\overline{X}_1 - \overline{X}_2$ ha, almeno approssimativamente, la distribuzione normale e, in base alla precedente proprietà, la media è $\mu_1 - \mu_2$ e la varianza è $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

Possiamo allora considerare la statistica

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (7.15)$$

che ha almeno approssimativamente la distribuzione normale standardizzata.

[4] Si ricordino le formule (3.23), pag.117.

Procedendo come già visto per ricavare l'intervallo di confidenza per la media possiamo asserire, con probabilità uguale a $1 - \alpha$, che è soddisfatta la disuguaglianza

$$-z_{\frac{\alpha}{2}} < \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\frac{\alpha}{2}}. \tag{7.16}$$

Dalla disuguaglianza (7.16), risolvendo rispetto a $\mu_1 - \mu_2$ si ottiene

$$\overline{X}_1 - \overline{X}_2 - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Pertanto, una volta estratti i campioni di ampiezza rispettivamente n_1 e n_2, , e calcolati i valori $\bar{x}_1$ e $\bar{x}_2$ delle medie dei due campioni, si ottiene il seguente **intervallo di confidenza per la differenza delle medie $\mu_1 - \mu_2$**, con **grado di fiducia** $(1-\alpha)\cdot 100\%$.

$$\bar{x}_1 - \bar{x}_2 - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7.17}$$

La formula (7.17) vale per popolazioni anche non normali, purché i campioni siano grandi ($n_1, n_2 \geq 30$).

Se le popolazioni da cui provengono i campioni hanno distribuzione normale, la (7.17) vale qualunque siano le dimensioni dei campioni.

I valori più comunemente usati per il grado di fiducia sono, come al solito, il 90%, il 95% o il 99%; i corrispondenti valori di $z_{\frac{\alpha}{2}}$ sono i seguenti

grado di fiducia del 90% $\quad z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

grado di fiducia del 95% $\quad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

grado di fiducia del 99% $\quad z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$

L'applicazione della (7.17) richiede la conoscenza delle varianze delle popolazioni; se invece le varianze σ_1^2 e σ_2^2 non sono note, nel caso di grandi campioni possono essere sostituite con le varianze campionarie s_1^2 e s_2^2.

Esempio 17

Un campione di 200 lampadine della marca A ha mostrato una durata media di 1500 ore ed uno scarto quadratico medio di 100 ore; un campione di 150 lampadine della marca B ha mostrato invece una durata media di 1300 ore ed uno scarto quadratico medio di 90 ore. Trovare gli intervalli di confidenza al 95% e al 99% per la differenza di durata di tutte le lampadine delle marche A e B.

I dati del problema sono i seguenti

$n_1 = 200 \qquad \bar{x}_1 = 1500 \qquad s_1 = 100$

$n_2 = 150 \qquad \bar{x}_2 = 1300 \qquad s_2 = 90$

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Applicando la formula (7.17) si deve sostituire alla varianza della popolazione, che non è nota, la varianza del campione; si ottiene l'intervallo di confidenza

$$1500-1300-1.96\cdot\sqrt{\frac{100^2}{200}+\frac{90^2}{150}}<\mu_1-\mu_2<1500-1300+1.96\cdot\sqrt{\frac{100^2}{200}+\frac{90^2}{150}}$$

$$180<\mu_1-\mu_2<220$$

b – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = z_{0.025} = 2.576$ e si ottiene l'intervallo di confidenza

$$1500-1300-2.576\cdot\sqrt{\frac{100^2}{200}+\frac{90^2}{150}}<\mu_1-\mu_2<1500-1300+2.576\cdot\sqrt{\frac{100^2}{200}+\frac{90^2}{150}}$$

$$173<\mu_1-\mu_2<227$$

Esempio 18

Nella fase di test di un nuovo farmaco due gruppi simili di pazienti, A e B, composti rispettivamente di 50 e 100 individui, hanno partecipato alla sperimentazione: il primo gruppo è stato sottoposto ad una cura con un nuovo tipo di sonnifero, mentre il secondo è stato curato con un tipo convenzionale di sonnifero. Per i pazienti del gruppo A, il numero medio di ore di sonno per notte è stato di 7.5 ore con uno scarto quadratico medio di 0.25 ore. Per i pazienti del gruppo B, il numero medio di ore di sonno è stato di 6.7 ore con uno scarto quadratico medio di 0.30 ore.
Trovare gli intervalli di confidenza al 95% e al 99% per la differenza tra i numeri medi di ore di sonno.

I dati del problema sono i seguenti

$$n_1 = 50 \qquad \bar{x}_1 = 7.5 \qquad s_1 = 0.25$$

$$n_2 = 100 \qquad \bar{x}_2 = 6.7 \qquad s_2 = 0.30$$

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ e si ottiene l'intervallo di confidenza

$$7.5-6.7-1.96\cdot\sqrt{\frac{0.25^2}{50}+\frac{0.30^2}{100}}<\mu_1-\mu_2<7.5-6.7+1.96\cdot\sqrt{\frac{0.25^2}{50}+\frac{0.30^2}{100}}$$

$$0.70<\mu_1-\mu_2<0.90$$

b – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = z_{0.025} = 2.576$ e si ottiene l'intervallo di confidenza

$$7.5-6.7-2.576\cdot\sqrt{\frac{0.25^2}{50}+\frac{0.30^2}{100}}<\mu_1-\mu_2<7.5-6.7+2.576\cdot\sqrt{\frac{0.25^2}{50}+\frac{0.30^2}{100}}$$

$$0.68<\mu_1-\mu_2<0.92$$

7.7 Intervalli di confidenza per la differenza fra due medie (varianze incognite)

L'applicazione della (7.17) richiede la conoscenza delle varianze delle popolazioni; se le varianze σ_1^2 e σ_2^2 non sono note, nel caso di grandi campioni possono essere sostituite con le varianze campionarie s_1^2 e s_2^2 (vedere il § precedente e gli esempi 17, 18).
Nel caso si tratti invece di piccoli campioni e le varianze non siano note, per stimare la differenza fra le medie delle due popolazioni si può far ricorso alla distribuzione *t*, ma occorre che siano verificate alcune ipotesi.
Innanzi tutto le due popolazioni devono avere distribuzione normale; inoltre occorre distinguere due casi: il caso in cui le varianze delle due popolazioni sono uguali e il caso in cui sono diverse.

In queste lezioni sarà trattato solo il caso di due popolazioni normali con la stessa varianza.
Se le due popolazioni normali hanno la stessa varianza (incognita), le due varianze campionarie S_1^2 e S_2^2, che si calcolano dai campioni indipendenti estratti dalle due popolazioni, possono essere considerate come stime della stessa quantità, la varianza comune alle due distribuzioni.
Basandoci su questa osservazione si può ricavare una **stima congiunta della varianza** comune, calcolando la media ponderata delle due varianze campionarie con la seguente formula

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{7.18}$$

Ciascuna delle due varianze campionarie è ponderata con il suo grado di libertà. Se i due campioni hanno la stessa ampiezza, la stima congiunta è la media aritmetica delle due varianze campionarie; se invece hanno ampiezze diverse, la media ponderata è maggiormente influenzata dall'informazione fornita dal campione più grande.
Per **piccoli campioni** (***n* < 30**), nell'ipotesi che le popolazioni da cui si estraggono i campioni abbiano distribuzione normale con la stessa varianza, indicando con $\overline{X}_1$ e $\overline{X}_2$ le medie campionarie e con S^2 la stima congiunta della varianza, si può dimostrare che la statistica

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7.19}$$

ha la distribuzione *t* con grado di libertà $n_1 + n_2 - 2$.
Pertanto, con procedimento analogo a quello del § precedente, una volta estratti i campioni di ampiezza rispettivamente n_1 e n_2, , e calcolati i valori $\bar{x}_1$ e $\bar{x}_2$ delle medie dei due campioni, i valori s_1^2 e s_2^2 delle due varianze, e il valore s^2 della stima congiunta della varianza, si ottiene il seguente **intervallo di confidenza per la differenza delle medie $\mu_1 - \mu_2$**, con **grado di fiducia** $(1-\alpha)\cdot 100\%$, per **piccoli campioni** estratti da due popolazioni normali con la stessa varianza.

$$\bar{x}_1 - \bar{x}_2 - t_{\frac{\alpha}{2}} \cdot \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + t_{\frac{\alpha}{2}} \cdot \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{7.20}$$

I valori più comunemente usati per il grado di fiducia sono, come al solito, il 90%, il 95% o il 99%; i corrispondenti valori di $t_{\frac{\alpha}{2}}$ sono i seguenti

grado di fiducia del 90% $\quad t_{\frac{\alpha}{2}} = t_{0.05}$

grado di fiducia del 95% $\quad t_{\frac{\alpha}{2}} = t_{0.025}$

grado di fiducia del 99% $\quad t_{\frac{\alpha}{2}} = t_{0.005}$

Questi valori possono essere letti sulla tabella della distribuzione *t* in corrispondenza al grado di libertà $\nu = n_1 + n_2 - 2$. Il valore del grado di libertà può essere maggiore di 29: in tal caso si utilizzano i valori critici dell'ultima riga della tabella della distribuzione *t*.

Esempio 19
Nella tabella 3 sono riportate le lunghezze in cm di due campioni A e B di oggetti dello stesso tipo prodotti da due macchine diverse.

A	8.26	8.13	8.35	8.07	8.34		
B	7.95	7.89	7.90	8.14	7.92	7.84	7.94

Tabella 3

Calcolare gli intervalli di confidenza per la differenza fra le medie con grado di fiducia del 95% e del 99%, supponendo che le popolazioni da cui provengono i campioni abbiano distribuzione normale con la stessa varianza.

In base ai dati della tabella si ha

$$n_1 = 5 \qquad \bar{x}_1 = 8.23 \qquad s_1^2 = 0.01575$$
$$n_2 = 7 \qquad \bar{x}_2 = 7.94 \qquad s_2^2 = 0.00910$$

La stima congiunta della varianza con la formula (7.18) è

$$s^2 = \frac{4 \cdot 0.01575 + 6 \cdot 0.00910}{5 + 7 - 2} = 0.01176$$

Il grado di libertà della distribuzione *t* è

$$\nu = n_1 + n_2 - 2 = 5 + 7 - 2 = 10 .$$

a – Per il grado di fiducia del 95% il valore critico è $t_{0.025} = 2.228$ e con la (7.15) si trova l'intervallo di confidenza seguente

$$8.23 - 7.94 - 2.228\sqrt{0.01176\left(\frac{1}{5} + \frac{1}{7}\right)} < \mu_1 - \mu_2 < 8.23 - 7.94 + 2.228\sqrt{0.01176\left(\frac{1}{5} + \frac{1}{7}\right)}$$

$$0.148 < \mu_1 - \mu_2 < 0.432$$

b – Per il grado di fiducia del 99% il valore critico è invece $t_{0.005} = 3.169$ e si trova l'intervallo di confidenza seguente

$$8.23 - 7.94 - 3.169\sqrt{0.01176\left(\frac{1}{5} + \frac{1}{7}\right)} < \mu_1 - \mu_2 < 8.23 - 7.94 + 3.169\sqrt{0.01176\left(\frac{1}{5} + \frac{1}{7}\right)}$$

$$0.088 < \mu_1 - \mu_2 < 0.492$$

7.8 Intervalli di confidenza per la differenza fra due proporzioni

Spesso si è interessati alla stima della differenza fra le proporzioni di due popolazioni. Possiamo voler confrontare, per esempio, due gruppi di età, due gruppi di diverso sesso o due gruppi diagnostici rispetto alla proporzione di coloro che possiedono una qualche caratteristica di interesse.

Uno stimatore puntuale corretto della differenza fra le proporzioni p_1 e p_2 di due popolazioni è fornito dalla differenza fra le proporzioni campionarie $\hat{p}_1 - \hat{p}_2$.

Se, come abbiamo già visto nel caso dell'intervallo di confidenza per la proporzione, le ampiezze n_1 e n_2 dei campioni sono grandi e le proporzioni delle popolazioni non sono troppo vicine a 0 o a 1 (ossia sono soddisfatte condizioni del tipo $np \geq 5$ e $n(1-p) \geq 5$), si può ricorrere all'approssimazione della distribuzione binomiale con la distribuzione normale per ricavare l'intervallo di confidenza per differenza fra due proporzioni.

Si può dimostrare che la statistica

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \tag{7.21}$$

ha approssimativamente la distribuzione normale standardizzata, per valori sufficientemente grandi di n_1 e n_2.

Con un procedimento analogo a quello seguito per ricavare l'intervallo di confidenza per la proporzione (§ 7.5, pag. 201) si può ricavare il seguente **intervallo di confidenza per la differenza fra due proporzioni $p_1 - p_2$**, con grado di fiducia $(1-\alpha)\cdot 100\%$, valido per **grandi campioni**

$$(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad (7.22)$$

Osserviamo che le quantità p_1 e p_2 che compaiono al denominatore nell'espressione (7.21) sono state approssimate con le rispettive proporzioni campionarie $\hat{p}_1 = \frac{x_1}{n_1}$ e $\hat{p}_2 = \frac{x_2}{n_2}$, ottenendo così un intervallo di confidenza approssimato.
Il valore critico $z_{\frac{\alpha}{2}}$ viene scelto con la stessa regola già indicata per l'intervallo di confidenza per la media, nel caso dei grandi campioni.

Esempio 20
In un campione casuale di 600 adolescenti e 400 adulti che seguono un certo programma televisivo, 300 adolescenti e 100 adulti hanno espresso un parere favorevole al programma stesso.
Trovare gli intervalli di confidenza al 95% e al 99% per la differenza fra le proporzioni degli adulti favorevoli e degli adolescenti favorevoli al programma.

I dati del problema sono i seguenti

$$\hat{p}_1 = \frac{300}{600} = 0.5 \qquad \hat{p}_2 = \frac{100}{400} = 0.25$$

a – Per il grado di fiducia del 95% con la (7.22) si trova l'intervallo di confidenza

$$0.5 - 0.25 - 1.96\sqrt{\frac{0.5\cdot 0.5}{600} + \frac{0.25\cdot 0.75}{400}} < p_1 - p_2 < 0.5 - 0.25 + 1.96\sqrt{\frac{0.5\cdot 0.5}{600} + \frac{0.25\cdot 0.75}{400}}$$

$$0.19 < p_1 - p_2 < 0.31$$

Esempio 21
Una macchina per lo stampaggio di parti in plastica viene sottoposta a una modifica nel processo di lavorazione. In un campione di 85 pezzi scelti prima della modifica, 10 sono difettosi, mentre in un campione di 85 pezzi scelti dopo la modifica 8 sono difettosi.
Trovare un intervallo di confidenza con grado di fiducia del 95% per la differenza fra le proporzioni di pezzi difettosi prima e dopo l'intervento.

I dati del problema sono i seguenti

$$\hat{p}_1 = \frac{10}{85} = 0.118 \qquad \hat{p}_2 = \frac{8}{85} = 0.094$$

Per il grado di fiducia del 95% con la (7.22) si trova l'intervallo di confidenza

$$0.118 - 0.094 - 1.96\sqrt{\frac{0.118\cdot(1-0.118)}{85} + \frac{0.094\cdot(1-0.094)}{85}} < p_1 - p_2 <$$

$$< 0.118 - 0.094 + 1.96\sqrt{\frac{0.118\cdot(1-0.118)}{85} + \frac{0.094\cdot(1-0.094)}{85}}$$

$$-0.068 < p_1 - p_2 < 0.117$$

Questo intervallo comprende lo zero, perciò, basandoci su questi due campioni, sembra improbabile che la modifica nel processo di lavorazione abbia diminuito la proporzione di pezzi difettosi.

7.9 Intervalli di confidenza per la varianza e per lo scarto quadratico medio

Nel calcolo dell'intervallo di confidenza per la media di un grande campione si è osservato che, se lo scarto quadratico medio della popolazione non è noto, esso può essere sostituito con lo scarto quadratico medio campionario.
In certi casi è però necessario determinare intervalli di confidenza per la varianza o per lo scarto quadratico medio.
Nella maggior parte delle applicazioni pratiche, le stime per intervallo per σ e σ^2 sono basate sullo scarto quadratico medio campionario s e sulla varianza campionaria s^2.
Si consideri una popolazione avente distribuzione normale, e si estraggano da questa popolazione campioni di ampiezza n.
In base al teorema 4, Cap. 6, si può affermare che la statistica

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \tag{7.23}$$

ha la distribuzione χ^2 con grado di libertà $\nu = n - 1$.
Come già osservato, la distribuzione χ^2 non è simmetrica (si veda l'esempio 20, Cap. 6); usando code di uguale area e indicando con $\frac{\alpha}{2}$ l'area di ciascuna coda (figura 3), si ha che

$$P\left(\chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)\cdot S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}\right) = 1-\alpha$$

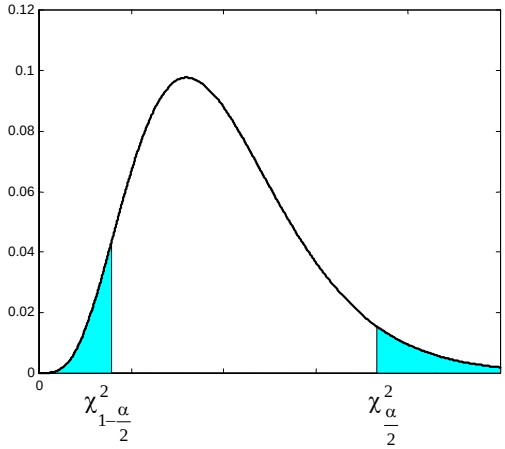


Figura 3

In altre parole si può asserire con probabilità $1 - \alpha$, ossia con grado di fiducia $(1-\alpha)\cdot 100\%$, che vale la disuguaglianza

$$\chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)\cdot S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}$$

Pertanto, indicando con s^2 la varianza di un campione di ampiezza n estratto da una popolazione normale, e risolvendo questa disuguaglianza rispetto a σ^2 si ottiene l'**intervallo di confidenza per la varianza σ^2** con grado di fiducia $(1-\alpha)\cdot 100\%$

$$\frac{(n-1)\cdot s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi^2_{1-\frac{\alpha}{2}}} \tag{7.24}$$

Estraendo la radice quadrata di ciascun membro della disuguaglianza, si ottiene l'intervallo di confidenza per lo scarto quadratico medio.

Il metodo descritto per trovare gli intervalli di confidenza per la varianza si applica solo a campioni estratti da popolazioni normali.
L'assunzione che la popolazione abbia distribuzione normale è molto importante: infatti i risultati ottenuti ignorando tale ipotesi possono portare a gravi errori.
Si osservi inoltre che l'intervallo di confidenza non è simmetrico, come invece accade per gli intervalli di confidenza per la media o per la proporzione; ciò è dovuto al fatto che la distribuzione χ^2 non è simmetrica.
I valori più comunemente usati per $1 - \alpha$ sono 0.90, 0.95 e 0.99, a cui corrispondono i gradi di fiducia del 90%, del 95% e del 99%; i corrispondenti valori di $\chi^2_{\frac{\alpha}{2}}$ e di $\chi^2_{1-\frac{\alpha}{2}}$ sono

grado di fiducia del 90%	$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.05}$	$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.95}$
grado di fiducia del 95%	$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025}$	$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975}$
grado di fiducia del 99%	$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005}$	$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995}$

Questi valori possono essere letti sulla tabella della distribuzione χ^2 in corrispondenza al grado di libertà $\nu = n - 1$.

Esempio 22
In una scuola è stato scelto a caso un campione di 16 studenti dell'ultimo anno e si è misurata l'altezza di ciascuno di essi. La varianza campionaria della misura delle altezze è $s^2 = 37.09$ cm^2.
Trovare gli intervalli di confidenza al 95% e al 99% per la varianza della popolazione costituita da tutti gli studenti dell'ultimo anno della scuola.

Poiché si tratta di misure, si può ragionevolmente ipotizzare che la popolazione da cui proviene il campione abbia distribuzione normale.
a – Per il grado di fiducia del 95% e il grado di libertà $\nu = n - 1 = 15$, si ha

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975} = 6.262 \qquad \chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025} = 27.488$$

Con la formula (7.24) si ottiene l'intervallo di confidenza per la varianza con grado di fiducia del 95%

$$\frac{15 \cdot 37.09}{27.488} < \sigma^2 < \frac{15 \cdot 37.09}{6.262}$$

$$20.23 < \sigma^2 < 88.84$$

b – Per il grado di fiducia del 99% e il grado di libertà $\nu = n - 1 = 15$, si ha

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995} = 4.601 \qquad \chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005} = 32.801$$

Con la formula (7.24) si ottiene l'intervallo di confidenza per la varianza con grado di fiducia del 99%

$$\frac{15 \cdot 37.09}{32.801} < \sigma^2 < \frac{15 \cdot 37.09}{4.601}$$

$$16.96 < \sigma^2 < 120.92$$

Il corrispondente intervallo di confidenza per lo scarto quadratico medio è

$$4.11 < \sigma < 11.00$$

Esempio 23
Lo scarto quadratico medio della durata di un campione di 25 lampadine è $s = 100$ ore.
Trovare l'intervallo di confidenza al 95% per la varianza della popolazione.

Poiché si tratta di misure, si può ragionevolmente ipotizzare che la popolazione da cui proviene il campione abbia distribuzione normale.
Per il grado di fiducia del 95% e il grado di libertà $\nu = n - 1 = 24$, si ha

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975} = 12.401 \qquad \chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025} = 39.364$$

Con la formula (7.24) si ottiene l'intervallo di confidenza per la varianza con grado di fiducia del 95%

$$\frac{24 \cdot 100^2}{39.364} < \sigma^2 < \frac{24 \cdot 100^2}{12.401}$$

$$6096.94 < \sigma^2 < 19353.28$$

Il corrispondente intervallo di confidenza per lo scarto quadratico medio è

$$78.08 < \sigma < 139.12$$

Esempio 24
Le misure della durata in ore di 10 batterie sono le seguenti

140	136	150	144	148	152	138	141	143	151

Trovare un intervallo di confidenza al 99% per la varianza e per lo scarto quadratico medio della popolazione.

Calcoliamo la media campionaria e la varianza campionaria

$$\bar{x} = \frac{140+136+150+144+148+152+138+141+143+151}{10} = 144.3$$

$$s^2 = \frac{1}{9}\left[140^2 + 136^2 + 150^2 + 144^2 + 148^2 + 152^2 + 138^2 + 141^2 + \right.$$
$$\left. + 143^2 + 151^2 - 10 \cdot 144.3^2 \right] = 32.233$$

Per il grado di fiducia del 99% e il grado di libertà $\nu = n - 1 = 9$, si ha

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995} = 1.735 \qquad \chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005} = 23.589$$

Con la formula (7.24) si ottiene l'intervallo di confidenza per la varianza con grado di fiducia del 99%

$$\frac{9 \cdot 32.233}{23.589} < \sigma^2 < \frac{9 \cdot 32.233}{1.735}$$

$$12.29 < \sigma^2 < 167.21$$

Per lo scarto quadratico medio si ha

$$3.50 < \sigma < 12.94 .$$

Esempio 25
Cinque studenti effettuano in modo indipendente il calcolo approssimato del numero π e trovano i seguenti valori

3.12 3.16 2.94 3.20 3.33

Trovare un intervallo di confidenza per il numero π (ossia per la media) e un intervallo di confidenza per lo scarto quadratico medio, con grado di fiducia del 95% e del 99%.

Ipotizziamo che il campione sia tratto da una popolazione normale.
Calcoliamo la media campionaria e la varianza campionaria

$$\bar{x} = \frac{3.12 + 3.16 + 2.94 + 3.20 + 3.33}{5} = 3.15$$

$$s^2 = \frac{1}{4}\left[3.12^2 + 3.16^2 + 2.94^2 + 3.20^2 + 3.33^2 - 5 \cdot 3.15^2\right] = 0.02$$

a – Intervalli di confidenza per la media.

Per il grado di fiducia del 95% e il grado di libertà $\nu = 4$ si ha $t_{\frac{\alpha}{2}} = t_{0.025} = 2.776$

Con la formula (7.8) si ottiene l'intervallo di confidenza per la media

$$3.15 - 2.776 \cdot \frac{\sqrt{0.02}}{\sqrt{5}} < \mu < 3.15 + 2.776 \cdot \frac{\sqrt{0.02}}{\sqrt{5}}$$

$$2.97 < \mu < 3.33$$

Per il grado di fiducia del 99% e il grado di libertà $\nu = 4$ si ha $t_{\frac{\alpha}{2}} = t_{0.005} = 4.604$

L'intervallo di confidenza al 99% per la media è

$$3.15 - 4.604 \cdot \frac{\sqrt{0.02}}{\sqrt{5}} < \mu < 3.15 + 4.604 \cdot \frac{\sqrt{0.02}}{\sqrt{5}}$$

$$2.85 < \mu < 3.45$$

b – Intervalli di confidenza per la varianza.

Per il grado di fiducia del 95% e il grado di libertà $\nu = n - 1 = 4$, si ha

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975} = 0.484 \qquad \chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025} = 11.143$$

Con la formula (7.24) si ottiene l'intervallo di confidenza per la varianza con grado di fiducia del 95%

$$\frac{4 \cdot 0.02}{11.143} < \sigma^2 < \frac{4 \cdot 0.02}{0.484}$$

$$0.00718 < \sigma^2 < 0.1653$$

L'intervallo di confidenza per lo scarto quadratico medio è

$$0.084 < \sigma < 0.41$$

Per il grado di fiducia del 99% e il grado di libertà $\nu = n - 1 = 4$, si ha

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975} = 0.207 \qquad \chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025} = 14.860$$

Con la formula (7.24) si ottiene l'intervallo di confidenza per la varianza con grado di fiducia del 99%

$$\frac{4 \cdot 0.02}{14.860} < \sigma^2 < \frac{4 \cdot 0.02}{0.207}$$

$$0.00538 < \sigma^2 < 0.386$$

L'intervallo di confidenza per lo scarto quadratico medio è

$$0.073 < \sigma < 0.63$$

La formula (7.24) per trovare l'intervallo di confidenza per la varianza e lo scarto quadratico medio, pur essendo valida sia per piccoli che per grandi campioni, viene di solito utilizzata solo per piccoli campioni e, come già sottolineato, nel caso in cui la popolazione da cui proviene il campione sia normale.

Per grandi campioni estratti da una popolazione normale, si può dimostrare che la distribuzione campionaria S della deviazione standard σ può essere approssimata con una distribuzione normale avente media σ e deviazione standard $\frac{\sigma}{\sqrt{2n}}$, ossia la statistica

$$Z = \frac{S - \sigma}{\frac{\sigma}{\sqrt{2n}}} \tag{7.25}$$

ha approssimativamente la distribuzione normale standardizzata, per *n* sufficientemente grande.
Si può pertanto asserire che, con probabilità $1 - \alpha$, vale la disuguaglianza

$$-z_{\frac{\alpha}{2}} < \frac{S - \sigma}{\frac{\sigma}{\sqrt{2n}}} < z_{\frac{\alpha}{2}}$$

Risolvendo la disuguaglianza rispetto a σ, e indicando con *s* lo scarto quadratico medio di un campione di ampiezza *n*, si trova l'**intervallo di confidenza per lo scarto quadratico medio σ**, per **grandi campioni**, con probabilità $1 - \alpha$, o con grado di fiducia $(1-\alpha)\cdot 100\%$

$$\frac{s}{1+\frac{z_{\frac{\alpha}{2}}}{\sqrt{2n}}} < \sigma < \frac{s}{1-\frac{z_{\frac{\alpha}{2}}}{\sqrt{2n}}} \tag{7.26}$$

I valori di $z_{\frac{\alpha}{2}}$ in base al grado di fiducia fissato sono i seguenti

grado di fiducia del 90% $\quad z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

grado di fiducia del 95% $\quad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

grado di fiducia del 99% $\quad z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$

Esempio 26
Determinare un intervallo di confidenza con grado di fiducia del 95% per lo scarto quadratico medio della popolazione da cui è stato estratto il campione studiato nell'esempio 2, Cap 1.

Per questo campione di ampiezza $n = 80$ (grande campione) si è calcolato (esempio 34, Cap. 1) la varianza campionaria

$$s^2 = 31.96$$

Per il grado di fiducia del 95% si ha $z_{\frac{\alpha}{2}} = 1.96$.

Con la formula (7.26) si trova l'intervallo di confidenza per lo scarto quadratico medio con grado di fiducia del 95%.

$$\frac{\sqrt{31.96}}{1+\frac{1.96}{\sqrt{160}}} < \sigma < \frac{\sqrt{31.96}}{1-\frac{1.96}{\sqrt{160}}}$$

$$4.89 < \sigma < 6.69$$

Esempio 27
Lo scarto quadratico medio della durata di un campione di 200 lampadine è $s = 100$ ore.
Trovare l'intervallo di confidenza al 95% per lo scarto quadratico medio dell'intera popolazione.

Poiché l'ampiezza del campione è $n = 200$, si tratta di un grande campione; per il grado di fiducia del 95% si ha $z_{\frac{\alpha}{2}} = 1.96$ e con la formula (7.26) si trova l'intervallo di confidenza

$$\frac{100}{1+\frac{1.96}{\sqrt{400}}} < \sigma < \frac{100}{1-\frac{1.96}{\sqrt{400}}}$$

$$91 < \sigma < 111$$

Esempio 28
Un campione di 32 misurazioni del punto di bollitura di una sostanza chimica ha scarto quadratico medio s = 0.83°C.
Determinare un intervallo di confidenza al 99% per lo scarto quadratico medio σ.

Poiché l'ampiezza del campione è n = 32, si può usare la formula (7.26); per il grado di fiducia del 99% si ha $z_{\frac{\alpha}{2}}$ = 2.576.

L'intervallo di confidenza per lo scarto quadratico medio σ è

$$\frac{0.83}{1+\frac{2.576}{\sqrt{64}}} < \sigma < \frac{0.83}{1-\frac{2.576}{\sqrt{64}}}$$

$$0.62 < \sigma < 1.23$$

7.10 Intervalli di confidenza per il rapporto di due varianze

Per confrontare fra loro due varianze si costruisce il loro rapporto

$$\frac{\sigma_1^2}{\sigma_2^2}$$

Se le due varianze sono uguali, il loro rapporto sarà uguale a 1; di solito però non si conoscono le varianze delle popolazioni studiate, e il confronto avverrà sulla base delle varianze campionarie, ossia si procede a una stima del rapporto delle varianze delle due popolazioni.
Si considerino due popolazioni avente distribuzione normale, e si estraggano da queste popolazioni campioni indipendenti di ampiezza rispettivamente n_1 e n_2. Le varianze campionarie siano rispettivamente S_1^2 e S_2^2.
In base al teorema 5, Cap. 6, si può affermare che la statistica

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \tag{7.27}$$

ha la **distribuzione *F*** di parametri $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$.
Si può osservare che la distribuzione F non è simmetrica perciò, con lo stesso tipo di procedimento già utilizzato per ricavare gli intervalli di confidenza per la varianza, usando code di uguale area e indicando con $\frac{\alpha}{2}$ l'area di ciascuna coda, si ha che

$$P\left(F_{1-\frac{\alpha}{2}} < \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} < F_{\frac{\alpha}{2}} \right) = 1 - \alpha .$$

In altre parole si può asserire con probabilità $1 - \alpha$, ossia con grado di fiducia $(1-\alpha)\cdot 100\%$, che vale la disuguaglianza

$$F_{1-\frac{\alpha}{2}} < \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} < F_{\frac{\alpha}{2}}$$

Risolvendo questa disuguaglianza rispetto a $\frac{\sigma_2^2}{\sigma_1^2}$ si ha

$$\frac{S_2^2}{S_1^2}F_{1-\frac{\alpha}{2}} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2}F_{\frac{\alpha}{2}}$$

e prendendo i reciproci dei tre termini si ha

$$\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}}$$

Pertanto, estraendo due campioni indipendenti di ampiezza n_1 e n_2 da due popolazioni normali e indicando con s_1^2 e s_2^2 le varianze dei due campioni, dove s_1^2 è la più grande delle due varianze, si ottiene l'**intervallo di confidenza per il rapporto di due varianze** $\frac{\sigma_1^2}{\sigma_2^2}$ con grado di fiducia $(1-\alpha)\cdot 100\%$

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}} \qquad (7.28)$$

Il metodo descritto per trovare gli intervalli di confidenza per il rapporto di due varianze si applica solo a campioni estratti da popolazioni normali.
Anche in questo caso la verifica dell'ipotesi di normalità delle due popolazioni è di grande importanza.
I valori più comunemente usati per $1-\alpha$ sono 0.90, 0.95 e 0.99, a cui corrispondono i gradi di fiducia del 90%, del 95% e del 99%; i corrispondenti valori di $F_{\frac{\alpha}{2}}$ e di $F_{1-\frac{\alpha}{2}}$ sono

grado di fiducia del 90%	$F_{\frac{\alpha}{2}} = F_{0.05}$	$F_{1-\frac{\alpha}{2}} = F_{0.95}$
grado di fiducia del 95%	$F_{\frac{\alpha}{2}} = F_{0.025}$	$F_{1-\frac{\alpha}{2}} = F_{0.975}$
grado di fiducia del 99%	$F_{\frac{\alpha}{2}} = F_{0.005}$	$F_{1-\frac{\alpha}{2}} = F_{0.995}$

I valori $F_{\frac{\alpha}{2}}$ possono essere letti sulla tavola della distribuzione F in corrispondenza ai gradi di libertà $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$; i valori $F_{1-\frac{\alpha}{2}}$ si possono ricavare dalla stessa tavola facendo uso della formula seguente (formula (6.8), pag. 187).

$$F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2) = \frac{1}{F_{\frac{\alpha}{2}}(\nu_2,\nu_1)} \qquad (7.29)$$

Esempio 29
Si vuole studiare la variabilità dei diametri delle sfere d'acciaio prodotte da due diverse macchine. A tale scopo si estraggono due campioni di sfere prodotte dalle due macchine, di ampiezza rispettivamente $n_1 = 11$ e $n_2 = 16$; le varianze dei due campioni sono $s_1^2 = 0.40$ e $s_2^2 = 0.35$.

Assumendo che le due popolazioni da cui provengono i campioni abbiano distribuzione normale, trovare gli intervalli di confidenza al 90% e al 95% per il rapporto fra le varianze delle popolazioni.

$$\nu_1 = n_1 - 1 = 10 \qquad \nu_2 = n_2 - 1 = 15$$
$$s_1^2 = 0.40 \qquad s_2^2 = 0.35$$

a – Per il grado di fiducia del 90%, con le tavole e facendo uso della formula (7.29) si ha

$$1 - \alpha = 0.90 \qquad \frac{\alpha}{2} = 0.05$$
$$F_{\frac{\alpha}{2}}(10,15) = F_{0.05}(10,15) = 2.54$$
$$F_{1-\frac{\alpha}{2}}(10,15) = F_{0.95}(10,15) = \frac{1}{F_{0.05}(15,10)} = \frac{1}{2.85} = 0.35$$

Applicando la formula (7.28) si trova l'intervallo di confidenza con grado di fiducia del 90%

$$\frac{\frac{0.40}{0.35}}{2.54} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\frac{0.40}{0.35}}{0.35}$$
$$0.44 < \frac{\sigma_1^2}{\sigma_2^2} < 3.27$$

b – Per il grado di fiducia del 95% si ha invece

$$1 - \alpha = 0.95 \qquad \frac{\alpha}{2} = 0.025$$
$$F_{\frac{\alpha}{2}}(10,15) = F_{0.025}(10,15) = 3.06$$
$$F_{1-\frac{\alpha}{2}}(10,15) = F_{0.975}(10,15) = \frac{1}{F_{0.025}(15,10)} = \frac{1}{3.52} = 0.28$$

Applicando la formula (7.28) si trova l'intervallo di confidenza con grado di fiducia del 95%

$$\frac{\frac{0.40}{0.35}}{3.06} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\frac{0.40}{0.35}}{0.28}$$
$$0.37 < \frac{\sigma_1^2}{\sigma_2^2} < 4.09$$

Si osservi che aumentando il grado di fiducia, cresce l'ampiezza dell'intervallo, ossia la stima è meno precisa.

8. Test di ipotesi

8.1 Introduzione

Come è già stato messo in evidenza, uno degli scopi più importanti di un'analisi statistica è quello di utilizzare dei dati provenienti da un campione per fare inferenza sulla popolazione da cui è stato tratto il campione.
Nel Cap. 7 si è visto ad esempio come, utilizzando la media campionaria, si può stimare il valore del corrispondente parametro della popolazione.
Ci sono altri problemi in cui si sottopone a test un'ipotesi su un parametro di una popolazione, con lo scopo di decidere, esaminando un campione tratto dalla popolazione, se l'affermazione riguardante il parametro è vera o falsa.
Ad esempio il responsabile della produzione in un'azienda può ipotizzare che le confezioni prodotte abbiano un peso medio di 250g; un medico può ipotizzare che un certo farmaco sia efficace nel 90% dei casi in cui viene usato. Con la verifica delle ipotesi si può determinare se tali congetture sono compatibili con i dati disponibili dal campione.

Definizioni 1
Un'ipotesi formulata in termini di parametri di una popolazione, come media o varianza, è detta **ipotesi statistica**.
Il procedimento che consente di rifiutare o accettare un'ipotesi statistica utilizzando i dati di un campione, viene chiamato **test di ipotesi**.

8.2 Ipotesi statistiche

Per illustrare i concetti generali riguardanti la verifica delle ipotesi, consideriamo i seguenti esempi.

Esempio 1
Si vuole sottoporre a test l'affermazione di un produttore di vernici secondo cui il tempo medio di asciugatura di una nuova vernice è non superiore a $\mu = 20$ minuti.
A questo scopo si prende un campione di 35 lattine di vernice, si effettuano 35 prove di verniciatura con la vernice delle diverse confezioni e si calcola il tempo medio di asciugatura, con l'intenzione di rifiutare l'affermazione del produttore se la media osservata supera il valore di 20 minuti, o di accettarla in caso contrario.

Esempio 2
Si vuole verificare se le lattine di caffè confezionate automaticamente da una ditta contengono in media il peso dichiarato $\mu = 250$ g. A tale scopo si estrae un campione di 30 lattine, se ne pesa il contenuto e si calcola il peso medio, per stabilire se il peso medio differisce da 250g.

La verifica delle ipotesi statistiche inizia con la definizione del problema in termini di ipotesi sul parametro oggetto di studio.
Per prima cosa si stabilisce l'ipotesi da sottoporre a test, detta **ipotesi nulla**, indicata con $\mathbf{H_0}$.
Oltre all'ipotesi nulla occorre specificare anche un'adeguata **ipotesi alternativa**, indicata con $\mathbf{H_1}$, ossia un'affermazione che contraddice l'ipotesi nulla.
Nell'esempio 1 l'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu \leq 20$ minuti
H_1: $\mu > 20$ minuti.

Nell'esempio 2 l'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu = 250$ g
H_1: $\mu \neq 250$ g.

Originariamente il termine "nulla" nell'ipotesi nulla era usato con il significato di "nessuna differenza" o "la differenza è nulla", come illustrano i seguenti esempi.
Se si vuole stabilire se un metodo di insegnamento di una lingua straniera è più efficiente di un altro, si ipotizza che i due metodi siano ugualmente efficienti; se si vuole verificare se un farmaco è più efficace di un altro, si ipotizza che siano ugualmente efficaci. Questo in altre parole significa ipotizzare che non ci sia nessuna differenza fra i due metodi o fra i due farmaci: per questo motivo l'ipotesi si dice "nulla".
In generale attualmente il termine "ipotesi nulla" viene usato per ogni ipotesi sottoposta a test con lo scopo di decidere se deve essere rifiutata in favore dell'ipotesi alternativa.

Regole per la scelta delle ipotesi
Un problema importante nel predisporre un test di ipotesi è la scelta delle ipotesi: come si può decidere qual è l'ipotesi nulla e quale deve essere l'ipotesi alternativa?
Non c'è purtroppo una risposta semplice alla domanda, perché la scelta dipende anche da fattori soggettivi: chi effettua il test ha in genere convinzioni e idee personali su quanto intende mostrare. Tuttavia si possono indicare alcune linee guida; facciamo riferimento per comodità a un test sulla media di una singola popolazione, ma gli stessi principi si possono applicare a ogni test riguardante uno o più parametri.
Se ad esempio il test ha come scopo di decidere se il valore della media μ della popolazione è diverso dal valore 100 , l'ipotesi nulla e l'ipotesi alternativa saranno della forma

H_0: $\mu = 100$
H_1: $\mu \neq 100$

Se invece il test ha lo scopo di decidere se il valore di μ è minore di 100, allora l'ipotesi nulla e l'ipotesi alternativa saranno

H_0: $\mu \geq 100$
H_1: $\mu < 100$

Se infine il test ha lo scopo di decidere se il valore di μ è maggiore di 100, allora l'ipotesi nulla e l'ipotesi alternativa saranno

H_0: $\mu \leq 100$
H_1: $\mu > 100$

In conclusione:
1 – nell'ipotesi alternativa viene messo ciò che si spera o ci si aspetta di poter concludere come risultato del test;
2 – l'ipotesi nulla è posta con lo scopo di essere screditata, quindi ciò che si oppone alla conclusione che il ricercatore cerca di raggiungere rappresenta l'ipotesi nulla;
3 – nell'ipotesi nulla deve sempre comparire un segno di uguaglianza ($=$, $\leq$ o $\geq$);
4 – le due ipotesi sono complementari, ossia considerate insieme esauriscono tutte le possibilità riguardanti il valore che può assumere il parametro in esame.

Gli esempi seguenti illustrano la scelta dell'ipotesi nulla e dell'ipotesi alternativa in varie situazioni, nelle quali il parametro sottoposto a test è la media.

Esempio 3
Si supponga di voler stabilire se possiamo concludere che il tempo medio richiesto per svolgere una certa operazione è minore di 30 minuti. In tal caso si scelgono le ipotesi

H_0: $\mu \geq 30$ minuti
H_1: $\mu < 30$ minuti.

Esempio 4
Il contenuto dichiarato dal produttore delle bottiglie di acqua minerale di una certa marca è 920ml.
Un'associazione di consumatori sostiene che in realtà le bottiglie contengono in media una quantità inferiore di acqua. In questo caso le ipotesi sono

H_0: $\mu \geq 920$ ml
H_1: $\mu < 920$ ml.

Esempio 5
Un ingegnere suggerisce alcune modifiche che si potrebbero apportare a una linea produttiva per aumentare il numero di pezzi prodotti giornalmente.
Per decidere se applicare queste modifiche occorre che i dati sperimentali indichino con forte evidenza che la macchina modificata è più produttiva di quella originaria.
Se μ_0 è il numero medio di pezzi prodotti prima della modifica, si scelgono le ipotesi

H_0: $\mu \leq \mu_0$
H_1: $\mu > \mu_0$.

Osservazione
E' importante sottolineare che con la verifica delle ipotesi, e in generale con l'inferenza statistica, non si arriva alla dimostrazione di un'ipotesi; si ha solo un'indicazione del fatto che l'ipotesi sia o meno avvalorata dai dati disponibili: quando non si rifiuta un'ipotesi nulla, non si dice che essa è vera, ma che può essere vera; in altre parole se non rifiutiamo l'ipotesi nulla, possiamo solo concludere che il campione non fornisce prove sufficienti a garantirne il rifiuto, ma ciò non implica alcuna dimostrazione.
Riassumendo, le **possibili conclusioni per un test di ipotesi** sono:
1 – se l'ipotesi nulla H_0 è rifiutata, si conclude che l'ipotesi alternativa H_1 è probabilmente vera;
2 – se l'ipotesi nulla non è rifiutata si conclude che i dati non forniscono una sufficiente evidenza per sostenere l'ipotesi alternativa.

8.3 Tipi di errore e livello di significatività

Dopo aver formulato le ipotesi, occorre specificare quale risultato del campione porterà al rifiuto dell'ipotesi nulla.
Ricordiamo che le statistiche campionarie media e varianza sono stimatori corretti del corrispondente parametro della popolazione. Poiché il valore della statistica è calcolato da un campione, anche se l'ipotesi nulla è vera, è però molto probabile che la statistica differisca di una certa quantità dal valore del parametro della popolazione, per effetto del caso; ciò nonostante, se l'ipotesi nulla è vera, ci aspettiamo che la statistica campionaria sia vicina al parametro della popolazione.
Se ciò accade non ci sono prove sufficienti per rifiutare l'ipotesi nulla. Se nell'esempio 1, la media campionaria fosse ad esempio di 20.50 minuti, potremmo ragionevolmente concludere che l'ipotesi nulla è vera, ossia l'affermazione del produttore è vera, perché il valore campionario è "abbastanza vicino" al valore $\mu = 20$ minuti.
Analogamente, nel caso dell'esempio 2, se la media campionaria fosse di 245 g o di 255 g, potremmo ragionevolmente decidere di accettare l'ipotesi nulla che il peso medio sia $\mu = 250$ g, perché la differenza dal peso dichiarato è piccola; se invece la differenza dal peso medio fosse "troppo grande" potremmo decidere di rifiutare l'ipotesi. In ogni caso si prende una decisione basata sul fatto che si ritiene che i campioni estratti dalla popolazione siano rappresentativi della stessa.
Tuttavia il processo decisionale non può certo essere basato sui termini "abbastanza vicino" o "troppo grande" usati negli esempi; la teoria della verifica dei test di ipotesi si basa sullo studio della distribuzione campionaria di una statistica, detta statistica test.
La **statistica test** è una statistica che viene calcolata dai dati del campione e può assumere tanti valori quanti sono i possibili campioni estraibili dalla popolazione, quindi il particolare valore calcolato dipende dal campione estratto.
Un esempio di statistica test è la quantità

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

La distribuzione di campionamento della statistica test è, di solito, una distribuzione nota, come la distribuzione normale o la distribuzione *t*, e ricorriamo a queste distribuzioni per sottoporre a verifica un'ipotesi nulla; ad esempio la statistica test *Z* ha la distribuzione normale standardizzata.
Utilizzando le proprietà della distribuzione di campionamento della statistica soggetta a test, si può identificare un intervallo di valori di quella statistica che verosimilmente non si presentano se l'ipotesi nulla è vera.
La distribuzione di campionamento della statistica test è divisa in due regioni, una **regione di rifiuto** e una **regione di accettazione**, delimitate da uno o più valori, detti **valori critici**.

Definizioni 2
La **regione di rifiuto** corrisponde all'insieme dei valori di una statistica test che conducono al rifiuto dell'ipotesi nulla.
L'insieme dei valori che portano invece all'accettazione dell'ipotesi nulla si chiama **regione di accettazione**.
I **valori critici** sono i valori della statistica test che separano le regioni di rifiuto e di accettazione.

Se la statistica test, in base ai dati del campione, assume un valore che cade nella regione di rifiuto, l'ipotesi nulla deve essere rifiutata; se al contrario il valore cade nella regione di accettazione, l'ipotesi nulla non può essere rifiutata.
La regione di rifiuto può essere vista come l'insieme dei valori della statistica test che non è probabile che si verifichino quando l'ipotesi nulla è vera, mentre è probabile che si verifichino quando l'ipotesi nulla è falsa. Pertanto, se il campione porta a un valore della statistica test che cade nella regione di rifiuto, rifiutiamo l'ipotesi nulla perché non è probabile che sia vera.

I test di ipotesi possono essere classificati in due gruppi: **test a una coda** (o **test unilaterale**) e **test a due code** (o **test bilaterale**).
Quando la regione di rifiuto è costituita da un intervallo, il **test** si dice **a una coda**: questo caso è illustrato dalle figure 3 e 4, pag. 226-227; quando invece la regione di rifiuto è costituita da due intervalli, ossia da due code della distribuzione, il **test** si dice **a due code**: questo caso è illustrato dalla figura 5, pag. 227.
Un semplice modo per stabilire di che tipo è un test, senza dover conoscere la/le regioni di rifiuto è il seguente: per un test a due code nell'ipotesi alternativa compare il segno $\neq$, mentre per un test a una coda compare uno dei segni $>$ oppure $<$.

Quando si usa una statistica campionaria per prendere una decisione sul parametro della popolazione si corre sempre il rischio di giungere a una conclusione sbagliata. Questo dipende dal fatto che un'informazione parziale, ottenuta da un campione, è usata per trarre conclusioni sull'intera popolazione. Nella verifica di ipotesi si individuano due **tipi di errore**.
Per illustrare questo problema riprendiamo in esame l'esempio 1. Supponiamo di aver scelto la regione di accettazione, stabilendo di accettare l'ipotesi nulla se la media del campione non supera i 20.50 minuti.
C'è una prima possibilità che la media del campione superi i 20.50 minuti stabiliti, mentre la media effettiva della popolazione è $\mu = 20$ minuti; c'è anche una seconda possibilità che la media del campione possa essere minore o uguale a 20.50 minuti, ma la media effettiva non sia $\mu = 20$ minuti, ma sia ad esempio $\mu = 21$ minuti.
La situazione appena descritta in questo esempio è tipica dei test di ipotesi: anche se si fa il test in modo corretto, si possono commettere questi due tipi di errore, che possono portare a conseguenze dannose.

Definizioni 3
Se l'ipotesi H_0 è vera, ma viene erroneamente rifiutata, si commette un **errore del I tipo**; la probabilità di commettere tale errore è indicata con α.
Se l'ipotesi H_0 è falsa, ma erroneamente non viene rifiutata, si commette un **errore del II tipo**; la probabilità di commettere questo tipo di errore è indicata con β.

I risultati delle decisioni a cui si perviene con un test di ipotesi possono essere riassunti nel seguente schema. A seconda della decisione presa, si può verificare uno dei due tipi di errore[1]

	H_0 vera	H_0 falsa
Rifiutiamo H_0	Errore del I tipo Probabilità (errore I tipo) = α	Decisione corretta
Accettiamo H_0	Decisione corretta	Errore del II tipo Probabilità (errore II tipo) = β

Un'analogia che può chiarire le idee precedenti è quella del processo a un imputato. In tribunale una persona sottoposta a processo viene ritenuta innocente fino a prova contraria. L'ipotesi nulla H_0 è quindi "l'imputato è innocente"; l'ipotesi alternativa H_1 è "l'imputato è colpevole".
L'errore del I tipo è condannare un innocente, l'errore del II tipo è assolvere un colpevole.
Riassumiamo questi concetti con lo schema seguente.

	Imputato innocente	Imputato colpevole
Imputato condannato	Errore del I tipo	Decisione corretta
Imputato assolto	Decisione corretta	Errore del II tipo

Scegliere come ipotesi nulla H_0 "l'imputato è innocente" significa ritenere che condannare un innocente sia un errore più grave che assolvere un colpevole.
In generale l'errore di I tipo è quello considerato più grave: questo significa che l'ipotesi nulla H_0 va formulata in modo che quello che si ritiene sia l'errore più grave coincida con l'errore di I tipo.
Servendoci ancora degli esempi 1 e 2, calcoliamo la probabilità α di commettere un errore del I tipo; usiamo a tale scopo le proprietà della distribuzione della media campionaria.

Esempio 1 – parte 2
Assumiamo che sia noto dall'esperienza che lo scarto quadratico medio del tempo di asciugatura della vernice è $\sigma = 2$ minuti e studiamo la probabilità di commettere un errore del I tipo, ossia la probabilità α che la media del campione superi 20.5 minuti, anche se la media effettiva della popolazione è $\mu \le 20$ minuti[2].
Come è noto dal Cap. 6, la distribuzione della media campionaria per grandi campioni ($n \ge 30$) è approssimativamente normale, quindi la probabilità suddetta è data dall'area della regione rappresentata nella figura 1.

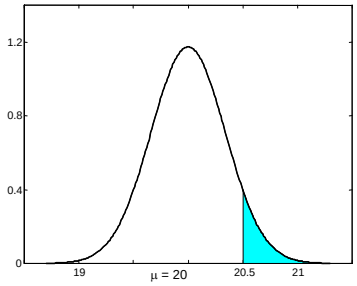


Figura 1

La regione a destra del valore 20.5 è la regione di rifiuto, quella a sinistra è la regione di accettazione: se il valore della media campionaria cade a destra di 20.5 l'ipotesi nulla viene rifiutata, altrimenti non viene rifiutata.

[1] Per ricordare le probabilità associate ai due tipi di errore, si osservi che α è la prima lettera dell'alfabeto greco e si usa per indicare la probabilità dell'errore di I tipo, β è la seconda lettera e si usa per indicare l'errore di II tipo.
[2] Si ricordi che l'ipotesi nulla e l'ipotesi alternativa in questo esempio sono
H_0: $\mu \le 20$ minuti
H_1: $\mu > 20$ minuti.

Se la popolazione da cui proviene il campione è sufficientemente grande da poterla considerare infinita[3], applicando il teorema 1, Cap. 6, pag. 177, si calcola la deviazione standard della distribuzione della media campionaria

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{35}} = 0.34$$

Standardizzando il valore $\bar{x} = 20.5$ si ha

$$Z = \frac{20.5 - 20}{0.34} = 1.47 .$$

Utilizzando le tavole della distribuzione normale, si trova che l'area della regione a destra di 20.5 è

$$P(Z > 1.47) = 1 - P(Z < 1.47) = 1 - 0.9292 = 0.0708$$

quindi la probabilità di rifiutare erroneamente l'ipotesi nulla è

$$\alpha = 0.0708$$

Esempio 2 – parte 2

Assumiamo che lo scarto quadratico medio della popolazione sia σ = 15g e studiamo la probabilità α che la media del campione non sia compresa fra 245g e 255g, anche se la media effettiva della popolazione è μ = 250g [4].

La probabilità che si vuole calcolare è data dalla somma delle due aree rappresentate nella figura 2

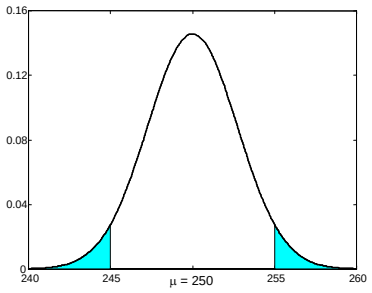


Figura 2

La regione di rifiuto in questo caso è costituita dai valori a sinistra di 245g e dai valori a destra di 255g; se il valore della media campionaria cade nell'intervallo (245, 255), che è la regione di accettazione, l'ipotesi nulla viene accettata, altrimenti viene rifiutata.

Seguendo il procedimento già descritto nell'esempio precedente si trova

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{30}} \cong 2.74 .$$

La regione di accettazione è un intervallo simmetrico rispetto a μ = 250; standardizzando il valore $\bar{x} = 255$ si ha

$$Z = \frac{255 - 250}{2.74} = 1.82 .$$

Utilizzando le tavole della distribuzione normale, si trova che l'area della regione colorata è

$$P(|Z| > 1.82) = 2 \cdot [1 - P(Z < 1.82)] = 2 \cdot (1 - 0.9656) = 0.0688$$

quindi la probabilità di rifiutare erroneamente l'ipotesi nulla è α = 0.0688 .

[3] Si ricordi quanto detto nel Cap. 6, pag. 177, a proposito della correzione per popolazioni finite e si veda lo schema riassuntivo a pag. 178, punto 2 b.

[4] Si ricordi che l'ipotesi nulla e l'ipotesi alternativa in questo esempio sono

H_0: μ = 250 g

H_1: $\mu \neq$ 250 g.

Definizione 4
La probabilità α di commettere un errore del I tipo, ossia di rifiutare un'ipotesi nulla vera, è detta **livello di significatività**.

Negli esempi 1 e 2 (parte 2) si è mostrato come calcolare la probabilità α di commettere un errore del I tipo, per regioni di rifiuto scelte arbitrariamente; questo non è però il procedimento seguito di solito nelle applicazioni.
Il metodo usato più frequentemente consiste invece nello specificare un valore per il livello di significatività α e poi identificare la regione di rifiuto che soddisfa tale valore.
Poiché l'errore di I tipo è quello considerato più grave, si scelgono per α valori piccoli; i valori più usati sono $\alpha = 0.01$ e $\alpha = 0.05$.
In corrispondenza al livello di significatività α, il valore $(1-\alpha)\cdot 100\%$ coincide con il **grado di fiducia** già introdotto per gli intervalli di confidenza.
Se si sceglie ad esempio un livello di significatività $\alpha = 0.05$, ossia del 5%, ci sarà una probabilità del 5% di rifiutare un'ipotesi che non avrebbe dovuto essere rifiutata; in altre parole siamo fiduciosi al 95% di aver preso la decisione giusta.

Definizione 5
La probabilità di commettere un errore del II tipo, indicata con β, viene anche chiamata **rischio del consumatore**.

Si può controllare il rischio connesso a un errore del I tipo scegliendo un valore di α piccolo, ad esempio $\alpha = 0.01$: questo deve essere fatto se si ritiene che le conseguenze di un errore del I tipo siano gravi. Tuttavia, per una fissata ampiezza del campione, al diminuire di α, aumenta β, ossia ad una riduzione dell'errore del I tipo si accompagna un aumento dell'errore del II tipo. Quindi nei casi in cui è molto importante evitare, per quanto possibile, un errore del II tipo troppo grande, è meglio scegliere come valore di α un valore non troppo piccolo, ad esempio $\alpha = 0.05$.
Un modo per controllare e ridurre l'errore del II tipo consiste nell'aumentare la dimensione del campione. Un'elevata dimensione del campione consente di solito di individuare anche piccole differenze tra la statistica campionaria e il parametro della popolazione. Si tenga presente però che aumentare di molto l'ampiezza del campione potrebbe essere troppo costoso.
Per un fissato valore di α l'aumento dell'ampiezza del campione riduce il rischio del consumatore β, quindi aumenta la probabilità $1 - \beta$ di rifiutare l'ipotesi nulla quando è falsa, e dovrebbe essere rifiutata. La probabilità $1 - \beta$ si chiama anche **potenza del test**.
La scelta dei valori di α e β dipende dai costi che ciascun errore comporta (vedere esempio 8).
Nell'esempio 2, relativo alla produzione delle confezioni di caffè, si commette un errore del I tipo quando si conclude che il peso medio del caffè contenuto nelle confezioni prodotte non è uguale a 250g quando invece lo è; si commette un errore del II tipo quando si conclude che il peso medio del caffè è uguale a 250g quando non lo è. La scelta dei valori di α e β dipende dai costi che ciascun errore comporta. Se un cambiamento del processo produttivo fosse molto costoso, si dovrebbe essere certi della sua necessità. Il rischio comportato da un errore di I tipo in questo caso è il più grave e si dovrebbe cercare di contenerlo. Se si vuole invece essere sicuri di cogliere una differenza anche piccola dalla media di 250g, si deve considerare come molto grave il rischio associato a un errore del II tipo e cercare di limitarlo scegliendo un valore più grande per α.

Riassumiamo nello schema seguente i passi in cui si articola un test di ipotesi.

Schema riassuntivo – Test di ipotesi
1 – Si scelgono l'ipotesi nulla e l'ipotesi alternativa.
2 – Si sceglie il livello di significatività α a cui si vuole eseguire il test.
3 – In funzione del tipo di test (una coda o due code) e del valore di α scelto, si determinano i valori critici e la regione di rifiuto.

4 – Si sceglie l'ampiezza campionaria, si raccoglie il campione, si calcola dai dati del campione il valore della statistica test e si vede se appartiene o no alla regione di rifiuto.
5 – Si prende la decisione: rifiutare o non rifiutare l'ipotesi nulla al livello di significatività stabilito.

E' opportuno sottolineare che, quando l'ipotesi nulla non è rifiutata, non si dovrebbe dire che tale ipotesi viene accettata, bensì che l'ipotesi nulla non viene rifiutata: questo perché è possibile che si commetta un errore del II tipo; poiché spesso la probabilità di commettere un errore del II tipo è abbastanza elevata, non ci si dovrebbe impegnare troppo dicendo che si accetta l'ipotesi nulla.
Tuttavia, anche se impropriamente, spesso si usa il termine "si accetta l'ipotesi nulla".

8.4 Test di ipotesi sulla media (varianza nota)

Descriviamo il procedimento per eseguire un test di ipotesi sulla media di una popolazione avente varianza σ^2 nota.
Il test si basa sulla statistica

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

dove n è l'ampiezza del campione e μ_0 è il valore della media assunto nell'ipotesi nulla
Il test qui illustrato è essenzialmente un **test per grandi campioni** ($n \geq 30$); in tal caso la distribuzione della media campionaria può essere approssimata dalla distribuzione normale e la variabile aleatoria Z ha approssimativamente la distribuzione normale standardizzata.
Nel caso particolare in cui il campione è estratto da una popolazione con distribuzione normale, la variabile Z ha distribuzione normale standardizzata, qualunque sia l'ampiezza del campione (vedere esempi 13 e 14).
Sia, come al solito, z_α il valore di Z per cui l'area a destra di z_α al di sotto della curva normale standardizzata è uguale a α.
Nelle figure seguenti si illustrano le regioni di rifiuto per un dato livello di significatività α, a seconda delle ipotesi nulla e alternativa stabilite.
Nei primi due casi si fa un test a una coda, nel terzo caso un test a due code.

1° caso – Test a una coda (figura 3)

Ipotesi nulla	H_0:	$\mu \leq \mu_0$.
Ipotesi alternativa	H_1:	$\mu > \mu_0$.
Regione di rifiuto[5]	$Z > z_\alpha$	
Regione di accettazione	$Z < z_\alpha$	

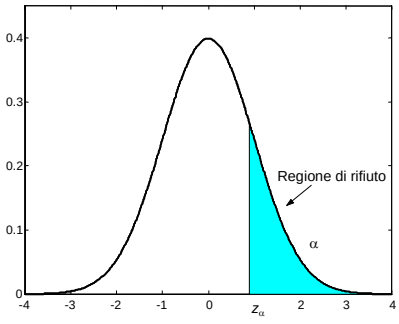


Figura 3

[5] L'utilizzo o meno del segno di uguale nelle regioni di rifiuto e di accettazione, in questo e nei casi seguenti, è assolutamente ininfluente, dal momento che la distribuzione normale è una distribuzione continua.

2° caso – Test a una coda (figura 4)

Ipotesi nulla H_0: $\mu \geq \mu_0$.

Ipotesi alternativa H_1: $\mu < \mu_0$.

Regione di rifiuto $Z < -z_\alpha$

Regione di accettazione $Z > -z_\alpha$

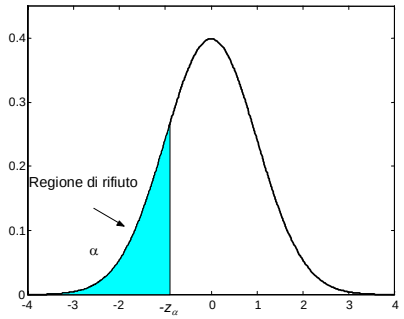


Figura 4

3° caso – Test a due code (figura 5)

Ipotesi nulla H_0: $\mu = \mu_0$.

Ipotesi alternativa H_1: $\mu \neq \mu_0$.

Regione di rifiuto $Z < -z_{\frac{\alpha}{2}}$ oppure $Z > z_{\frac{\alpha}{2}}$

Regione di accettazione $-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}$

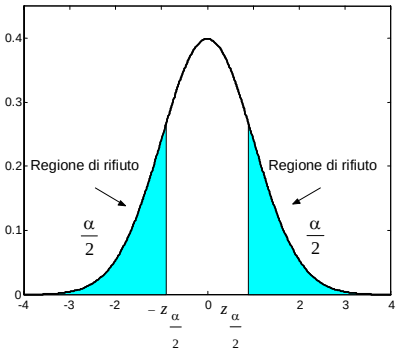


Figura 5

I valori z_α e $z_{\frac{\alpha}{2}}$ sono i valori critici del test nei tre casi; tali valori possono essere letti sulla tavola dei percentili della distribuzione normale standardizzata.

Nella tabella 1 riassumiamo i valori comunemente usati per il livello di significatività α e i corrispondenti valori critici z_α e $z_{\frac{\alpha}{2}}$ per i test a una e a due code.

Test	Ipot. nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu \leq \mu_0$	$\mu > \mu_0$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$\mu \geq \mu_0$	$\mu < \mu_0$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$\mu = \mu_0$	$\mu \neq \mu_0$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Tabella 1

Esempio 6

Una ditta produttrice di lampadine sostiene che la durata media delle lampadine prodotte è di 1600 ore, con uno scarto quadratico medio $\sigma = 120$ ore.
Estraendo un campione di 100 lampadine si è calcolata una durata media di 1570 ore.
Stabilire se l'affermazione del produttore è corretta, usando come ipotesi alternativa che la durata media sia
a – inferiore a quella dichiarata;
b – diversa da quella dichiarata.
Usare in entrambi i casi il livello di significatività $\alpha = 0.05$ e il livello di significatività $\alpha = 0.01$.

a –
Ipotesi nulla H_0: $\mu \geq 1600$
Ipotesi alternativa H_1: $\mu < 1600$
Livello di significatività $\alpha = 0.05$
Il test è a una coda; il valore critico per questo livello di significatività è $z_\alpha = -1.645$.
La regola di decisione consiste nel rifiutare l'ipotesi se il valore della statistica *Z* ottenuto dai dati del campione è minore di -1.645.
Il campione ha le seguenti caratteristiche

$$n = 100 \qquad \bar{x} = 1570$$

Il valore della statistica test è

$$Z = \frac{1570 - 1600}{\frac{120}{\sqrt{100}}} = -2.50 .$$

Dato che il valore trovato $Z = -2.50$ è minore del valore critico $z_\alpha = -1.645$, si rifiuta l'ipotesi nulla al livello di significatività $\alpha = 0.05$, ossia del 5%.
Livello di significatività $\alpha = 0.01$
Il test è a una coda; il valore critico per questo livello di significatività è $z_\alpha = -2.326$.
Anche in questo caso il valore $Z = -2.50$ è minore del valore critico $z_\alpha = -2.326$, perciò si rifiuta l'ipotesi nulla al livello di significatività $\alpha = 0.01$, ossia dell'1%.
b –
Ipotesi nulla H_0: $\mu = 1600$
Ipotesi alternativa H_1: $\mu \neq 1600$
Livello di significatività $\alpha = 0.05$
Il test è a due code; i valori critici per questo livello di significatività sono

$$z_{\frac{\alpha}{2}} = -1.96 \text{ e } z_{\frac{\alpha}{2}} = 1.96 .$$

Il valore $Z = -2.50$ cade al di fuori dell'intervallo avente come estremi i valori critici, cioè appartiene alla regione di rifiuto, perciò si rifiuta l'ipotesi nulla al livello di significatività $\alpha = 0.05$, ossia del 5%.
Livello di significatività $\alpha = 0.01$.
I valori critici per questo livello di significatività sono $z_{\frac{\alpha}{2}} = -2.576$ e $z_{\frac{\alpha}{2}} = 2.576$.

Il valore $Z = -2.50$ cade fra questi estremi, perciò non si rifiuta l'ipotesi nulla al livello di significatività $\alpha = 0.01$, ossia dell'1%.

Esempio 7

La lunghezza della corda contenuta nei rotoli prodotti da una macchina ha una distribuzione avente varianza $\sigma^2 = 27.4\,\text{m}^2$. La ditta produttrice afferma che la lunghezza media è $\mu = 300$m.
Viene prelevato un campione di 100 rotoli e calcolata la lunghezza media, pari a $\bar{x} = 299.2$.

Verificare se il produttore afferma il vero, oppure se la lunghezza è inferiore, al livello di significatività dell'1%.

Ipotesi nulla H_0: $\mu \geq 300$
Ipotesi alternativa H_1: $\mu < 300$
Livello di significatività $\alpha = 0.01$.
Il test è a una coda; il valore critico per questo livello di significatività è $z_\alpha = -2.326$. La regione di rifiuto è $z < -2.326$.
Si ha

$$n = 100 \qquad \bar{x} = 299.2$$

$$\sigma^2 = 27.4$$

Il valore della statistica test è

$$Z = \frac{299.2 - 300}{\frac{\sqrt{27.4}}{\sqrt{100}}} = -1.53 .$$

Il valore $Z = -1.53$ non appartiene alla regione di rifiuto, quindi l'ipotesi nulla non viene rifiutata al livello di significatività dell'1%.

Esempio 8
La precisione di una macchina che produce componenti di dimensioni specificate viene controllata con periodiche verifiche a campione: la dimensione media richiesta è μ = 3.5mm, con una varianza $\sigma^2 = 0.22\ \text{mm}^2$.
a – Valutare se il processo è da ritenersi sotto controllo oppure no, quando la media riscontrata su un campione di 60 pezzi è $\bar{x} = 3.42$ mm.
b – Ripetere la valutazione nel caso che il campione sia di 150 pezzi, con media $\bar{x} = 3.41$ mm.

Si sceglie come ipotesi nulla di ritenere che il processo sia sotto controllo e non sia quindi necessario alcun intervento

H_0: $\mu = 3.5$

L'ipotesi alternativa è che il processo sia fuori controllo

H_1: $\mu \neq 3.5$

e in questo caso occorre attuare qualche intervento per riportarlo sotto controllo. Si effettua quindi un test a due code.
Se il processo è sotto controllo, cioè H_0 è vera, ma erroneamente lo riteniamo fuori controllo, cioè rifiutiamo H_0, commettiamo un errore del I tipo; la probabilità di compiere tale errore è pari al livello di significatività α.
L'errore del II tipo consiste invece nel concludere che il processo è sotto controllo, cioè H_0 è vera, quando non lo è; la probabilità di commettere questo errore è indicata con β;
La scelta dei valori di α e β dipende dai costi che ciascun errore comporta.
Se un cambiamento del processo produttivo è molto costoso, si dovrebbe essere ben sicuri della sua necessità, quindi si deve scegliere un valore di α piccolo.
Se invece ci poniamo dal punto di vista del consumatore e vogliamo essere sicuri di cogliere uno spostamento anche piccolo dalla media ipotizzata, allora il rischio β del consumatore deve essere basso e dobbiamo scegliere un valore più elevato di α.
a – $n = 60$ $\bar{x} = 3.42$ $\sigma^2 = 0.22$
Il valore della statistica test è

$$Z = \frac{3.42 - 3.5}{\frac{\sqrt{0.22}}{\sqrt{60}}} = -1.32 .$$

Livello di significatività $\alpha = 0.05$
Regione di rifiuto $Z < -1.96$ e $Z > 1.96$
Il valore $Z = -1.32$ non appartiene alla regione di rifiuto, quindi l'ipotesi nulla non viene rifiutata al livello di significatività del 5%; il processo si ritiene sotto controllo.

Livello di significatività $\alpha = 0.01$
Regione di rifiuto $Z < -2.576$ e $Z > 2.576$
Il valore $Z = -1.32$ non appartiene alla regione di rifiuto, quindi l'ipotesi nulla non viene rifiutata al livello di significatività dell'1%; anche in questo caso il processo si ritiene sotto controllo.

b – $n = 150$ $\bar{x} = 3.42$ $\sigma^2 = 0.2209$
Il valore della statistica test è

$$Z = \frac{3.41 - 3.5}{\frac{\sqrt{0.22}}{\sqrt{150}}} = -2.35 .$$

Livello di significatività $\alpha = 0.05$
Regione di rifiuto $Z < -1.96$ e $Z > 1.96$
Il valore $Z = -2.35$ appartiene alla regione di rifiuto, quindi l'ipotesi nulla viene rifiutata al livello di significatività del 5%; il processo si ritiene fuori controllo e si devono intraprendere delle modifiche al processo produttivo.

Livello di significatività $\alpha = 0.01$
Regione di rifiuto $Z < -2.576$ e $Z > 2.576$
Il valore $Z = -2.35$ appartiene alla regione di accettazione, quindi l'ipotesi nulla viene accettata al livello di significatività dell'1%; il processo si ritiene sotto controllo e non si intraprendono modifiche al processo produttivo.
Il rischio più basso per il consumatore si ha nel caso in cui $n = 150$ e $\alpha = 0.05$. Il punto di vista del produttore è ovviamente diverso.

Esempio 9
I carichi di rottura dei cavi prodotti da un'azienda hanno una media pari a 1800kg e uno scarto quadratico medio σ =100kg. Si afferma che mediante una nuova tecnica di costruzione il carico di rottura può essere reso maggiore. Per sottoporre a test questa affermazione si provano 50 cavi e si trova che il carico di rottura medio è di 1850kg.
E' possibile accettare l'affermazione ad un livello di significatività dell'1%?

Si assume come ipotesi nulla che non ci sia nessun cambiamento

H_0: $\mu \leq 1800$

e come ipotesi alternativa che ci sia un aumento nel carico di rottura, ossia

H_1: $\mu > 1800$.

Si effettua un test ad una coda; per il livello di significatività $\alpha = 0.01$ il valore critico è $z_\alpha = 2.326$ e la regione di rifiuto è costituita dai valori $Z > 2.326$.
Il valore della statistica test è

$$Z = \frac{1850 - 1800}{\frac{100}{\sqrt{50}}} = 3.54 .$$

Dato che il valore trovato $Z = 3.54$ è maggiore del valore critico $z_\alpha = 2.326$, appartiene alla regione di rifiuto, perciò l'ipotesi nulla può essere rifiutata al livello di significatività $\alpha = 0.01$.

Esempio 10

Un campione di 36 osservazioni avente media $\bar{x} = 86.2$ proviene da una distribuzione avente varianza $\sigma^2 = 100$. In passato la media della distribuzione era $\mu = 83$, ma si ipotizza che recentemente la media possa essere cambiata.
Usando il livello di significatività del 5%, sottoporre a test l'ipotesi nulla opportuna nei seguenti casi:
a – supporre di non sapere, nel caso che la media sia cambiata, se è aumentata o diminuita;
b – supporre di sapere che, nel caso che la media sia cambiata, essa può solo essere aumentata.

a – Nel primo caso l'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu = 83$
H_1: $\mu \neq 83$

Si effettua un test a due code; per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è costituita dai valori $Z < -1.96$ e $Z > 1.96$.
Il valore della statistica test è

$$Z = \frac{86.2 - 83}{\frac{10}{\sqrt{36}}} = 1.92$$

Il valore $Z = 1.92$ non appartiene alla regione di rifiuto, perciò si decide di non rifiutare l'ipotesi nulla. In altre parole non c'è un'evidenza significativa, al livello del 5%, che la media sia cambiata.

b – Nel secondo caso l'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu \leq 83$
H_1: $\mu > 83$

Si effettua un test a una coda; per il livello di significatività $\alpha = 0.05$ il valore critico é $z_\alpha = 1.645$; la regione di rifiuto è costituita dai valori $Z > 1.645$.
Il valore $Z = 1.92$ appartiene alla regione di rifiuto, perciò si decide di rifiutare l'ipotesi nulla. In altre parole si ha un'evidenza significativa, al livello del 5%, che la media è aumentata.
Si noti che le decisioni prese sono diverse nei due casi, e ciò dipende dal fatto che l'ipotesi nulla viene testata contro alternative diverse.

Il test descritto in questo paragrafo richiede che sia noto il valore σ dello scarto quadratico medio; se σ non è conosciuto, ma il campione è grande, si può sostituire σ con il valore s dello scarto quadratico medio del campione, commettendo un errore di approssimazione.

Esempio 11

Una ditta produttrice di pneumatici afferma che la durata media di un certo tipo di pneumatici per auto è di almeno 50000km.
Per sottoporre a test questa affermazione un campione di 40 pneumatici viene sottoposto a prove su strada e si misura una durata media $\bar{x} = 48900$ km, con uno scarto quadratico medio $s = 2500$km.
Sottoporre a test l'affermazione, con un livello di significatività $\alpha = 0.01$.

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu \geq 50000$
H_1: $\mu < 50000$.

Si effettua un test ad una coda; per il livello di significatività $\alpha = 0.01$ il valore critico è $z_\alpha = -2.326$ e la regione di rifiuto è costituita dai valori $Z < -2.326$.
Lo scarto quadratico medio della popolazione non è noto e viene sostituito con lo scarto quadratico medio del campione.

Il valore della statistica test è

$$Z = \frac{48900 - 50000}{\frac{2500}{\sqrt{40}}} = -2.78 .$$

Il valore $Z = -2.78$ appartiene alla regione di rifiuto, perciò l'ipotesi nulla deve essere rifiutata al livello di significatività $\alpha = 0.01$, e l'affermazione del produttore non può essere accettata.

Esempio 12

In un dato anno il voto medio all'esame di maturità classica è stato di 73/100.
In una commissione che ha esaminato 70 candidati, si è registrato un voto medio di 76.2/100 con uno scarto quadratico medio $s = 14$.
Verificare l'ipotesi che non ci sia differenza significativa tra la media generale e la media del campione, al livello di significatività del 5%.

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu = 73$
H_1: $\mu \neq 73$.

Il test è a due code e al livello di significatività del 5% la regione di rifiuto è costituita dai valori $Z < -1.96$ e $Z > 1.96$.
Lo scarto quadratico medio della popolazione non è noto e viene sostituito con lo scarto quadratico medio del campione $s = 14$.
Il valore della statistica test è

$$Z = \frac{76.2 - 73}{\frac{14}{\sqrt{70}}} = 1.91 .$$

Il valore $Z = 1.91$ non appartiene alla regione di rifiuto, perciò al livello di significatività del 5% l'ipotesi nulla non deve essere rifiutata; concludiamo quindi che la differenza tra il risultato generale e il risultato della particolare commissione è dovuta a fluttuazioni casuali, ossia con una probabilità del 95% la differenza non è imputabile né ai candidati, né alla commissione d'esame.
E' evidente che nella scelta delle ipotesi non si vuole indagare su una maggiore o minore severità della commissione.

Se la popolazione da cui proviene il campione è normale, questo test può essere applicato anche nel caso di piccoli campioni con varianza σ^2 nota.

Esempio 13

Supponiamo che i punteggi di un test sul quoziente di intelligenza di una certa popolazione di adulti si distribuiscano normalmente con uno scarto quadratico medio $\sigma = 15$.
Un campione di 25 adulti estratti da questa popolazione ha un punteggio medio di 105.
Sottoporre a test l'ipotesi che il punteggio medio sia 100, con un livello di significatività del 5%.

Poiché la popolazione da cui proviene il campione ha distribuzione normale con scarto quadratico medio noto $\sigma = 15$, quanto detto per i grandi campioni è valido anche per un piccolo campione.
L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu = 100$
H_1: $\mu \neq 100$.

Si effettua un test a due code; per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è costituita dai valori $Z < -1.96$ e $Z > 1.96$. Il valore della statistica test è

$$Z = \frac{105 - 100}{\frac{15}{\sqrt{25}}} = 1.67 .$$

Il valore $Z = 1.67$ non appartiene alla regione di rifiuto, perciò non si rifiuta l'ipotesi nulla.

Esempio 14

Da una popolazione normale avente scarto quadratico medio $\sigma = 2$, si estrae un campione di ampiezza $n = 10$. Il valor medio del campione sia $\bar{x} = 18.58$.

Sottoporre a test l'ipotesi nulla

$$H_0: \quad \mu = 20$$

scegliendo come ipotesi alternativa

$$H_1: \quad \mu \neq 20$$

ai livelli di significatività dell'1% e del 5%.

Poiché la popolazione da cui proviene il campione è normale, si può effettuare il test per grandi campioni anche se l'ampiezza del campione è $n = 10$.

a – Livello di significatività $\alpha = 0.01$.

Si effettua un test a due code; per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è costituita dai valori $Z < -2.576$ e $Z > 2.576$.

Il valore della statistica test è

$$Z = \frac{18.58 - 20}{\frac{2}{\sqrt{10}}} = -2.245 .$$

Il valore $Z = -2.245$ non appartiene alla regione di rifiuto, perciò si decide di non rifiutare l'ipotesi nulla.

b – Livello di significatività $\alpha = 0.05$.

Si effettua un test a due code; per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è costituita dai valori $Z < -1.96$ e $Z > 1.96$.

Il valore $Z = -2.245$ appartiene alla regione di rifiuto, perciò si decide di rifiutare l'ipotesi nulla.

Nel caso trattato in questo esempio si possono dunque trarre le seguenti conclusioni:

a – I dati campionari non consentono di rifiutare l'ipotesi nulla al livello di significatività dell'1%.

b – I dati campionari consentono di rifiutare l'ipotesi nulla al livello di significatività del 5%.

Come si vede, la decisione che si prende non dipende solo dai dati campionari, ma anche dal livello di significatività fissato. In questo caso, la differenza fra la media del campione $\bar{x} = 18.58$ e il valore ipotizzato $\mu = 20$ per il parametro della popolazione viene ritenuta statisticamente significativa al livello del 5%, ma non al livello dell'1%.

Queste conclusioni ci portano alle seguenti considerazioni.

Ogni test di ipotesi porta al confronto di due numeri, il valore della statistica Z, che può essere calcolato in base ai dati campionari, e il valore critico (o i due valori critici nel test a due code), che invece dipende dal livello di significatività fissato. Nell'esempio sono stati confrontati il valore della statistica $Z = -2.245$ e i valori critici $z_{\frac{\alpha}{2}} = -1.96$ e $z_{\frac{\alpha}{2}} = 1.96$. Se tra i valori suddetti vale una certa disuguaglianza, si rifiuta l'ipotesi, altrimenti non si rifiuta.

Poiché, come abbiamo visto nell'esempio precedente, un livello α diverso può condurre a una decisione diversa (rifiutare/non rifiutare) risulta interessante determinare qual è il valore α che fa da spartiacque fra le due diverse conclusioni.

Nell'esempio precedente ci poniamo la seguente domanda: fissati i dati del campione, e quindi il valore di Z, qual è il più piccolo livello di significatività α per cui si rifiuta l'ipotesi nulla?

Nel caso dell'esempio la regione di rifiuto è costituita dai valori Z tali che

$$|Z| > z_{\frac{\alpha}{2}}$$

Il più piccolo valore di α per cui si rifiuta l'ipotesi si trova risolvendo l'equazione

$$2.245 = z_{\frac{\alpha}{2}}$$

Da qui segue[6]

$$\frac{\alpha}{2} = P(Z > 2.245) = 1 - P(Z < 2.245) = 1 - 0.9877 = 0.0123$$

$$\alpha = 0.0246$$

Questo significa che, con i dati campionari disponibili, il livello di significatività che fa da spartiacque tra la decisione di rifiutare l'ipotesi nulla e quella di non rifiutarla è il livello del 2.46%: questo livello è quindi il più piccolo livello a cui i dati disponibili permettono di rifiutare l'ipotesi nulla.

Definizione 6
In un test di ipotesi, dopo aver effettuato il campionamento e aver calcolato il valore della statistica test necessario per eseguire il test, si dice ***p*-value** il più piccolo valore del livello di significatività α per cui i dati campionari consentono di rifiutare l'ipotesi nulla.

Un *p*-value quasi uguale a zero significa che siamo praticamente certi di non sbagliare rifiutando l'ipotesi nulla; un *p*-value dell'ordine dei soliti livelli di significatività indica che la decisione se rifiutare o no l'ipotesi nulla è critica e dipende in modo cruciale dalla scelta del livello di significatività; un *p*-value maggiore indica invece che, a qualsiasi livello ragionevole di significatività, sbagliamo rifiutando l'ipotesi nulla; in questo caso si può anche dire che il test ci porta ad accettare l'ipotesi.
Una regola generale utile da ricordare è la seguente: se il *p*-value è minore o uguale ad α, rifiutiamo l'ipotesi nulla; se il *p*-value è maggiore di α, non rifiutiamo l'ipotesi nulla.
Riportare il *p*-value al termine di un test dà maggiori informazioni rispetto al riportare solo la decisione presa al livello di significatività scelto.
Il *p*-value può essere difficile da calcolare con precisione usando le tavole, ma viene di solito fornito dai più diffusi software statistici.
Per i test basati sulla distribuzione normale, come nel caso dell'esempio 12, il *p*-value è relativamente facile da calcolare. Se Z_0 è il valore della statistica test, calcolato in base ai dati campionari, allora il *p*-value può essere ottenuto in base alle seguenti formule

$$p-\text{value} = \begin{cases} 1 - P(Z < Z_0) & \text{per il test a una coda con} \quad H_0 : \mu \le \mu_0 \quad H_1 : \mu > \mu_0 \\ P(Z < Z_0) & \text{per il test a una coda con} \quad H_0 : \mu \ge \mu_0 \quad H_1 : \mu < \mu_0 \\ 2\left[1 - P(Z < |Z_0|)\right] & \text{per il test a due code con} \quad H_0 : \mu = \mu_0 \quad H_1 : \mu \ne \mu_0 \end{cases}$$

Esempio 15
Riprendiamo in esame i risultati ottenuti nell'esempio 8.
Le conclusioni tratte nel caso b sono piuttosto critiche e questo viene evidenziato dal *p*-value; si ha infatti

$$Z_0 = -2.35$$

$$p-\text{value} = 2\left[1 - P(Z < 2.35)\right] = 2(1 - 0.9906) = 0.0188$$

Il livello minimo che consente di rifiutare l'ipotesi nulla è del 1.88%.
Nel caso a invece le conclusioni non sono critiche; si ha infatti

$$Z_0 = -1.32$$

$$p-\text{value} = 2\left[1 - P(Z < 1.32)\right] = 2(1 - 0.9066) = 0.1868$$

In questo caso a ogni ragionevole livello di significatività possiamo accettare l'ipotesi nulla.

[6] Il valore della probabilità $P(Z < 2.245)$ è stato ottenuto come media dei due valori adiacenti disponibili sulle tavole

$$\frac{0.9875 + 0.9878}{2} = 0.98765 \cong 0.9877$$

8.5 Test di ipotesi sulla media (varianza incognita)

Esaminiamo ora il caso in cui il campione usato per effettuare il test proviene da una popolazione di cui non è nota la varianza σ^2.
Come già osservato nel paragrafo precedente, se σ non è noto, ma il campione è grande, si può sostituire σ con il valore *s* dello scarto quadratico medio del campione.
Se invece il campione è piccolo, e la popolazione da cui proviene il campione ha distribuzione normale, si può usare il teorema 3, Cap. 6; sulla base di tale teorema la statistica test

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

è una variabile aleatoria avente la distribuzione *t* con grado di libertà $\nu = n - 1$.
I criteri per i test a una e a due code basati sull'uso di questa distribuzione sono analoghi a quelli già descritti nel paragrafo precedente, con z_α e $z_{\frac{\alpha}{2}}$ sostituiti da t_α e $t_{\frac{\alpha}{2}}$; questi valori critici per un dato livello di significatività α dipendono dal grado di libertà e devono essere letti di volta in volta sulle tavole della distribuzione *t*.
Nella tabella 2 riassumiamo i valori comunemente usati per il livello di significatività α e i corrispondenti valori critici t_α e $t_{\frac{\alpha}{2}}$ per i test a una e a due code.

Test	Ipot. nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu \le \mu_0$	$\mu > \mu_0$	0.01	$t_\alpha = t_{0.01}$	$T > t_{0.01}$
			0.05	$t_\alpha = t_{0.05}$	$T > t_{0.05}$
una coda	$\mu \ge \mu_0$	$\mu < \mu_0$	0.01	$t_\alpha = -t_{0.01}$	$T < -t_{0.01}$
			0.05	$t_\alpha = -t_{0.05}$	$T < -t_{0.05}$
due code	$\mu = \mu_0$	$\mu \ne \mu_0$	0.01	$t_{\frac{\alpha}{2}} = t_{0.005}$ $t_{\frac{\alpha}{2}} = -t_{0.005}$	$T > t_{0.005}$ $T < -t_{0.005}$
			0.05	$t_{\frac{\alpha}{2}} = t_{0.025}$ $t_{\frac{\alpha}{2}} = -t_{0.025}$	$T > t_{0.025}$ $T < -t_{0.025}$

Tabella 2

Esempio 16
Le bottiglie di vino poste in vendita contengono usualmente 750ml di vino.
Si effettua un controllo su un campione di 6 bottiglie e si misurano i seguenti valori in ml

747.0	751.5	752.0	747.5	747.0	749.0

Stabilire se questi dati confermano con un livello di significatività del 5% l'affermazione che le bottiglie hanno un contenuto medio pari a quanto dichiarato.

Se il test è effettuato per tutelare l'interesse del consumatore, l'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu \ge 750$
H_1: $\mu < 750$.

Calcolando la media e la varianza del campione si ottengono i seguenti valori

$$\bar{x} = \frac{747.0 + 751.5 + 752.0 + 747.5 + 747.0 + 749.0}{6} = 749$$

$$s^2 = \frac{1}{5} \cdot \left(747.0^2 + 751.5^2 + 752.0^2 + 747.5^2 + \right.$$
$$\left. + 747.0^2 + 749.0^2 - 6 \cdot 749^2 \right) = 5.1$$

Il valore della statistica test è

$$T = \frac{749 - 750}{\frac{\sqrt{5.1}}{\sqrt{6}}} = -1.08 .$$

Il test è a una coda, e per il livello di significatività del 5% e il grado di libertà $\nu = 5$ il valore critico è

$$t_\alpha = -t_{0.05} = -2.015$$

La regione di rifiuto è data dai valori $T < -2.015$.
Il valore $T = -1.08$ appartiene alla regione di accettazione, perciò non si rifiuta l'ipotesi nulla: non c'è un'evidenza significativa, al livello del 5%, che le bottiglie contengano meno di 750ml di vino.

Esempio 17

Una prova del carico di rottura di 6 cavi d'acciaio costruiti da una ditta ha mostrato un carico di rottura medio $\bar{x} = 7750$ kg e uno scarto quadratico medio s = 145kg, mentre il costruttore afferma che il carico di rottura medio è di 8000kg.
E' possibile sostenere che l'affermazione del costruttore non è corretta e che il carico di rottura è inferiore, ai livelli di significatività del 5% e dell'1%?

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu \geq 8000$
H_1: $\mu < 8000$.

Il valore della statistica test è

$$T = \frac{7750 - 8000}{\frac{145}{\sqrt{6}}} = -4.22 .$$

Il test è a una coda, e per il livello di significatività del 5% e il grado di libertà $\nu = n - 1 = 5$, il valore critico è

$$t_\alpha = -t_{0.05} = -2.015$$

La regione di rifiuto è data dai valori $T < -2.015$. Il valore $T = -4.22$ appartiene alla regione di rifiuto, perciò rifiutiamo l'ipotesi nulla al livello di significatività del 5%.
Per il livello di significatività dell'1% e il grado di libertà $\nu = n - 1 = 5$, il valore critico è

$$t_\alpha = -t_{0.01} = -3.365$$

La regione di rifiuto è data dai valori $T < -3.365$. Il valore $T = -4.22$ appartiene alla regione di rifiuto, perciò anche al livello di significatività dell'1% rifiutiamo l'ipotesi nulla.
In conclusione non possiamo sostenere che l'affermazione del costruttore sia giustificata per nessuno dei due livelli di significatività.

Esempio 18

Si estrae un campione di 8 confezioni di detersivo in polvere da una grossa produzione.
I pesi in g delle 8 confezioni sono

1998.5	2000.4	1999.9	2005.8	2011.5	2007.6	2001.3	2002.4

Assumendo che popolazione da cui proviene il campione abbia distribuzione normale, verificare se al livello di significatività del 5%, si può affermare che il peso medio delle confezioni di questa produzione è maggiore di 2000g.

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu \le 2000$

H_1: $\mu > 2000$.

Calcolando la media e la varianza del campione si ottengono i seguenti valori

$$\bar{x} = 2003.4 \qquad s^2 = 19.95$$

Il valore della statistica test è

$$T = \frac{2003.4 - 2000}{\frac{\sqrt{19.95}}{\sqrt{8}}} = 2.15 .$$

Il test è a una coda, e per il livello di significatività del 5% e il grado di libertà $\nu = n - 1 = 7$, il valore critico è

$$t_\alpha = t_{0.05} = 1.895$$

La regione di rifiuto è data dai valori $T > 1.895$. Il valore $T = 2.15$ appartiene alla regione di rifiuto, perciò rifiutiamo l'ipotesi nulla e concludiamo che c'è una significativa evidenza, al livello del 5%, che il contenuto delle scatole sia maggiore di 2000g.

Esempio 19

Il contenuto dichiarato delle bottiglie di una certa bibita è 330ml.

Scegliendo un campione di 20 bottiglie, si riscontra un contenuto medio $\bar{x} = 328$ ml, con uno scarto quadratico medio s = 3.2 ml.

In base a questi dati si può ritenere che la ditta produttrice inganni il consumatore? Si assuma che la quantità di liquido contenuta nelle bottiglie segua approssimativamente la distribuzione normale e si scelga il livello di significatività dell'1%.

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\mu \ge 330$

H_1: $\mu < 330$.

Il valore della statistica test è

$$T = \frac{328 - 330}{\frac{3.2}{\sqrt{20}}} = -2.8 .$$

Il test è a una coda, e per il livello di significatività dell'1% e il grado di libertà $\nu = 19$, il valore critico è

$$t_\alpha = -t_{0.05} = -2.539$$

La regione di rifiuto è data dai valori $T < -2.539$. Il valore $T = -2.8$ appartiene alla regione di rifiuto, perciò rifiutiamo l'ipotesi nulla e concludiamo che c'è una significativa evidenza, al livello dell'1%, che ci sia una frode da parte del produttore.

Per il livello di significatività del 5% il valore critico è

$$t_\alpha = -1.729$$

La regione di rifiuto è data dai valori $T < -1.729$. Il valore $T = -2.8$ appartiene alla regione di rifiuto, perciò anche al livello di significatività del 5% rifiutiamo l'ipotesi nulla, concludendo ancora che c'è una significativa evidenza di frode.

Riassumiamo nella tabella 3 i vari procedimenti da seguire per effettuare un test di ipotesi sulla media μ di una popolazione.

Procedimento	Ipotesi	Statistica test	Distribuzione della statistica test
1	$n \geq 30$ varianza σ^2 nota	$Z = \dfrac{\overline{X} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$	Distribuzione normale
2	$n \geq 30$ varianza σ^2 incognita	$Z = \dfrac{\overline{X} - \mu_0}{\dfrac{s}{\sqrt{n}}}$	Distribuzione normale
3	$n < 30$ popolaz. normale varianza σ^2 nota	$Z = \dfrac{\overline{X} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$	Distribuzione normale
4	$n < 30$ popolaz. normale varianza σ^2 incognita	$T = \dfrac{\overline{X} - \mu_0}{\dfrac{s}{\sqrt{n}}}$	Distribuzione *t* di Student (gradi libertà = $n-1$)

Tabella 3

Osservazione
Il procedimento 4, descritto per piccoli campioni, può essere usato anche per grandi campioni da popolazione normale, al posto del procedimento 2 indicato nella tabella 3: il procedimento esatto è quello basato sulla distribuzione *t* (procedimento 4), mentre l'altro è approssimato (si approssima lo scarto quadratico medio σ con lo scarto *s* del campione). In pratica entrambi i procedimenti sono adatti e portano essenzialmente alle stesse decisioni.

8.6 Test di ipotesi sulla proporzione

Consideriamo in questo paragrafo il problema della verifica di ipotesi sulla proporzione di una popolazione. In alcuni casi si deve sottoporre a test l'ipotesi che la proporzione della popolazione assuma un determinato valore p_0.
Per risolvere problemi di questo tipo si conta il numero *X* di volte in cui la caratteristica osservata si presenta nel campione di ampiezza *n* e si calcola la proporzione campionaria: in altre parole si osserva il numero di successi in *n* prove o proporzione di successi; si ha quindi a che fare con la distribuzione binomiale e si fa un test di ipotesi sul parametro *p* di una popolazione binomiale.
Quando il numero *n* di elementi del campione è sufficientemente grande, il test di ipotesi sulla proporzione può essere basato sulla distribuzione normale.
E' infatti noto che, indicando con *p* la proporzione di successi in *n* prove bernoulliane, se si verifica che $np \geq 5$ e $n(1-p) \geq 5$, la distribuzione binomiale di parametri *n* e *p* può essere approssimata con la distribuzione normale (vedere Cap. 5, §5.5).
Per sottoporre a test l'ipotesi nulla

$$H_0: \quad p = p_0$$

(o, in modo analogo, le ipotesi $p \leq p_0$, $p \geq p_0$) si utilizza la statistica

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

che ha approssimativamente la distribuzione normale standardizzata, per *n* sufficientemente grande, e si procede nel modo già illustrato per i test per la media nel caso dei grandi campioni[7].

[7] Anche in questo caso, come già visto a proposito degli intervalli di confidenza per la proporzione, si dovrebbe effettuare la correzione di continuità, ma, quando *n* è grande, gli effetti di tale correzione sono in generale trascurabili.

Nella tabella 4 (che non differisce sostanzialmente dalla tabella 1) riassumiamo per comodità i valori comunemente usati per il livello di significatività α e i corrispondenti valori critici z_α e $z_{\frac{\alpha}{2}}$ per i test a una e a due code.

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$p \le p_0$	$p > p_0$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$p \ge p_0$	$p < p_0$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$p = p_0$	$p \ne p_0$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Tabella 4

Esempio 20

Si effettuano 500 lanci di una moneta e si ottiene 267 volte testa.

a – Decidere se la moneta è truccata oppure no, con un livello di significatività del 5%.

b – Ripetere il calcolo nel caso che il numero di volte in cui si ottiene testa sia 280.

Per una moneta non truccata la probabilità che esca testa è 0.5.

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $p = p_0 = 0.5$

H_1: $p \ne 0.5$.

Si effettua un test a due code con il livello di significatività $\alpha = 0.05$; la regione di rifiuto è costituita dai valori $Z < -1.96$ e $Z > 1.96$.

a – Si ha

$$n = 500 \qquad x = 267 \qquad p_0 = 0.5$$

$$Z = \frac{267 - 500 \cdot 0.5}{\sqrt{500 \cdot 0.5 \cdot 0.5}} = 1.52$$

Il valore $Z = 1.52$ cade nella regione di accettazione, perciò l'ipotesi nulla non può essere rifiutata; in conclusione la moneta non può ritenersi truccata, al livello di significatività del 5%.

b – Si ha

$$n = 500 \qquad x = 280 \qquad p_0 = 0.5$$

$$Z = \frac{280 - 500 \cdot 0.5}{\sqrt{500 \cdot 0.5 \cdot 0.5}} = 2.68$$

Il valore $Z = 2.68$ cade nella regione di rifiuto, perciò l'ipotesi nulla deve essere rifiutata; in conclusione la moneta può ritenersi truccata, al livello di significatività del 5%.

Esempio 21

Una ditta farmaceutica asserisce che un suo farmaco è efficace nel 90% dei casi. In un campione di 200 persone che lo hanno usato, il farmaco si è rivelato efficace in 160 casi. Stabilire se l'affermazione della ditta farmaceutica è legittima con un livello di significatività uguale a 0.01.

Si assume come ipotesi nulla

H_0: $p \ge 0.9$

e come ipotesi alternativa

H_1: $p < 0.9$.

In questo caso interessa infatti stabilire se l'efficacia del farmaco è minore di quanto affermato; si effettua perciò un test a una coda e la regione di rifiuto è data dai valori $Z < -2.326$.

Si ha

$$n = 200 \qquad x = 160 \qquad p_0 = 0.9$$

$$Z = \frac{160 - 200 \cdot 0.9}{\sqrt{200 \cdot 0.9 \cdot (1 - 0.9)}} = -4.71$$

Il valore $Z = -4.71$ cade nella regione di rifiuto, perciò si rifiuta l'ipotesi nulla, al livello di significatività dell'1%, concludendo che l'affermazione non è legittima.

Esempio 22

Un fabbricante dichiara che almeno il 95% della merce fornita a una ditta è conforme alle esigenze del cliente.
Un esame di un campione di 200 esemplari della merce rivela che 18 esemplari sono difettosi.
Sottoporre a test la dichiarazione del fabbricante al livello di significatività $\alpha = 0.01$ e $\alpha = 0.05$.

Si assume come ipotesi nulla

H_0: $p \geq 0.95$

e come ipotesi alternativa

H_1: $p < 0.95$.

Si effettua un test a una coda e si ha

$$n = 200 \qquad 200 - x = 18 \qquad x = 182 \qquad p_0 = 0.95$$

$$Z = \frac{182 - 200 \cdot 0.95}{\sqrt{200 \cdot 0.95 \cdot (1 - 0.95)}} = -2.60$$

a – Per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è data dai valori $Z < -2.326$.
Il valore $Z = -2.60$ cade nella regione di rifiuto, perciò si rifiuta l'ipotesi nulla, al livello di significatività dell'1%, concludendo che l'affermazione del fabbricante è falsa.
b – Per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z < -1.645$.
Il valore $Z = -2.60$ cade nella regione di rifiuto, perciò anche per questo livello di significatività si rifiuta l'ipotesi nulla, concludendo che l'affermazione del fabbricante è falsa.

Esempio 23

Una compagnia aerea afferma che non più del 6% dei bagagli smarriti viene definitivamente perso.
Sottoporre a test questa affermazione, sapendo che su un campione di 200 persone che hanno subito lo smarrimento del bagaglio, 17 non l'hanno più ritrovato; scegliere il livello di significatività dell'1%.

Si assumono come ipotesi

H_0: $p \leq 0.06$
H_1: $p > 0.06$.

Si effettua un test a una coda e per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è data dai valori $Z > 2.326$.
Si ha

$$n = 200 \qquad x = 17 \qquad p_0 = 0.06$$

$$Z = \frac{17 - 200 \cdot 0.06}{\sqrt{200 \cdot 0.06 \cdot (1 - 0.06)}} = 1.49$$

Il valore $Z = 1.49$ cade nella regione di accettazione, perciò al livello di significatività dell'1% l'affermazione della compagnia aerea non può essere contestata.

Esempio 24

Il responsabile del personale di un'azienda intende ridurre il turn-over, nei primi due anni di lavoro, dei dipendenti addetti alle vendite. Si supponga che da analisi precedenti risulti che il 30% dei dipendenti si licenzi entro i primi due anni.

Si decide di sottoporre 120 neo assunti a un nuovo corso sulle tecniche di vendita; di questi, al termine del secondo anno dall'assunzione, 29 non lavorano più nell'azienda.
Si può affermare che il turn-over si sia ridotto con la partecipazione al corso di addestramento?

Si assumono come ipotesi

H_0: $p \geq 0.30$
H_1: $p < 0.30$

Si effettua un test a una coda e per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è data dai valori $Z < -2.326$.
Si ha

$$n = 120 \qquad x = 29 \qquad p_0 = 0.30$$

$$Z = \frac{29 - 120 \cdot 0.30}{\sqrt{120 \cdot 0.30 \cdot (1 - 0.30)}} = -1.39$$

Il valore $Z = -1.-39$ cade nella regione di accettazione, perciò al livello di significatività dell'1% si può affermare che il turn-over non si è ridotto.
Calcolo del *p*-value

$$p - \text{value} = P(Z < Z_0) = P(Z < -1.39) = 1 - 0.9177 = 0.0823$$

Il valore del *p*-value ci consente di accettare l'ipotesi nulla.

8.7 Test di ipotesi sulla differenza fra due medie (varianze note)

Descriviamo il procedimento per eseguire un test di ipotesi sulla differenza fra le medie di due popolazioni; questo test viene effettuato quando si vogliono confrontare le medie di due popolazioni diverse.
Questa situazione si può verificare in molte indagini comparative: si vuole confrontare la produttività di una macchina con quella di un'altra; si vuole sapere se la popolazione di una certa città ha un reddito medio superiore a quello di un'altra, e così via.

Consideriamo due popolazioni aventi medie μ_1 e μ_2, e varianze σ_1^2 e σ_2^2; vogliamo sottoporre a test una delle ipotesi nulle

H_0: $\mu_1 - \mu_2 \leq d$
H_0: $\mu_1 - \mu_2 \geq d$
H_0: $\mu_1 - \mu_2 = d$

dove d è una costante specificata, basandoci sulle medie di due campioni casuali indipendenti di ampiezza n_1 e n_2. In analogia con il test sulla media già esaminati nei § 8.4 e 8.5, si effettua il test dell'ipotesi nulla, scegliendo la corrispondente ipotesi alternativa fra le seguenti

H_1: $\mu_1 - \mu_2 > d$
H_1: $\mu_1 - \mu_2 < d$
H_1: $\mu_1 - \mu_2 \neq d$

Nei primi due casi si fa un test a una coda, nel terzo un test a due code.
Il test dipende dalla differenza fra le medie campionarie $\overline{X}_1 - \overline{X}_2$ e, in base a quanto già illustrato nel § 7.6, pag. 205, si basa sulla statistica

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Il test qui illustrato è essenzialmente un **test per grandi campioni** ($n \geq 30$); in tal caso la distribuzione della differenza fra le medie campionarie può essere approssimata dalla distribuzione normale e la variabile aleatoria Z ha approssimativamente la distribuzione normale standardizzata.

Nel caso particolare in cui i campioni siano estratti da due popolazioni aventi distribuzione normale con varianze note, la variabile Z ha la distribuzione normale standardizzata, qualunque siano le ampiezze dei campioni (vedere esempi 28, 29).
Sia, come al solito, z_α il valore di Z per cui l'area a destra di z_α al di sotto della curva normale standardizzata è uguale a α.
Nella tabella 5 riassumiamo i valori comunemente usati per il livello di significatività α e i corrispondenti valori critici z_α e $z_{\frac{\alpha}{2}}$ per i test a una e a due code.

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu_1 - \mu_2 \le d$	$\mu_1 - \mu_2 > d$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$\mu_1 - \mu_2 \ge d$	$\mu_1 - \mu_2 < d$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \ne d$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Tabella 5

Anche se d può avere un qualunque valore, nella maggior parte dei problemi il suo valore è zero e si sottopone a test una delle ipotesi nulle

H_0: $\mu_1 \le \mu_2$
H_0: $\mu_1 \ge \mu_2$
H_0: $\mu_1 = \mu_2$

contro la corrispondente ipotesi alternativa

H_1: $\mu_1 > \mu_2$
H_1: $\mu_1 < \mu_2$
H_1: $\mu_1 \ne \mu_2$.

Per poter effettuare il test qui descritto si richiede la conoscenza delle varianze delle popolazioni; nella maggior parte dei casi le varianze σ_1^2 e σ_2^2 non sono note, e nel caso di grandi campioni possono essere sostituite con le varianze campionarie s_1^2 e s_2^2, commettendo un errore di approssimazione.

Esempio 25
Un tema d'esame è stato assegnato a due gruppi di studenti composti rispettivamente da 40 e 50 studenti. Il voto medio del primo gruppo è stato 74/30 con uno scarto quadratico medio s = 8; il voto medio del secondo gruppo è stato invece 78/30 con uno scarto quadratico medio s = 7.
C'è differenza fra le due classi al livello di significatività α = 0.05?

Si assumono come ipotesi nulla e come ipotesi alternativa

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \ne \mu_2$.

Si ha

$$n_1 = 40 \qquad \bar{x}_1 = 74 \qquad s_1 = 8$$
$$n_2 = 50 \qquad \bar{x}_2 = 78 \qquad s_2 = 7$$
$$Z = \frac{74 - 78}{\sqrt{\frac{8^2}{40} + \frac{7^2}{50}}} = -2.49$$

Si effettua un test a due code e per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z < -1.96$ e $Z > 1.96$.
Dato che il valore trovato $Z = -2.49$ appartiene alla regione di rifiuto, si rifiuta l'ipotesi nulla al livello di significatività $\alpha = 0.05$, ossia si decide che i due gruppi sono significativamente diversi.

Esempio 26
Un campione di 100 lampadine della marca A ha mostrato una durata media di 1190 ore ed uno scarto quadratico medio di 90 ore; un campione di 75 lampadine della marca B ha mostrato invece una durata media di 1230 ore ed uno scarto quadratico medio di 120 ore. C'è differenza tra i tempi di durata media delle due marche di lampadine ai livelli di significatività $\alpha = 0.05$ e $\alpha = 0.01$?.

Si assumono come ipotesi

$$H_0: \quad \mu_1 = \mu_2$$
$$H_1: \quad \mu_1 \neq \mu_2.$$

Si ha

$$n_1 = 100 \qquad \bar{x}_1 = 1190 \qquad s_1 = 90$$
$$n_2 = 75 \qquad \bar{x}_2 = 1230 \qquad s_2 = 120$$
$$Z = \frac{1190 - 1230}{\sqrt{\frac{90^2}{100} + \frac{120^2}{75}}} = -2.42$$

a – Si effettua un test a due code e per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z < -1.96$ e $Z > 1.96$.
Dato che il valore trovato $Z = -2.42$ appartiene alla regione di rifiuto, si rifiuta l'ipotesi nulla al livello di significatività $\alpha = 0.05$, ossia si decide che le durate medie sono diverse.
b – Per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è data dai valori $Z < -2.576$ e $Z > 2.576$.
Dato che il valore trovato $Z = -2.42$ appartiene alla regione di accettazione, non si rifiuta l'ipotesi nulla, ossia al livello di significatività $\alpha = 0.01$ si ritiene che le durate medie siano uguali.
Le diverse conclusioni raggiunte ai due livelli di significatività suggeriscono la necessità di ulteriori indagini.
Con il procedimento illustrato per il calcolo del *p*-value nel caso dei test basati sulla distribuzione normale (pag. 234), si può calcolare il *p*-value e si trova

$$Z_0 = -2.42$$
$$p - \text{value} = 2\left[1 - P\left(Z < 2.42\right)\right] = 2\left(1 - 0.9922\right) = 0.0156$$

Questo valore conferma la situazione critica: rifiutare o no l'ipotesi nulla dipende in modo cruciale dal livello di significatività.

Esempio 27
Nel precedente problema sottoporre a test l'ipotesi che le lampadine della marca B sono superiori a quelle della marca A usando i due livelli di significatività.

In questo caso si assumono come ipotesi nulla e come ipotesi alternativa

$$H_0: \quad \mu_1 \geq \mu_2$$
$$H_1: \quad \mu_1 < \mu_2.$$

e si effettua un test a una coda.
a – Per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z < -1.645$.
Il valore $Z = -2.42$ appartiene alla regione di rifiuto, perciò si rifiuta l'ipotesi nulla, concludendo che la marca B è superiore alla marca A.
b – Per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è data dai valori $Z < -2.326$.

Il valore $Z = -2.42$ appartiene alla regione di rifiuto, perciò si rifiuta l'ipotesi nulla, concludendo anche a questo livello di significatività che la marca B è superiore alla marca A.
Il *p*-value in questo caso è

$$p - \text{value} = P(Z < -2.42) = 1 - P(Z < 2.42) = 1 - 0.9922 = 0.0078$$

Le conclusioni raggiunte in questo e nel precedente test non sono in contraddizione tra loro perché l'ipotesi alternativa nei due test effettuati è diversa.

Esempio 28

Si intendono confrontare i tempi di asciugatura di due vernici aventi composizione chimica che differisce per un componente. A tale scopo si effettuano 10 prove con vernice del primo tipo e 10 prove con vernice del secondo tipo e si misurano i relativi tempi di asciugatura, trovando i valori medi $\bar{x}_1 = 121$ minuti e $\bar{x}_2 = 112$ minuti.

Si può ritenere che le popolazioni abbiano distribuzione normale con scarto quadratico medio $\sigma_1 = \sigma_2 = 8$ minuti.

Sottoporre a test l'ipotesi che la prima vernice asciughi più rapidamente della seconda al livello di significatività $\alpha = 0.05$.

Poiché le popolazioni hanno distribuzione normale, si può usare la distribuzione normale anche per piccoli campioni.
Si assumono come ipotesi

H_0: $\mu_1 \le \mu_2$
H_1: $\mu_1 > \mu_2$.

Si ha

$$n_1 = 10 \qquad \bar{x}_1 = 121$$
$$n_2 = 10 \qquad \bar{x}_2 = 112$$
$$\sigma_1 = \sigma_2 = 8$$
$$Z = \frac{121 - 112}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2.52$$

Si effettua un test a una coda e al livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z > 1.645$; il valore $Z = 2.52$ appartiene alla regione di rifiuto, e concludiamo che la seconda vernice asciuga più rapidamente della prima.

Esempio 29

Nella tabella sono riportate le misure del peso in g di due campioni di 10 oggetti dello stesso tipo prodotti da due macchine diverse; gli oggetti sono scelti a caso da due popolazioni aventi entrambe la distribuzione normale, con varianze $\sigma_1^2 = 1.8$ e $\sigma_2^2 = 1.3$.

Sottoporre a test l'ipotesi che le due popolazioni abbiano la stessa media. Scegliere il livello di significatività $\alpha = 0.05$.

Campione 1	37.2	39.7	37.2	38.8	37.7	36.6	37.5	40.5	38.2	36.6
Campione 2	35.6	35.0	34.9	36.0	36.6	36.1	35.8	34.9	38.6	36.5

Per i due campioni si ha

Campione 1	$\sum_{i=1}^{10} x_i = 380$	$\bar{x} = 38$
Campione 2	$\sum_{i=1}^{10} x_i = 360$	$\bar{x} = 36$

Poiché le popolazioni hanno distribuzione normale, si può usare la statistica test *Z* anche per piccoli campioni.
Si assumono come ipotesi nulla e come ipotesi alternativa

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$.

Si ha

$$Z = \frac{38 - 36}{\sqrt{\frac{1.8}{10} + \frac{1.3}{10}}} = 3.59$$

Si effettua un test a due code e al livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z < -1.96$ o $Z > 1.96$; il valore $Z = 3.59$ appartiene alla regione di rifiuto, e concludiamo che le due popolazioni hanno media diversa.

8.8 Test di ipotesi sulla differenza fra due medie (varianze incognite)

Esaminiamo ora il test di ipotesi sulla differenza fra due medie nel caso in cui non siano note le varianze delle due popolazioni; come già detto nel § precedente, nel caso di grandi campioni le varianze incognite possono essere sostituite con i valori delle varianze campionarie dei due campioni.
Se invece si usano piccoli campioni, per stimare la differenza fra le medie delle due popolazioni si può far ricorso alla distribuzione *t*, ma le due popolazioni devono avere distribuzione normale; inoltre, come già visto nel § 7.7, pag. 207, occorre distinguere due casi: il caso in cui le varianze delle due popolazioni sono uguali e il caso in cui sono diverse.
In queste lezioni sarà trattato solo il caso in cui le varianze sono uguali; il fatto che le varianze di due popolazioni siano uguali può, a sua volta, essere oggetto di un test statistico, che sarà discusso nel § 8.11. Perciò, volendo eseguire un test sulla differenza di due medie nel caso in cui le varianze siano incognite e i campioni piccoli, l'indagine può procedere in due tempi: prima si verifica l'ipotesi di uguaglianza delle varianze, poi, se l'uguaglianza è verificata, si applica il test sulla differenza fra le medie.
Nel caso in cui le due popolazioni normali hanno la stessa varianza, si ricava la **stima congiunta della varianza** comune con la seguente formula ((7.18), pag. 208)

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{8.1}$$

dove n_1 e n_2 sono le ampiezze dei due campioni e S_1^2 e S_2^2 sono le rispettive varianze campionarie.
Per **piccoli campioni** (***n* < 30**), nell'ipotesi che le popolazioni da cui si estraggono i campioni abbiano distribuzione normale con la stessa varianza, si può dimostrare che la statistica

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

ha la distribuzione *t* con grado di libertà $\nu = n_1 + n_2 - 2$.
I test a una e a due code basati sull'uso di questa distribuzione sono analoghi a quelli già descritti nel paragrafo precedente, con z_α e $z_{\frac{\alpha}{2}}$ sostituiti da t_α e $t_{\frac{\alpha}{2}}$.
I valori di t_α e $t_{\frac{\alpha}{2}}$ per un dato livello di significatività α dipendono dal grado di libertà e devono essere letti di volta in volta sulle tavole della distribuzione *t*.
Il valore del grado di libertà $\nu = n_1 + n_2 - 2$ può essere maggiore di 29: in tal caso si utilizzano i

valori critici dell'ultima riga della tabella della distribuzione *t*.
Nella tabella 6 riassumiamo i valori comunemente usati per il livello di significatività α e i corrispondenti valori critici t_α e $t_{\frac{\alpha}{2}}$ per i test a una e a due code.

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu_1 - \mu_2 \le d$	$\mu_1 - \mu_2 > d$	0.01	$t_\alpha = t_{0.01}$	$T > t_{0.01}$
			0.05	$t_\alpha = t_{0.05}$	$T > t_{0.05}$
una coda	$\mu_1 - \mu_2 \ge d$	$\mu_1 - \mu_2 < d$	0.01	$t_\alpha = -t_{0.01}$	$T < -t_{0.01}$
			0.05	$t_\alpha = -t_{0.05}$	$T < -t_{0.05}$
due code	$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \ne d$	0.01	$t_{\frac{\alpha}{2}} = t_{0.005}$ $t_{\frac{\alpha}{2}} = -t_{0.005}$	$T > t_{0.005}$ $T < -t_{0.005}$
			0.05	$t_{\frac{\alpha}{2}} = t_{0.025}$ $t_{\frac{\alpha}{2}} = -t_{0.025}$	$T > t_{0.025}$ $T < -t_{0.025}$

Tabella 6

Esempio 30
Nella tabella sono riportate le lunghezze in cm di due campioni di oggetti dello stesso tipo prodotti da due macchine diverse (esempio 19, Cap. 7, pag. 208).

Campione 1	8.26	8.13	8.35	8.07	8.34		
Campione 2	7.95	7.89	7.90	8.14	7.92	7.84	7.94

Sottoporre a test l'ipotesi che gli oggetti prodotti abbiano lunghezze significativamente diverse al livello di significatività $\alpha = 0.05$, supponendo che le popolazioni da cui provengono i campioni abbiano distribuzione normale con la stessa varianza.

Si assumono come ipotesi nulla e come ipotesi alternativa

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \ne \mu_2$.

In base ai dati della tabella si ha

$$n_1 = 5 \qquad \bar{x}_1 = 8.23 \qquad s_1^2 = 0.01575$$
$$n_2 = 7 \qquad \bar{x}_2 = 7.94 \qquad s_2^2 = 0.00910$$

La stima congiunta della varianza con la formula (8.1) è

$$S^2 = \frac{4 \cdot 0.01575 + 6 \cdot 0.00910}{5 + 7 - 2} = 0.01176$$

Il grado di libertà della distribuzione *t* è

$$\nu = n_1 + n_2 - 2 = 5 + 7 - 2 = 10$$

La statistica *T* ha il valore

$$T = \frac{8.23 - 7.94}{\sqrt{0.01176 \cdot \left(\frac{1}{5} + \frac{1}{7}\right)}} = 4.57$$

Per il livello di significatività $\alpha = 0.05$ il valore critico è $t_{0.025} = 2.228$ e la regione di rifiuto è data dai valori $T < -2.228$ e $T > 2.228$.

Il valore $T = 4.57$ appartiene alla regione di rifiuto, perciò l'ipotesi nulla deve essere rifiutata e si conclude che le lunghezze sono diverse al livello di significatività $\alpha = 0.05$.
Usando il livello di significatività $\alpha = 0.01$ il valore critico è $t_{0.005} = 3.169$ e anche in questo caso l'ipotesi nulla deve essere rifiutata.

Esempio 31
Due tipi di soluzioni chimiche sono state provate per misurarne il pH (grado di acidità della soluzione). L'analisi di 6 campioni della prima soluzione ha mostrato un pH medio di 7.52, con uno scarto quadratico medio di 0.032; l'analisi di 5 campioni della seconda soluzione ha mostrato un pH medio di 7.49 con uno scarto quadratico medio di 0.024. Stabilire se le due soluzioni abbiano valori uguali o diversi del pH usando il livello di significatività $\alpha = 0.05$.

Per poter usare il test di ipotesi basato sulla distribuzione *t* bisogna supporre che le distribuzioni delle due popolazioni siano normali con la stessa varianza (vedere esempio 43)..
Si assumono come ipotesi nulla e come ipotesi alternativa

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

In base ai dati si ha

$$n_1 = 6 \qquad \bar{x}_1 = 7.52 \qquad s_1 = 0.032$$
$$n_2 = 5 \qquad \bar{x}_2 = 7.49 \qquad s_2 = 0.024$$

La stima congiunta della varianza con la formula (8.1) è

$$S^2 = \frac{5 \cdot 0.032^2 + 4 \cdot 0.024^2}{6 + 5 - 2} = 0.000825$$

Il grado di libertà della distribuzione *t* è

$$\nu = n_1 + n_2 - 2 = 6 + 5 - 2 = 9$$

La statistica *T* ha il valore

$$T = \frac{7.52 - 7.49}{\sqrt{0.000825 \cdot \left(\frac{1}{6} + \frac{1}{5}\right)}} = 1.72$$

Per il livello di significatività $\alpha = 0.05$ il valore critico è $t_{0.025} = 2.262$ e la regione di rifiuto è data dai valori $T < -2.262$ e $T > 2.262$.
Il valore $T = 1.72$ appartiene alla regione di accettazione, perciò l'ipotesi nulla non può essere rifiutata e si conclude che al livello di significatività $\alpha = 0.05$ le due soluzioni hanno lo stesso grado di acidità.
Usando il livello di significatività $\alpha = 0.01$ il valore critico è $t_{0.005} = 3.250$ e anche in questo caso l'ipotesi nulla non può essere rifiutata.

Esempio 32
L'osservazione dei guasti occorsi a due tipi di macchine fotocopiatrici ha registrato che 25 guasti della prima macchina hanno richiesto un tempo medio di riparazione di 90.8 minuti, con uno scarto quadratico medio di 21.4 minuti, mentre 25 guasti della seconda macchina hanno richiesto un tempo medio di riparazione di 83.2 minuti con uno scarto quadratico medio di 19.3 minuti.
Eseguire un test, al livello di significatività del 5%, sull'ipotesi nulla di uguaglianza fra i tempi medi di riparazione.

Supponiamo che i tempi medi di riparazione seguano una distribuzione normale e che le varianze delle due popolazioni siano uguali (vedere esempio 44).
Si assumono come ipotesi

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

In base ai dati si ha

$$n_1 = 25 \qquad \bar{x}_1 = 90.8 \qquad s_1 = 21.4$$

$$n_2 = 25 \qquad \bar{x}_2 = 83.2 \qquad s_2 = 19.3$$

La stima congiunta della varianza con la formula (8.1) è

$$S^2 = \frac{24 \cdot 21.4^2 + 24 \cdot 19.3^2}{25 + 25 - 2} = 415.225$$

Il grado di libertà della distribuzione *t* è

$$\nu = n_1 + n_2 - 2 = 48$$

La statistica *T* ha il valore

$$T = \frac{90.8 - 83.2}{\sqrt{415.225 \cdot \left(\frac{1}{25} + \frac{1}{25}\right)}} = 1.32$$

Per il livello di significatività $\alpha = 0.05$ il valore critico è $t_{0.025} = 1.96$ e la regione di rifiuto è data dai valori $T < -1.96$ e $T > 1.96$.
Il valore $T = 1.32$ appartiene alla regione di accettazione, perciò l'ipotesi nulla non deve essere rifiutata e si conclude che al livello di significatività $\alpha = 0.05$ i tempi medi di riparazione sono uguali.

8.9 Test di ipotesi sulla differenza fra due proporzioni

Un altro problema statistico piuttosto comune è quello di voler confrontare tra loro le proporzioni di due popolazioni: ad esempio, ci chiediamo se in due gruppi diversi di persone la proporzione di coloro che hanno una certa caratteristica sia uguale o diversa. Come al solito, il senso della domanda è il seguente: la differenza fra le proporzioni rilevate su due campioni casuali estratti dalle due popolazioni è statisticamente significativa, o invece si può ritenere solo effetto del caso?
Consideriamo quindi il test sulla differenza fra due proporzioni p_1 e p_2 ed esaminiamo in particolare il caso dei grandi campioni.
Si estraggono due campioni di ampiezza rispettivamente n_1 e n_2 (grandi campioni) e siano X_1 e X_2 i numeri di volte in cui la caratteristica osservata si presenta nei due campioni; le proporzioni campionarie $\hat{P}_1 = \frac{X_1}{n_1}$ e $\hat{P}_2 = \frac{X_2}{n_2}$ sono stimatori corretti delle proporzioni p_1 e p_2 delle due popolazioni.

La statistica

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

ha approssimativamente la distribuzione normale standardizzata, per valori sufficientemente grandi di n_1 e n_2.
In particolare per sottoporre a test l'ipotesi nulla

$$H_0: \quad p_1 = p_2$$

ci serviamo del fatto che $p_1 = p_2 = p$, e la statistica test diventa

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Come stima congiunta della proporzione *p* della popolazione si usa il valore

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}.$$

(Si procede in modo analogo se l'ipotesi nulla è H_0: $p_1 \le p_2$ oppure H_0: $p_1 \ge p_2$)
Nella tabella 7 riassumiamo i valori comunemente usati per il livello di significatività α e i corrispondenti valori critici z_α e $z_{\frac{\alpha}{2}}$ per i test a una e a due code.

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$p_1 \le p_2$	$p_1 > p_2$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$p_1 \ge p_2$	$p_1 < p_2$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$p_1 = p_2$	$p_1 \ne p_2$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Tabella 7

Il test qui descritto si applica all'ipotesi nulla $p_1 = p_2$ (o alle ipotesi simili riportate nella tabella 7), ma può essere modificato per applicarlo anche al caso più generale $p_1 - p_2 = d$.

Esempio 33
Due gruppi di 100 persone, tutte sofferenti della stessa malattia, partecipano a uno studio per la sperimentazione di un nuovo farmaco. Al gruppo A viene somministrato il farmaco, che non viene somministrato al gruppo B (detto gruppo di controllo); per ogni altra terapia i due gruppi vengono trattati nello stesso modo.
Si osserva che nei due gruppi guariscono dalla malattia rispettivamente 78 e 65 persone. Si sottoponga a test l'ipotesi che il farmaco è efficace nel curare la malattia ai due livelli di significatività $\alpha = 0.01$ e $\alpha = 0.05$.

Si assumono come ipotesi nulla e come ipotesi alternativa

H_0: $p_1 \le p_2$
H_1: $p_1 > p_2$.

In base ai dati si ha

$$\hat{p}_1 = \frac{78}{100} = 0.78 \qquad \hat{p}_2 = \frac{65}{100} = 0.65 \qquad \hat{p} = \frac{78+65}{200} = 0.715$$

La statistica Z ha il valore

$$Z = \frac{0.78 - 0.65}{\sqrt{0.715 \cdot (1 - 0.715) \cdot \left(\frac{1}{100} + \frac{1}{100}\right)}} = 2.03$$

Per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è costituita dai valori $Z > 2.326$; il valore $Z = 2.03$ appartiene alla regione di accettazione, perciò si conclude che il farmaco è inefficace e le differenze sono dovute al caso.
Per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è costituita dai valori $Z > 1.645$; il valore $Z = 2.03$ appartiene alla regione di rifiuto, perciò si conclude che il farmaco è efficace.
Si noti che le conclusioni tratte con il test dipendono da quanto si vuole rischiare di sbagliare. Se i risultati sono in realtà dovuti al caso e concludiamo erroneamente che sono dovuti al farmaco (errore di primo tipo), potremmo procedere a somministrare il farmaco a molti individui, solo per accorgerci, dopo qualche tempo, che il farmaco stesso è in realtà inutile.
Possiamo invece concludere che il farmaco è inefficace, quando in realtà invece è utile (errore di secondo tipo), decidendo di non somministrarlo ai malati e questa conclusione è pericolosa, specialmente nel caso di malattie gravi.

Esempio 34
Risolvere il problema precedente nel caso in cui ogni gruppo è composto da 200 persone e ne guariscono rispettivamente 156 e 130.

I valori di $\hat{p}_1$, $\hat{p}_2$ e $\hat{p}$ sono gli stessi di prima; il valore di Z diventa

$$Z = \frac{0.78 - 0.65}{\sqrt{0.715 \cdot (1 - 0.715) \cdot \left(\frac{1}{200} + \frac{1}{200} \right)}} = 2.88$$

Ad entrambi i livelli di significatività l'ipotesi nulla deve essere rifiutata, perché il valore $Z = 2.88$ appartiene alle regioni di rifiuto (che sono le stesse di prima). Ciò mette il rilievo il fatto che, aumentando l'ampiezza dei campioni, possiamo aumentare l'affidabilità della decisione.

Esempio 35
Due campioni rispettivamente di 300 votanti della regione A e di 200 votanti della regione B, hanno mostrato che il 56% e il 48% sono favorevoli ad un certo candidato.
Al livello di significatività $\alpha = 0.05$ provare che
a – c'è differenza nella preferenza fra le due regioni;
b – il candidato è preferito nella regione A.

Si ha

$$\hat{p}_1 = 0.56 \qquad \hat{p}_2 = 0.48$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \cdot \hat{p}_1 + n_2 \cdot \hat{p}_2}{n_1 + n_2} = \frac{300 \cdot 0.56 + 200 \cdot 0.48}{500} = 0.528$$

$$Z = \frac{0.56 - 0.48}{\sqrt{0.528 \cdot (1 - 0.528) \cdot \left(\frac{1}{300} + \frac{1}{200} \right)}} = 1.75$$

a – Se si vuole determinare se c'è differenza fra le regioni dobbiamo decidere fra le ipotesi

H_0: $p_1 = p_2$
H_1: $p_1 \neq p_2$

Per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z < -1.96$ e $Z > 1.96$; il valore $Z = 1.75$ appartiene alla regione di accettazione, perciò concludiamo che non c'è differenza significativa fra le due regioni.
b – Se si vuole determinare se il candidato è preferito nella regione A dobbiamo decidere fra le ipotesi

H_0: $p_1 \leq p_2$
H_1: $p_1 > p_2$

e si effettua un test a una coda.
Per il livello di significatività $\alpha = 0.05$ la regione di rifiuto è data dai valori $Z > 1.645$; il valore $Z = 1.75$ appartiene alla regione di rifiuto, perciò concludiamo che il candidato è preferito nella regione A.

Esempio 36
A due campioni di telespettatori, formati rispettivamente da 500 maschi e da 600 femmine, è stato chiesto se sono interessati a vedere le partite di calcio in TV; ha risposto sì il 75% dei maschi e il 60% delle femmine.
Verificare l'ipotesi che la differenza rispetto al sesso sia significativa al livello dell'1%.

Sottoponiamo a test le ipotesi

H_0: $p_1 = p_2$
H_1: $p_1 \neq p_2$

effettuando un test a due code.

Si ha

$$\hat{p}_1 = 0.75 \qquad \hat{p}_2 = 0.60$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \cdot \hat{p}_1 + n_2 \cdot \hat{p}_2}{n_1 + n_2} = \frac{500 \cdot 0.75 + 600 \cdot 0.60}{500 + 600} = 0.668$$

$$Z = \frac{0.75 - 0.60}{\sqrt{0.668 \cdot (1 - 0.668) \cdot \left(\frac{1}{500} + \frac{1}{600} \right)}} = 5.26$$

Per il livello di significatività $\alpha = 0.01$ la regione di rifiuto è data dai valori Z <–2.576 e Z >2.576; il valore Z = 5.26 appartiene alla regione di rifiuto, perciò concludiamo che c'è differenza significativa fra maschi e femmine.
Il calcolo del *p*-value, che è molto prossimo a zero, garantisce che siamo praticamente certi di non sbagliare rifiutando l'ipotesi nulla

$$p - \text{value} = 2\left[1 - P(Z < 5.26)\right] = 2(1 - 0.999999928) = 0.00000014$$

8.10 Test di ipotesi sulla varianza e sullo scarto quadratico medio

Studiamo ora come effettuare un test sulla varianza, ossia come stabilire se la varianza di una popolazione è uguale a un dato valore σ_0^2. Questo tipo di test è utile quando si studia la variabilità di un prodotto, di un processo o di un'operazione.
Il test sull'ipotesi nulla[8]

$$H_0: \qquad \sigma^2 = \sigma_0^2$$

è basato sulle stesse ipotesi già richieste per gli intervalli di confidenza per la varianza. Si suppone che il campione di *n* elementi provenga da una popolazione avente la distribuzione normale e si usa come statistica la variabile

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

che ha la distribuzione χ^2 con grado di libertà $\nu = n - 1$.

In analogia con il test di ipotesi per la media, le regioni di rifiuto dipendono dall'ipotesi alternativa e il test può essere a una o a due code.
I valori critici che delimitano la regione di rifiuto dipendono dal grado di libertà ν e sono rispettivamente $\chi^2_{1-\alpha}$ o χ^2_{α} per i due tipi di test a una coda, $\chi^2_{\frac{\alpha}{2}}$ e $\chi^2_{1-\frac{\alpha}{2}}$ per il test a due code.

Questi valori possono essere letti sulla tavola della distribuzione χ^2 per il grado di libertà usato.

Si noti che per il test a due code si usano code di uguale ampiezza, come nel caso degli intervalli di confidenza per la varianza.
I valori comunemente usati per il livello di significatività sono, come al solito, $\alpha = 0.01$ e $\alpha = 0.05$.
Nella tabella 8 riassumiamo i valori comunemente usati per il livello di significatività α e le corrispondenti regioni di rifiuto per i test a una e a due code.

[8] Si ragiona in modo simile se l'ipotesi nulla è H_0: $\sigma^2 \le \sigma_0^2$ oppure H_0: $\sigma^2 \ge \sigma_0^2$.

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\sigma^2 \le \sigma_0^2$	$\sigma^2 > \sigma_0^2$	0.01	$\chi^2_\alpha = \chi^2_{0.01}$	$\chi^2 > \chi^2_{0.01}$
			0.05	$\chi^2_\alpha = \chi^2_{0.05}$	$\chi^2 > \chi^2_{0.05}$
una coda	$\sigma^2 \ge \sigma_0^2$	$\sigma^2 < \sigma_0^2$	0.01	$\chi^2_{1-\alpha} = \chi^2_{0.99}$	$\chi^2 < \chi^2_{0.99}$
			0.05	$\chi^2_{1-\alpha} = \chi^2_{0.95}$	$\chi^2 < \chi^2_{0.95}$
due code	$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	0.01	$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995}$ $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005}$	$\chi^2 < \chi^2_{0.995}$ $\chi^2 > \chi^2_{0.005}$
			0.05	$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975}$ $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025}$	$\chi^2 < \chi^2_{0.975}$ $\chi^2 > \chi^2_{0.025}$

Tabella 8

Esempio 37

E' noto che una certa popolazione normale ha media $\mu = 44$ e varianza $\sigma^2 = 22.5$
Da un'altra popolazione viene estratto il campione

16	10	12	8	0	12	10	6	10	8	4	2

Si può concludere al livello di significatività del 5% che la seconda popolazione abbia la stessa varianza della prima?

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\sigma^2 = 22.5$

H_1: $\sigma^2 \ne 22.5$.

Dai dati del campione si calcola la varianza campionaria $s^2 = 20.697$. Il valore della statistica test è

$$\chi^2 = \frac{11 \cdot 20.697}{22.5} = 10.12$$

Il test è a due code e la regione di rifiuto è $\chi^2 < \chi^2_{0.975}$ e $\chi^2 > \chi^2_{0.025}$; al livello di significatività del 5% e per il grado di libertà $\nu = 11$, sulle tavole della distribuzione χ^2 si trova

$$\chi^2_{0.975} = 3.816 \qquad \chi^2_{0.025} = 21.920$$

Il valore $\chi^2 = 10.12$ appartiene alla regione di accettazione, perciò l'ipotesi nulla non viene rifiutata, e si decide che le varianze delle due popolazioni sono uguali.

Esempio 38

Il peso di certi pacchetti confezionati automaticamente è distribuito secondo una distribuzione normale con scarto quadratico medio $\sigma = 0.25$g. L'esame di un campione di 20 confezioni ha permesso di calcolare uno scarto quadratico campionario $s = 0.32$g.
L'apparente aumento dello scarto quadratico medio, ossia della variabilità, è significativo al livello $\alpha = 0.05$? E al livello $\alpha = 0.01$?

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\sigma \le 0.25$

H_1: $\sigma > 0.25$.

Il valore di χ^2 è

$$\chi^2 = \frac{19 \cdot 0.32^2}{0.25^2} \cong 31.13$$

Il test è a una coda e la regione di rifiuto è $\chi^2 > \chi^2_\alpha$; al livello di significatività $\alpha = 0.05$ e per il grado di libertà $\nu = 19$, sulle tavole della distribuzione χ^2 si trova

$$\chi^2_\alpha = \chi^2_{0.05} = 30.144 .$$

Il valore $\chi^2 = 31.13$ appartiene alla regione di rifiuto, perciò l'ipotesi nulla viene rifiutata, e si decide che la variabilità è aumentata.
Al livello di significatività $\alpha = 0.01$ e per il grado di libertà $\nu = 19$, sulle tavole della distribuzione χ^2 si trova

$$\chi^2_\alpha = \chi^2_{0.01} = 36.191 .$$

Il valore $\chi^2 = 31.13$ appartiene in questo caso alla regione di accettazione, perciò l'ipotesi nulla viene accettata, e si decide che la variabilità non è aumentata e il risultato è dovuto al caso.
Concludiamo quindi che la variabilità potrebbe essere aumentata e sarebbe prudente effettuare un controllo sul buon funzionamento della macchina.

Esempio 39
Lo scarto quadratico medio delle temperature annuali di una città in un periodo di 100 anni è stato di 8°C. Misurando la temperatura media del quindicesimo giorno di ogni mese durante gli ultimi 15 anni si è riscontrato che lo scarto quadratico medio delle temperature annuali è stato di 5°C.
Sottoporre a test l'ipotesi che la temperatura della città sia diventata meno variabile che in passato, usando i livelli di significatività $\alpha = 0.05$ e $\alpha = 0.01$.

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\sigma \geq 8$
H_1: $\sigma < 8$.

Il valore della statistica test χ^2 è

$$\chi^2 = \frac{14 \cdot 5^2}{8^2} \cong 5.47$$

Il test è a una coda e la regione di rifiuto è $\chi^2 < \chi^2_\alpha$; al livello di significatività $\alpha = 0.05$ e per il grado di libertà $\nu = 14$, sulle tavole della distribuzione χ^2 si trova

$$\chi^2_\alpha = \chi^2_{0.95} = 6.575 .$$

Il valore $\chi^2 = 5.47$ appartiene alla regione di rifiuto, perciò l'ipotesi nulla viene rifiutata, concludendo che la diminuzione della variabilità della temperatura è significativa al livello del 5% .
Al livello di significatività $\alpha = 0.01$ e per il grado di libertà $\nu = 14$, sulle tavole si trova

$$\chi^2_\alpha = \chi^2_{0.99} = 4.660 .$$

Il valore $\chi^2 = 5.47$ appartiene in questo caso alla regione di accettazione, perciò l'ipotesi nulla viene accettata, concludendo che la variabilità della temperatura non è cambiata e il risultato è dovuto al caso. Le conclusioni a cui si giunge per i due livelli di significatività sono opposte: si tratta di un caso dubbio, per il quale occorrono ulteriori indagini.

Questo test basato sull'uso della distribuzione χ^2 è valido sia per piccoli che per grandi campioni, purché provenienti da una popolazione normale; in pratica viene però usato solo per piccoli campioni.

Infatti, se il campione è grande e proviene da popolazione normale, si può usare la statistica

$$Z = \frac{S - \sigma_0}{\frac{\sigma_0}{\sqrt{2n}}}$$

che ha approssimativamente la distribuzione normale standardizzata per *n* sufficientemente grande.
I valori critici definenti le regioni di rifiuto sono gli stessi usati per i test di ipotesi sulla media per grandi campioni; tali valori possono essere letti nella tabella 9 (analoga alla tabella 1).

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\sigma^2 \le \sigma_0^2$	$\sigma^2 > \sigma_0^2$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$\sigma^2 \ge \sigma_0^2$	$\sigma^2 < \sigma_0^2$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Tabella 9

Esempio 40
Si misura la temperatura di ebollizione di 100 campioni di un liquido e si trova uno scarto quadratico medio campionario $s = 0.099^\circ C$.
Si può affermare al livello di significatività $\alpha = 0.01$ che la varianza della distribuzione della popolazione da cui proviene il campione sia minore di 0.015?
Supporre che la popolazione abbia distribuzione almeno approssimativamente normale.

L'ipotesi nulla e l'ipotesi alternativa sono

H_0: $\sigma^2 \ge 0.015$

H_1: $\sigma^2 < 0.015$.

Si ha

$n = 100$ $s = 0.099$

$\sigma_0^2 = 0.015$ $\sigma_0 = 0.1225$

$$Z = \frac{0.099 - 0.1225}{\frac{0.1225}{\sqrt{200}}} = -2.71$$

Per il livello di significatività $\alpha = 0.01$ il valore critico è $z_\alpha = -2.326$ e la regione di rifiuto è data dai valori $Z < -2.326$; il valore $Z = -2.71$ appartiene alla regione di rifiuto, perciò possiamo rifiutare l'ipotesi nulla; dobbiamo perciò concludere che, al livello di significatività $\alpha = 0.01$ la varianza della popolazione è minore di 0.015.

8.11 Test di ipotesi sul rapporto di due varianze

Spesso si pone il problema di verificare se due popolazioni indipendenti hanno la stessa varianza.
Il confronto fra le varianze di due popolazioni può avere un significato a se stante: si pensi ad esempio all'esigenza di fare un confronto sull'accuratezza di un processo di produzione quando si usano due macchine diverse.

Il test può anche essere effettuato per verificare l'applicabilità del test sulla differenza fra le medie descritto nel § 8.8; tale test può infatti essere utilizzato solo se le varianze delle due popolazioni da cui si estraggono i campioni sono uguali. In questo caso il test sull'uguaglianza delle varianze diventa un prerequisito per applicarne un altro.

Il test per confrontare due varianze σ_1^2 e σ_2^2 si basa sul rapporto fra le due varianze campionarie.

Considerando due popolazioni aventi distribuzione normale, si estraggano da esse due campioni indipendenti di ampiezza rispettivamente n_1 e n_2. Le varianze dei due campioni siano s_1^2 e s_2^2, e si indichi con s_1^2 la più grande delle due varianze campionarie.

Se queste ipotesi sono verificate, in base al teorema 5, Cap. 6, si può affermare che la statistica

$$F = \frac{S_1^2}{S_2^2}$$

ha la distribuzione F con gradi di libertà $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$.

In analogia con il test di ipotesi per la media, le regioni di rifiuto dipendono dall'ipotesi alternativa e il test può essere a una o a due code; per il test a due code si usano code di uguale ampiezza, come nel caso del test di ipotesi per la varianza.

I valori critici che delimitano la regione di rifiuto dipendono dal grado di libertà ν e sono rispettivamente $F_{\alpha}(\nu_1,\nu_2)$ o $F_{1-\alpha}(\nu_1,\nu_2)$ per i due tipi di test a una coda, $F_{\frac{\alpha}{2}}(\nu_1,\nu_2)$ e $F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2)$ per il test a due code.

Questi valori possono essere letti sulla tavola della distribuzione F.

Per trovare i valori critici si deve usare la formula seguente (formula (6.8), Cap. 6, pag. 187)

$$F_{1-\alpha}(\nu_1,\nu_2) = \frac{1}{F_{\alpha}(\nu_2,\nu_1)} \tag{8.2}$$

I valori comunemente usati per il livello di significatività sono, come al solito, $\alpha = 0.01$ e $\alpha = 0.05$.

Nella tabella 10 riassumiamo i valori comunemente usati per il livello di significatività α e le corrispondenti regioni di rifiuto per i test a una e a due code.

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\sigma_1^2 \le \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	0.01	$F_{\alpha} = F_{0.01}$	$F > F_{0.01}$
			0.05	$F_{\alpha} = F_{0.05}$	$F > F_{0.05}$
una coda	$\sigma_1^2 \ge \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	0.01	$F_{1-\alpha} = F_{0.99}$	$F < F_{0.99}$
			0.05	$F_{1-\alpha} = F_{0.95}$	$F < F_{0.95}$
due code	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \ne \sigma_2^2$	0.01	$F_{1-\frac{\alpha}{2}} = F_{0.995}$ $F_{\frac{\alpha}{2}} = F_{0.005}$	$F < F_{0.995}$ $F > F_{0.005}$
			0.05	$F_{1-\frac{\alpha}{2}} = F_{0.975}$ $F_{\frac{\alpha}{2}} = F_{0.025}$	$F < F_{0.975}$ $F > F_{0.025}$

Tabella 10

Esempio 41

Da due popolazioni aventi distribuzione normale vengono estratti due campioni indipendenti aventi le seguenti caratteristiche

$$n_1 = 16 \qquad s_1^2 = 47.3$$

$$n_2 = 13 \qquad s_1^2 = 36.4$$

Sottoporre a test l'opportuna ipotesi nulla scegliendo come ipotesi alternativa

a – H_1: $\sigma_1^2 > \sigma_2^2$

b – H_1: $\sigma_1^2 < \sigma_2^2$

c – H_1: $\sigma_1^2 \neq \sigma_2^2$

Usare i livelli di significatività $\alpha = 0.05$ e $\alpha = 0.01$.

I gradi di libertà della distribuzione F sono

$$\nu_1 = n_1 - 1 = 15 \qquad \nu_2 = n_2 - 1 = 12$$

Livello di significatività $\alpha = 0.05$

a – H_0: $\sigma_1^2 \leq \sigma_2^2$

H_1: $\sigma_1^2 > \sigma_2^2$

Il valore critico per il test è

$$F_\alpha(\nu_1, \nu_2) = F_{0.05}(15,12) = 2.62$$

Il valore della statistica F è

$$F = \frac{47.3}{36.4} = 1.30$$

La regione di rifiuto è costituita dai valori $F > 2.62$, perciò l'ipotesi nulla non deve essere rifiutata e si può concludere che i dati non rivelano l'esistenza di una differenza significativa fra le varianze delle due popolazioni.

b – H_0: $\sigma_1^2 \geq \sigma_2^2$

H_1: $\sigma_1^2 < \sigma_2^2$

Il valore critico per il test è

$$F_{1-\alpha}(\nu_1, \nu_2) = F_{0.95}(15,12) = \frac{1}{F_{0.05}(12,15)} = \frac{1}{2.48} = 0.403$$

La regione di rifiuto è costituita dai valori $F < 0.403$, perciò l'ipotesi nulla non deve essere rifiutata.

c – H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

I valori critici per il test sono

$$F_{\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.025}(15,12) = 3.18$$

$$F_{1-\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.975}(15,12) = \frac{1}{F_{0.025}(12,15)} = \frac{1}{2.96} = 0.34$$

La regione di rifiuto è costituita dai valori $F < 0.34$ e dai valori $F > 3.18$, perciò l'ipotesi nulla non deve essere rifiutata.

Livello di significatività $\alpha = 0.01$

a – H_0: $\sigma_1^2 \leq \sigma_2^2$

H_1: $\sigma_1^2 > \sigma_2^2$

Il valore critico per il test è

$$F_\alpha(\nu_1, \nu_2) = F_{0.01}(15,12) = 4.01$$

La regione di rifiuto è costituita dai valori $F > 4.01$, perciò l'ipotesi nulla non deve essere rifiutata: i dati non rivelano l'esistenza di una differenza significativa fra le varianze delle due popolazioni.

b – H_0: $\sigma_1^2 \geq \sigma_2^2$

H_1: $\sigma_1^2 < \sigma_2^2$

Il valore critico per il test è

$$F_{1-\alpha}(\nu_1,\nu_2) = F_{0.99}(15,12) = \frac{1}{F_{0.01}(12,15)} = \frac{1}{3.67} = 0.27$$

La regione di rifiuto è costituita dai valori $F < 0.27$, perciò l'ipotesi nulla non deve essere rifiutata.

c – H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

I valori critici per il test sono

$$F_{\frac{\alpha}{2}}(\nu_1,\nu_2) = F_{0.005}(15,12) = 4.72$$

$$F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2) = F_{0.995}(15,12) = \frac{1}{F_{0.005}(12,15)} = \frac{1}{4.25} = 0.24$$

La regione di rifiuto è costituita dai valori $F < 0.24$ e dai valori $F > 4.72$, perciò l'ipotesi nulla non deve essere rifiutata.

Esempio 42

Nella tabella sono riportate le lunghezze in cm di due campioni A e B di oggetti dello stesso tipo prodotti da due macchine diverse.

A	8.26	8.13	8.35	8.07	8.34		
B	7.95	7.89	7.90	8.14	7.92	7.84	7.94

Per questi dati è stato calcolato un intervallo di confidenza per la differenza fra le medie, assumendo che le due popolazioni da cui provengono i campioni abbiano distribuzione normale con la stessa varianza (esempio 19, §7.7, pag. 208).

Sottoporre a test questa assunzione con livello di significatività $\alpha = 0.05$.

Per verificare se è ragionevole assumere che le varianze delle due popolazioni sono uguali, scegliamo come ipotesi nulla e come ipotesi alternativa

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

In base ai dati della tabella si ha

$$n_1 = 5 \qquad s_1^2 = 0.01575 \qquad n_2 = 7 \qquad s_2^2 = 0.00910$$

Il valore della statistica F è

$$F = \frac{0.01575}{0.00910} = 1.73$$

Si effettua un test a due code e i valori critici per il test sono

$$F_{\frac{\alpha}{2}}(\nu_1,\nu_2) = F_{0.025}(4,6) = 6.23$$

$$F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2) = F_{0.975}(4,6) = \frac{1}{F_{0.025}(6,4)} = \frac{1}{9.20} = 0.11$$

La regione di rifiuto è costituita dai valori $F < 0.11$ e dai valori $F > 6.23$, perciò l'ipotesi nulla non deve essere rifiutata: non c'è una ragione significativa per dubitare che le due varianze siano uguali.

Esempio 43
Consideriamo i dati dell'esempio 31. Per poter effettuare il test di ipotesi sulla differenza fra le medie, abbiamo ipotizzato che le due popolazioni abbiano la stessa varianza; effettuiamo il test sul rapporto delle varianze per stabilire se questa ipotesi è verificata.

Per verificare se è ragionevole assumere che le varianze delle due popolazioni sono uguali, scegliamo come ipotesi nulla e come ipotesi alternativa

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$
$$H_1: \quad \sigma_1^2 \neq \sigma_2^2$$

In base ai dati si ha

$$n_1 = 6 \qquad s_1 = 0.032 \qquad n_2 = 5 \qquad s_2 = 0.024$$

Il valore della statistica F è

$$F = \frac{0.032^2}{0.024^2} = 1.78$$

Si effettua un test a due code e i valori critici per il test sono

$$F_{\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.025}(5,4) = 9.36$$

$$F_{1-\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.975}(5,4) = \frac{1}{F_{0.025}(4,5)} = \frac{1}{7.39} = 0.14$$

La regione di rifiuto è costituita dai valori $F < 0.14$ e dai valori $F > 9.36$, perciò l'ipotesi nulla non deve essere rifiutata e non c'è una ragione significativa per dubitare che le due varianze siano uguali.

Esempio 44
Consideriamo i dati dell'esempio 32. Per poter effettuare il test di ipotesi sulla differenza fra le medie, abbiamo ipotizzato che le due popolazioni abbiano la stessa varianza; effettuiamo il test sul rapporto delle varianze per stabilire se questa ipotesi è verificata.

Per verificare se è ragionevole assumere che le varianze delle due popolazioni sono uguali, scegliamo come ipotesi nulla e come ipotesi alternativa

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$
$$H_1: \quad \sigma_1^2 \neq \sigma_2^2$$

In base ai dati si ha

$$n_1 = 25 \qquad s_1 = 21.4 \qquad n_2 = 25 \qquad s_2 = 19.3$$

Il valore della statistica F è

$$F = \frac{21.4^2}{19.3^2} = 1.23$$

Si effettua un test a due code e i valori critici per il test sono

$$F_{\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.025}(24,24) = 2.27$$

$$F_{1-\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.975}(24,24) = \frac{1}{F_{0.025}(26,24)} = \frac{1}{2.27} = 0.44$$

La regione di rifiuto è costituita dai valori $F < 0.44$ e dai valori $F > 2.27$: l'ipotesi nulla non deve essere rifiutata e non c'è una ragione significativa per dubitare che le due varianze siano uguali.

Esempio 45
Due macchine diverse producono filo metallico che deve avere diametro costante. Per controllare la qualità del processo, vengono eseguite misure del diametro in punti casuali diversi del filo prodotto dalle due macchine; il campione di 16 misure effettuate sulla prima macchina ha varianza $s_1^2 = 0.00385$, mentre il campione di 25 misure effettuate sulla seconda macchina ha varianza $s_2^2 = 0.00125$. Si può sostenere che le due macchine siano ugualmente accurate? In caso contrario la seconda macchina è più accurata della prima (ossia per la seconda c'è una minor variabilità)?

a – Per rispondere alla prima domanda, si effettua un test sull'uguaglianza delle due varianze; per verificare se è ragionevole assumere che le varianze delle due popolazioni sono uguali, scegliamo come ipotesi nulla e come ipotesi alternativa

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$
$$H_1: \quad \sigma_1^2 \neq \sigma_2^2$$

In base ai dati si ha

$$n_1 = 16 \qquad s_1^2 = 0.00385 \qquad n_2 = 25 \qquad s_2^2 = 0.00125$$

Il valore della statistica F è

$$F = \frac{0.00385}{0.00125} = 3.08$$

Si effettua un test a due code e i valori critici per il test al livello di significatività del 5% sono

$$F_{\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.025}(15,24) = 2.44$$

$$F_{1-\frac{\alpha}{2}}(\nu_1, \nu_2) = F_{0.975}(15,24) = \frac{1}{F_{0.025}(24,15)} = \frac{1}{2.70} = 0.37$$

La regione di rifiuto è costituita dai valori $F < 0.37$ e dai valori $F > 2.44$, perciò l'ipotesi nulla deve essere rifiutata e c'è una ragione significativa per dubitare che le due varianze siano uguali.
b – Per rispondere alla seconda domanda, si deve effettuare il test scegliendo come ipotesi

$$H_0: \quad \sigma_1^2 \leq \sigma_2^2$$
$$H_1: \quad \sigma_1^2 > \sigma_2^2$$

Si effettua un test a una coda e il valore critico per il test al livello di significatività del 5% è

$$F_{\alpha}(\nu_1, \nu_2) = F_{0.05}(15,24) = 2.11$$

La regione di rifiuto è costituita dai valori $F > 2.11$, perciò l'ipotesi nulla deve essere rifiutata e c'è una ragione significativa per affermare che la seconda macchina è più accurata della prima.

Esempio 46
In una scuola elementare è stato fatto un esame di grammatica. Il voto medio dei 25 bambini è stato di 72/100, con uno scarto quadratico medio $s_1 = 8$, mentre il voto medio delle 25 bambine è stato di 78/100 con uno scarto quadratico medio $s_2 = 6$.
Provare al livello di significatività del 5% l'ipotesi che le bambine siano migliori dei bambini in grammatica.

Per poter effettuare il test di ipotesi sulla differenza fra le medie occorre ipotizzare che le due popolazioni da cui sono estratti i campioni abbiano distribuzione normale con la stessa varianza.
Si effettua dapprima un test sull'uguaglianza delle due varianze, per verificare se è ragionevole assumere che le varianze delle due popolazioni sono uguali; scegliamo come ipotesi

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$
$$H_1: \quad \sigma_1^2 \neq \sigma_2^2$$

In base ai dati si ha

$$n_1 = 25 \qquad s_1^2 = 64 \qquad n_2 = 25 \qquad s_2^2 = 36$$

Il valore della statistica F è

$$F = \frac{64}{36} = 1.78$$

Si effettua un test a due code e i valori critici per il test al livello di significatività del 5% sono

$$F_{\frac{\alpha}{2}}(\nu_1,\nu_2) = F_{0.025}(24,24) = 2.27$$

$$F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2) = F_{0.975}(24,24) = \frac{1}{F_{0.025}(24,24)} = \frac{1}{2.27} = 0.44$$

La regione di rifiuto è costituita dai valori $F < 0.44$ e dai valori $F > 2.27$, perciò l'ipotesi nulla non può essere rifiutata e non c'è una ragione significativa per dubitare che le due varianze siano uguali.

Si può ora effettuare il test sulla differenza fra le medie.

Si assume come ipotesi nulla e come ipotesi alternativa

H_0: $\mu_1 \geq \mu_2$

H_1: $\mu_1 < \mu_2$.

In base ai dati della tabella si ha

$$n_1 = 25 \qquad \bar{x}_1 = 72 \qquad s_1^2 = 64$$

$$n_2 = 25 \qquad \bar{x}_2 = 78 \qquad s_2^2 = 36$$

La stima congiunta della varianza con la formula (8.1) è

$$S^2 = \frac{24 \cdot 64 + 24 \cdot 36}{48} = 50$$

Il grado di libertà della distribuzione t è

$$\nu = n_1 + n_2 - 2 = 48$$

La statistica t ha il valore

$$T = \frac{72 - 78}{\sqrt{50 \cdot \left(\frac{1}{25} + \frac{1}{25}\right)}} = -3$$

Per il livello di significatività $\alpha = 0.05$ il valore critico è $t_{0.025} = 1.96$ e la regione di rifiuto è data dai valori $T < -1.96$ e $T > 1.96$. Il valore $T = -3$ appartiene alla regione di rifiuto, perciò l'ipotesi nulla deve essere rifiutata e si conclude che al livello di significatività $\alpha = 0.05$ il voto medio delle bambine è superiore a quello dei bambini.

9. Test chi-quadro

9.1 Introduzione

Nei capitoli sulla stima e sulla verifica delle ipotesi abbiamo usato la distribuzione χ^2 per la costruzione di intervalli di confidenza e per il test di ipotesi per la varianza di una popolazione.
Questa distribuzione ha numerose altre applicazioni nella statistica; in particolare ne faremo uso nella verifica di ipotesi con dati disponibili sotto forma di frequenze.
Queste procedure di verifica delle ipotesi sono note come **test di adattamento** (o anche **goodness of fit**) e **test di indipendenza**.
In qualche modo entrambi i test chi-quadro che esaminiamo possono essere pensati come dei test sulla bontà dell'adattamento, nel senso che studiano la bontà dell'adattamento delle frequenze osservate rispetto a delle frequenze che si presume dovrebbero verificarsi, se i dati fossero generati da una qualche teoria o ipotesi.
Tuttavia il termine "bontà dell'adattamento" viene di solito usato in senso stretto, per riferirsi al confronto tra la distribuzione osservata su un campione e la distribuzione teorica che si ipotizza possa descrivere la popolazione da cui proviene il campione.

9.2 Test chi-quadro di adattamento

In questo paragrafo ci occupiamo di un metodo statistico utile per stabilire se un campione di dati osservati si adatta a una distribuzione teorica assegnata; ad esempio, potrebbe esserci motivo di credere che il numero di incidenti che si verificano in un certo periodo di tempo in un tratto di strada sia una variabile aleatoria avente distribuzione di Poisson: questa convinzione può essere verificata osservando per un certo periodo il numero di incidenti, ed eseguendo quindi un test che sia in grado di stabilire con un certo grado di fiducia se la popolazione possa avere la distribuzione ipotizzata.
I test statistici che servono a verificare se una certa distribuzione è compatibile con i dati del campione sono detti **test sulla bontà di adattamento**.
Per effettuare il test supponiamo di avere un campione di n osservazioni di una variabile, raggruppate in una tabella contenente k classi.
Le classi possono rappresentare:
– caratteristiche qualitative;
– valori assunti da una variabile discreta: ogni classe raggruppa tutte le osservazioni che assumono un dato valore, eventualmente una o due classi raggruppano le code;
– intervalli di valori assunti da una variabile continua.
In altri termini, la tabella rappresenta la distribuzione di frequenza assoluta di una variabile qualitativa o di una variabile numerica discreta o continua.
Per ciascuna classe supponiamo di avere, oltre alla **frequenza osservata** O_i , una **frequenza attesa** A_i , con cui si vuole confrontare la frequenza osservata; le frequenze attese sono quelle che si osserverebbero se i dati del campione fossero distribuiti esattamente secondo la distribuzione ipotizzata.
Per valutare quantitativamente la bontà dell'adattamento delle frequenze osservate alle frequenze attese si utilizza la **statistica test**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i} \tag{9.1}$$

che viene detta il **chi-quadro** calcolato dal campione.
Si dimostra che, per n sufficientemente grande, questa statistica ha approssimativamente la distribuzione χ^2, con grado di libertà $\nu = k - 1 - m$, dove m è il numero dei parametri della distribuzione teorica stimati servendosi dei dati del campione.

Se l'ipotesi nulla H_0 è che i dati si adattino alla distribuzione teorica ipotizzata, la regola di decisione sarà: si rifiuti l'ipotesi nulla se il valore della statistica χ^2 calcolato dai dati è maggiore del valore critico χ^2_α

$$\chi^2 > \chi^2_\alpha \tag{9.2}$$

α è il livello di significatività stabilito e il grado di libertà della distribuzione χ^2 è $\nu = k - 1 - m$; k indica il numero delle classi e m il numero dei parametri della distribuzione teorica stimati servendosi dei dati del campione.
Questa proceduta, detta **test chi-quadro di adattamento**, è valida purché le frequenze assolute attese siano tutte maggiori o uguali a 5. Questa condizione garantisce che la distribuzione della statistica χ^2 sia ben approssimata dalla distribuzione chi-quadro; quando, dopo aver calcolato le frequenze attese, si osserva che qualcuna di queste è minore di 5, bisogna accorpare opportunamente due o più classi contigue, in modo che la condizione sia verificata. Si ricordi che, dopo aver accorpato le classi, il numero di classi da considerare per calcolare il grado di libertà della distribuzione chi-quadro è quello ridotto e non quello originale.
Negli esempi seguenti viene illustrato il test di adattamento; in particolare in alcuni esempi esamineremo il caso della distribuzione binomiale, della distribuzione di Poisson e della distribuzione normale.

Esempio 1
Alle ultime elezioni amministrative in un comune si sono presentate quattro liste che hanno ottenuto le seguenti percentuali

Lista	1	2	3	4	*Totale*
Percentuale	26%	32%	15%	27%	100%

Nella sezione elettorale A del comune, su 350 voti validi, i voti sono risultati così suddivisi

Lista	1	2	3	4	*Totale*
Voti	80	120	60	90	350

Nella sezione elettorale B invece, su 320 voti validi, i voti sono risultati così suddivisi

Lista	1	2	3	4	*Totale*
Voti	65	120	40	95	320

Si può ritenere che i risultati elettorali delle due sezioni si adattino bene ai risultati complessivi, oppure le differenze sono statisticamente rilevanti?

– Sezione elettorale A
Costruiamo innanzi tutto una tabella contenente le frequenze osservate, ossia i voti della sezione, e le frequenze attese, ossia quelle che si osserverebbero, sui 350 voti della sezione, se questi voti fossero distribuiti esattamente secondo le percentuali di tutto l'elettorato; per ottenere le frequenze attese si trasforma ogni frequenza attesa in frequenza relativa attesa e poi in frequenza assoluta attesa. Ad esempio per la lista 1:
Percentuale attesa: 26%
Frequenza relativa attesa: 0.26
Frequenza assoluta attesa: $0.26 \cdot 350 = 91$

Si ottiene la seguente tabella 1

Lista	*Frequenze osservate* O_i	*Frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
1	80	91	$\frac{(80-91)^2}{91} = 1.33$
2	120	112	$\frac{(120-112)^2}{112} = 0.57$
3	60	52.5	$\frac{(60-52.5)^2}{52.5} = 1.07$
4	90	94.5	$\frac{(90-94.5)^2}{94.5} = 0.21$
Totale	350	350	3.18

Tabella 1

Per il calcolo del valore della statistica χ^2 con la (9.1) è utile aggiungere l'ultima colonna della precedente tabella: in tale colonna sono riportati i singoli addendi della sommatoria; per ottenere il valore di χ^2 basta sommare i valori della colonna, quindi il valore della statistica chi-quadro calcolato dal campione della sezione A è

$$\chi^2 = 1.33 + 0.57 + 1.07 + 0.21 = 3.18$$

Le classi sono 4 e nessun parametro è stato stimato dai dati del campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 4 - 1 - 0 = 3$$

Al livello di significatività del 5%, sulle tavole della distribuzione χ^2 si legge il valore critico

$$\chi^2_{0.05} = 7.815$$

L'ipotesi nulla è che i risultati della sezione A si adattino alla distribuzione complessiva dei voti; il test prevede che l'ipotesi nulla venga rifiutata se il valore della statistica chi-quadro è maggiore del valore critico. Nel nostro caso per la sezione A il valore è minore, perciò l'ipotesi nulla non può essere rifiutata e concludiamo che non c'è una differenza statisticamente rilevante fra i dati di questa sezione e i risultati complessivi.

– Sezione elettorale B

Ripetiamo tutto il calcolo con i dati della sezione B; la tabella delle frequenze osservate e delle frequenze attese è la seguente

Lista	*Frequenze osservate* O_i	*Frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
1	65	83.2	$\frac{(65-83.2)^2}{83.2} = 3.98$
2	120	102.4	$\frac{(120-102.4)^2}{102.4} = 3.03$
3	40	48	$\frac{(40-48)^2}{48} = 1.33$
4	95	86.4	$\frac{(95-86.4)^2}{86.4} = 0.86$
Totale	320	320	9.20

Tabella 2

Il valore della statistica χ^2 calcolato dai dati del campione della sezione B è

$$\chi^2 = 3.98 + 3.03 + 1.33 + 0.86 = 9.20$$

Poiché il valore del chi-quadro è maggiore del valore critico, l'ipotesi nulla viene rifiutata e concludiamo che i risultati della sezione B non sono rappresentativi dei risultati complessivi, ossia c'è una differenza statisticamente rilevante.

Esempio 2

Si effettuano 120 lanci di un dado e si osservano le seguenti uscite

N° uscito	1	2	3	4	5	6
Frequenza	25	17	15	23	24	16

Provare l'ipotesi che il dado non sia truccato, usando il livello di significatività del 5% (adattamento alla distribuzione uniforme discreta).

Se il dado non è truccato, le frequenze attese sono tutte uguali a 20 e si costruisce la tabella 3

N° uscito	*Frequenze osservate* O_i	*Frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
1	25	20	1.25
2	17	20	0.45
3	15	20	1.25
4	23	20	0.45
5	24	20	0.80
6	16	20	0.80
Totale	120	120	5.00

Tabella 3

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 5.00$$

Le classi sono 6 e nessun parametro è stato stimato dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 6 - 1 = 5$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 11.070$$

Il test è: si rifiuta l'ipotesi che il dado sia buono se il valore del chi-quadro è maggiore del valore critico; nel nostro caso è minore, perciò in base ai dati non possiamo rifiutare l'ipotesi, e concludiamo che, al livello di significatività del 5%, non c'è una significativa evidenza che il dado sia truccato.

Esempio 3

Una tabella di 250 numeri casuali di una cifra mostra la seguente distribuzione dei numeri da 0 a 9. La distribuzione osservata differisce significativamente dalla distribuzione attesa?

Numeri	0	1	2	3	4	5	6	7	8	9
Frequenze osservate	17	31	29	18	14	20	35	30	20	36

La tabella delle frequenze osservate e delle frequenze attese è la seguente tabella 4

Numero	*Frequenze osservate* O_i	*Frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
0	17	25	2.56
1	31	25	1.44
2	29	25	0.64
3	18	25	1.96
4	14	25	4.84
5	20	25	1.00
6	35	25	4.00
7	30	25	1.00
8	20	25	1.00
9	36	25	4.84
Totale	250	250	23.28

Tabella 4

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 23.28$$

Le classi sono 10 e nessun parametro è stato stimato dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 10 - 1 = 9$$

I valori critici ai livelli di significatività del 5% e dell'1% sono rispettivamente

$$\chi^2_{0.05} = 16.919 \qquad \chi^2_{0.01} = 21.666$$

Il test è: si rifiuta l'ipotesi che non vi sia differenza significativa dalla distribuzione attesa, se il valore del chi-quadro è maggiore del valore critico; nel nostro caso per entrambi i livelli di significatività è maggiore, perciò, in base ai dati, rifiutiamo l'ipotesi e concludiamo che la distribuzione osservata differisce significativamente dalla distribuzione attesa: la tabella dei numeri casuali deve essere giudicata con diffidenza!

Esempio 4

In base a una ricerca condotta in anni precedenti, si può ritenere che il numero di incidenti stradali per settimana, in un certo tratto di autostrada, segua la distribuzione di Poisson di parametro $\lambda = 0.4$. Nelle ultime 90 settimane si sono rilevati i seguenti dati

N° di incidenti per settimana	0	1	2	3 o più	*Totale*
N° di settimane in cui si è verificato	52	32	6	0	90

Possiamo affermare che il modello è ancora applicabile alla descrizione del fenomeno, oppure qualcosa è cambiato?

La distribuzione teorica con cui si vogliono confrontare i dati è la distribuzione di Poisson di parametro (valor medio) $\lambda = 0.4$. Usando questa distribuzione possiamo calcolare le seguenti probabilità[1]

$$P(X = 0) = e^{-0.4} = 0.6703$$

$$P(X = 1) = e^{-0.4} \cdot 0.4 = 0.2681$$

$$P(X = 2) = e^{-0.4} \cdot \frac{0.4^2}{2} = 0.0536$$

$$P(X \geq 3) = 1 - P(X < 3) = 1 - (0.6703 + 0.2681 + 0.0536) = 0.0080$$

Queste sono le probabilità con cui *X* appartiene alle quattro classi, ossia le frequenze relative attese;

[1] Queste probabilità possono anche essere ottenute usando le tavole della distribuzione di Poisson

le frequenze assolute attese si ottengono moltiplicando le frequenze relative attese per il numero di osservazioni, in questo caso 90. Si ottiene la tabella 5

Classe (N° incidenti per settimana)	*Frequenza relativa attesa*	*Frequenza assoluta attesa*	*Frequenza assoluta osservata*
$X = 0$	0.6703	60.33	52
$X = 1$	0.2681	24.13	32
$X = 2$	0.0536	4.82	6
$X \geq 3$	0.0080	0.72	0
Totale	1	90	90

Tabella 5

Si osserva che le ultime due classi hanno frequenze assolute attese minori di 5, perciò non possiamo usare questa tabella per effettuare il test; si accorpano allora le ultime due classi in un'unica classe con frequenza assoluta attesa pari a $4.82 + 0.72 = 5.54$ e frequenza assoluta osservata pari a $6 + 0 = 6$.
Otteniamo così la tabella 6, nella quale l'ultima colonna contiene gli addendi della sommatoria che definisce la statistica chi-quadro

Classe (N° incidenti per settimana)	*Frequenza assoluta attesa* A_i	*Frequenza assoluta osservata* O_i	$\frac{(O_i - A_i)^2}{A_i}$
$X = 0$	60.33	52	1.15
$X = 1$	24.13	32	2.57
$X \geq 2$	5.54	6	0.04
Totale	90	90	3.76

Tabella 6

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 1.15 + 2.57 + 0.04 = 3.76$$

Le classi, dopo l'accorpamento sono 3 e nessun parametro è stato stimato dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 3 - 1 = 2$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 5.991$$

Il test è: si rifiuta l'ipotesi di adattamento se il valore del chi-quadro è maggiore del valore critico; nel nostro caso è minore, perciò, in base ai dati, non possiamo rifiutare l'ipotesi nulla di adattamento, e concludiamo che, al livello di significatività del 5%, in base ai dati del campione non c'è evidenza statistica del fatto che la legge seguita dal numero settimanale di incidenti sia cambiata.

Esempio 5
Durante 400 intervalli di 5 minuti alla torre di controllo di un aeroporto arrivano 0, 1, 2,…,13 messaggi radio con le rispettive frequenze 3, 15, 47,…,1. I dati di questo campione sono raccolti nella tabella seguente

N° messaggi radio	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequenze osservate	3	15	47	76	68	74	46	39	15	9	5	2	0	1

Si vuole sottoporre a test l'ipotesi che questi dati confermino l'affermazione che il numero di messaggi radio che si ricevono in un intervallo di 5 minuti sia una variabile aleatoria avente la distribuzione di Poisson di parametro $\lambda = 4.6$.

Le frequenze relative attese possono essere ottenute usando la tavola della distribuzione di Poisson con parametro $\lambda = 4.6$; le corrispondenti frequenza assolute attese si ottengono moltiplicando le frequenze relative per 400; questi valori sono raccolti nella tabella 7, insieme con le frequenze osservate dai dati del campione

N° messaggi radio	*Frequenze osservate* O_i	*Frequenze relative attese (tavola Poisson)*	*Frequenze assolute attese* A_i
0	3	0.0101	4.04
1	15	0.0462	18.48
2	47	0.1063	42.52
3	76	0.1631	65.24
4	68	0.1875	75.00
5	74	0.1726	69.04
6	46	0.1322	52.88
7	39	0.0869	34.76
8	15	0.0500	20.00
9	9	0.0256	10.24
10	5	0.0117	4.68
11	2	0.0049	1.96
12	0	0.0019	0.76
13	1	0.0010	0.40
Totale	400	1	400

Tabella 7

Poiché ci sono delle classi che hanno frequenze assolute attese minori di 5, procediamo ad accorpare alcune classi e otteniamo la tabella 8.
Ricordiamo che solo le frequenze attese non devono essere minori di 5, e non quelle osservate.

N° messaggi radio	*Frequenze osservate* O_i	*Frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
0 e 1	18	22.52	0.91
2	47	42.52	0.47
3	76	65.24	1.77
4	68	75.00	0.65
5	74	69.04	0.36
6	46	52.88	0.90
7	39	34.76	0.52
8	15	20.00	1.25
9	9	10.24	0.15
10	8	7.80	0.01
Totale	400	400	6.99

Tabella 8

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 6.99$$

Le classi, dopo l'accorpamento sono 10 e nessun parametro è stato stimato dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 10 - 1 = 9$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 16.919$$

Il test è: si rifiuta l'ipotesi di adattamento se il valore del chi-quadro è maggiore del valore critico; nel nostro caso è minore, perciò, in base ai dati, non possiamo rifiutare l'ipotesi nulla di adattamento, e concludiamo che, al livello di significatività del 5%, c'è un buon adattamento dei dati alla distribuzione di Poisson con parametro $\lambda = 4.6$.

Esempio 6

Si ipotizza che il numero di difetti presenti in un circuito elettronico stampato segua una distribuzione di Poisson. In un campione casuale di 60 circuiti è stato osservato il numero di difetti presenti, ottenendo i seguenti dati

Numero di difetti	*Frequenza osservata*
0	32
1	15
2	9
3	4

I dati si adattano alla distribuzione ipotizzata?

Il valor medio della distribuzione di Poisson è incognito e deve essere calcolato dai dati.

$$\lambda = \frac{0 \cdot 32 + 1 \cdot 15 + 2 \cdot 9 + 3 \cdot 4}{60} = 0.75$$

Le frequenze attese possono essere calcolate con la distribuzione di Poisson di parametro $\lambda = 0.75$

$$P(X = 0) = e^{-0.75} = 0.4724$$

$$P(X = 1) = e^{-0.75} \cdot 0.75 = 0.3543$$

$$P(X = 2) = e^{-0.75} \cdot \frac{0.75^2}{2} = 0.1329$$

$$P(X \geq 3) = 1 - P(X < 3) = 1 - (0.4724 + 0.3543 + 0.1329) = 0.0404$$

Le corrispondenti frequenza assolute attese si ottengono moltiplicando le frequenze relative per 60; questi valori sono raccolti nella tabella 9, insieme con le frequenze osservate dai dati del campione

Numero di difetti	*Frequenze osservate* O_i	*Frequenze relative attese (distrib. Poisson)*	*Frequenze assolute attese* A_i
0	32	0.4724	28.34
1	15	0.3543	21.26
2	9	0.1329	7.97
≥ 3	4	0.0404	2.42

Tabella 9

Poiché l'ultima classe ha una frequenza assoluta attesa minore di 5, accorpiamo le ultime due classi e otteniamo la tabella 10

Numero di difetti	*Frequenze osservate* O_i	*Frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
0	32	28.34	0.47
1	15	21.26	1.84
≥ 2	13	10.39	0.66

Tabella 10

Il valore della statistica chi-quadro calcolato dai dati del campione è

$$\chi^2 = 0.47 + 1.84 + 0.66 = 2.97$$

Le classi dopo l'accorpamento sono 3 e il valor medio della distribuzione è stato stimato dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 3 - 1 - 1 = 1$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 3.841$$

L'ipotesi nulla è che i dati si adattino alla distribuzione di Poisson di parametro λ=0.75; dato che il valore della statistica calcolato dal campione è minore del valore critico, non possiamo rifiutare l'ipotesi nulla.

Esempio 7

Cinque monete sono state lanciate 1000 volte, e a ciascun lancio è stato osservato il numero di teste; nella tabella è riportato il numero di lanci nei quali sono state ottenute 0, 1, …, 5 teste.

N° teste	0	1	2	3	4	5
Frequenza osservata	38	144	342	287	164	25

Stabilire se le monete si possono ritenere non truccate.

Se le monete sono eque, il numero di teste su 5 monete in un singolo lancio ha una distribuzione binomiale di parametri $p = \frac{1}{2}$, n = 5. Le probabilità di avere 0, 1, …, 5 teste si possono ottenere dalla tavola della distribuzione binomiale; le corrispondenti frequenze assolute attese si ricavano moltiplicando per 1000 tali probabilità. Si ottiene così la tabella 11

Numero di teste	*Frequenze osservate* O_i	*Frequenze relative attese (binomiale)*	*Frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
0	38	0.0313	31.25	1.46
1	144	0.1562	156.25	0.96
2	342	0.3125	312.50	2.78
3	287	0.3125	312.50	2.08
4	164	0.1562	156.25	0.38
5	25	0.0313	31.25	1.25
Totale	1000	1	1000	8.91

Tabella 11

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 8.91$$

Le classi sono 5 e nessun parametro è stato stimato dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 5$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 11.070$$

Il valore del chi-quadro è minore del valore critico, perciò in base ai dati non possiamo rifiutare l'ipotesi nulla di adattamento, e concludiamo che, al livello di significatività del 5%, c'è un buon adattamento dei dati alla distribuzione binomiale; in altri termini non possiamo rifiutare l'ipotesi che la moneta sia equa.

Ci sono molte procedure statistiche che richiedono come ipotesi il fatto che la popolazione abbia la distribuzione normale: ad esempio, quando si effettua un test di ipotesi sulla media nel caso dei piccoli campioni, si richiede che la popolazione da cui si estrae il campione sia normale.
In queste situazioni il test chi-quadro di adattamento è uno strumento utile per verificare se queste procedure sono applicabili.
I seguenti esempi illustrano l'applicazione del test chi-quadro per l'adattamento alla distribuzione normale.

Esempio 8
La tabella 12 fornisce la distribuzione della pressione sanguigna sistolica (in mm di mercurio) per un campione casuale di 250 uomini di età fra i 30 e i 40 anni.
Stabilire al livello di significatività del 5% se i dati del campione si adattano a una distribuzione normale.

Pressione	*Frequenza osservata (n° di uomini)*
$80 < x \le 100$	3
$100 < x \le 110$	12
$110 < x \le 120$	52
$120 < x \le 130$	74
$130 < x \le 140$	67
$140 < x \le 150$	26
$150 < x \le 160$	12
$160 < x \le 180$	4

Tabella 12

L'ipotesi nulla è che la pressione sanguigna abbia una distribuzione normale. Per sottoporre a test questa ipotesi occorre calcolare dai dati la stima per la media e la varianza della popolazione.
Disponiamo i calcoli nella tabella 13

Pressione	*Valore centrale* x_i	*Frequenza osservata* f_i	$f_i x_i$	$f_i x_i^2$
$80 < x \le 100$	90	3	270	24300
$100 < x \le 110$	105	12	1260	132300
$110 < x \le 120$	115	52	5980	687700
$120 < x \le 130$	125	74	9250	1156250
$130 < x \le 140$	135	67	9045	1221075
$140 < x \le 150$	145	26	3770	546650
$150 < x \le 160$	155	12	1860	288300
$160 < x \le 180$	170	4	680	115600
Totale		250	32115	4172175

Tabella 13

Il valor medio, la varianza e lo scarto quadratico medio (dati raggruppati) sono

$$\bar{x} = \frac{32115}{250} = 128.46$$

$$s^2 = \frac{1}{249}\left[4172175 - \frac{(32115)^2}{250}\right] = 187.4783$$

$$s = 13.69$$

Calcoliamo ora le frequenze relative attese delle classi. Oltre alle classi indicate, nelle quali cadono le osservazioni, occorre considerare anche la classe $x \le 80$ e la classe $x \ge 180$; queste due classi hanno frequenze osservate nulle, ma le frequenze relative attese non sono nulle.

Per calcolare queste frequenze usiamo la funzione di ripartizione della variabile aleatoria normale X di media $\mu = 128.46$ e scarto quadratico medio $\sigma = 13.69$; passando alla variabile aleatoria standardizzata

$$Z = \frac{X - 128.46}{13.69}$$

e servendosi delle tavole della distribuzione normale standardizzata, si costruisce la tabella 14, contenente le frequenze relative attese e le frequenze assolute attese.
Ad esempio, per la classe $110 < x \le 120$ la frequenza relativa attesa (probabilità) si calcola nel modo seguente

$$X = 110 \quad \Rightarrow \quad Z = \frac{110 - 128.46}{13.69} = -1.35$$

$$X = 120 \quad \Rightarrow \quad Z = \frac{120 - 128.46}{13.69} = -0.62$$

$$P(110 < X < 120) = P(-1.35 < Z < -0.62) = P(0.62 < Z < 1.35) =$$
$$= 0.9115 - 0.7324 = 0.1791$$

La corrispondente frequenza attesa è quindi $250 \cdot 0.1791 = 44.775$

Classi	*Frequenza osservata* f_i	*frequenze relative attese (probabilità)*	*frequenze attese*
$x \le 80$	0	0.0002	0.050
$80 < x \le 100$	3	0.0186	4.650
$100 < x \le 110$	12	0.0697	17.425
$110 < x \le 120$	52	0.1791	44.775
$120 < x \le 130$	74	0.2762	69.050
$130 < x \le 140$	67	0.2557	63.925
$140 < x \le 150$	26	0.1423	35.575
$150 < x \le 160$	12	0.0475	11.875
$160 < x \le 180$	4	0.0106	2.650
$x > 180$	0	0.0001	0.025
Totale	250	1	250

Tabella 14

Poiché ci sono delle classi che hanno frequenze assolute attese minori di 5, procediamo ad accorpare alcune classi e otteniamo la tabella 15.

Classi	*Frequenza osservata* O_i	*frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
$x \le 110$	15	22.125	2.294
$110 < x \le 120$	52	44.775	1.166
$120 < x \le 130$	74	69.050	0.355
$130 < x \le 140$	67	63.925	0.148
$140 < x \le 150$	26	35.575	2.577
$x > 150$	16	14.550	0.145
Totale	250	250	6.685

Tabella 15

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 6.685$$

Le classi sono 6 e due parametri, valor medio e varianza della popolazione, sono stati stimati dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 6 - 1 - 2 = 3$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 7.815$$

Il valore del chi-quadro è minore del valore critico, perciò in base ai dati non possiamo rifiutare l'ipotesi nulla di adattamento, e concludiamo che, al livello di significatività del 5%, c'è un buon adattamento dei dati alla distribuzione normale.

Esempio 9

Sono state misurate le lunghezze di 150 sbarrette di metallo simili, e i dati sono stati raggruppati nella tabella seguente

Classi (lunghezza in mm)	*Frequenze osservate*
$27 < x \le 28$	3
$28 < x \le 29$	23
$29 < x \le 30$	53
$30 < x \le 31$	50
$31 < x \le 32$	21

Stabilire se in base a questi dati si può affermare che la lunghezza delle sbarrette segue una distribuzione normale.

L'ipotesi nulla è che la lunghezza abbia una distribuzione normale. Per sottoporre a test questa ipotesi occorre calcolare dai dati la stima per la media e la varianza della popolazione.
Usando i dati raggruppati si ottiene

$$\bar{x} = 29.92$$

$$s^2 = 0.9566 \qquad s = 0.978$$

Calcoliamo ora le frequenze relative attese delle classi. Oltre alle classi indicate, nelle quali cadono le osservazioni, occorre considerare anche la classe $x \le 27$ e la classe $x > 32$.
Per calcolare queste frequenze usiamo la funzione di ripartizione della variabile aleatoria normale X di media $\mu = 29.92$ e scarto quadratico medio $\sigma = 0.978$; passando alla variabile aleatoria standardizzata

$$Z = \frac{X - 29.92}{0.978}$$

e servendosi delle tavole della distribuzione normale standardizzata, si costruisce la tabella 16, contenente le frequenze relative attese e le frequenze assolute attese.
Ad esempio per la classe $27 < x \le 28$ la frequenza relativa attesa (probabilità) si calcola nel modo seguente

$$X = 27 \quad \Rightarrow \quad Z = \frac{27 - 29.92}{0.978} = -2.99$$

$$X = 28 \quad \Rightarrow \quad Z = \frac{28 - 29.92}{0.978} = -1.96$$

$$P(27 < X < 28) = P(-2.99 < Z < -1.96) = P(1.96 < Z < 2.99) =$$
$$= 0.9986 - 0.9750 = 0.0236$$

La corrispondente frequenza attesa è quindi $150 \cdot 0.0236 = 3.54$

Classi	*Frequenza osservata* f_i	*frequenze relative attese (probabilità)*	*frequenze attese*
$x \le 27$	0	0.0014	0.21
$27 < x \le 28$	3	0.0236	3.54
$28 < x \le 29$	23	0.1486	22.29
$29 < x \le 30$	53	0.3583	53.74
$30 < x \le 31$	50	0.3324	49.86
$31 < x \le 32$	21	0.1191	17.87
$x > 32$	0	0.0166	2.49
Totale	150	1	150

Tabella 16

Poiché ci sono delle classi che hanno frequenze assolute attese minori di 5, procediamo ad accorpare le prime tre classi e le ultime due e otteniamo la tabella 17

Classi	*Frequenza osservata* f_i	*frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
$x \le 29$	26	26.04	0.0001
$29 < x \le 30$	53	53.74	0.0102
$30 < x \le 31$	50	49.86	0.0004
$x > 31$	21	20.36	0.0201
Totale	150	150	0.0308

Tabella 17

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 0.0308$$

Le classi sono 4 e due parametri, valor medio e varianza della popolazione, sono stati stimati dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 4 - 1 - 2 = 1$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 3.841$$

Il valore del chi-quadro è minore del valore critico, perciò in base ai dati non possiamo rifiutare l'ipotesi nulla di adattamento, e concludiamo che, al livello di significatività del 5%, c'è un buon adattamento dei dati alla distribuzione normale.

Esempio 10

Sono state misurate le stature di 60 studenti e i dati sono stati raggruppati nella seguente distribuzione di frequenza

Classi (Statura in cm)	*Frequenza assoluta osservata*
$162 < x \le 165$	2
$165 < x \le 168$	13
$168 < x \le 171$	24
$171 < x \le 174$	15
$174 < x \le 177$	6

Verificare se la statura si può ritenere distribuita normalmente con media 170 cm e scarto quadratico medio 3 cm.

Calcoliamo le frequenze relative attese delle classi. Oltre alle classi indicate, nelle quali cadono i dati, occorre considerare anche la classe $x \le 162$ e la classe $x > 177$.
Per calcolare queste frequenze usiamo la funzione di ripartizione della variabile aleatoria normale X di media $\mu = 170$ e scarto quadratico medio $\sigma = 3$; passando alla variabile aleatoria standardizzata

$$Z = \frac{X - 170}{3}$$

e servendosi delle tavole della distribuzione normale standardizzata, si costruisce la tabella 18, contenente le frequenze relative attese e le frequenze assolute attese.

Classi	*Frequenza osservata*	*frequenze relative attese (probabilità)*	*frequenze attese*
$x \le 162$	0	0.0038	0.228
$162 < x \le 165$	2	0.0437	2.622
$165 < x \le 168$	13	0.2039	12.234
$168 < x \le 171$	24	0.3779	22.674
$171 < x \le 174$	15	0.2789	16.734
$174 < x \le 177$	6	0.0819	4.914
$x > 177$	0	0.0099	0.594
Totale	60	1	60

Tabella 18

Accorpiamo le prime tre classi e le ultime due, che hanno frequenze attese minori di 5 e otteniamo la tabella 19

Classi	*Frequenza osservata* O_i	*frequenze attese* A_i	$\frac{(O_i - A_i)^2}{A_i}$
$x \le 168$	15	15.084	0.0005
$168 < x \le 171$	24	22.674	0.075
$171 < x \le 174$	15	16.734	0.1797
$x > 174$	6	5.508	0.0439
Totale	60	60	0.3016

Tabella 19

Il valore della statistica chi-quadro calcolato dal campione è

$$\chi^2 = 0.3016$$

Le classi sono 4 e nessun parametro è stato stimato dal campione, perciò il grado di libertà è

$$\nu = k - 1 - m = 3$$

Il valore critico al livello di significatività del 5% è

$$\chi^2_{0.05} = 7.815$$

Il valore del chi-quadro è minore del valore critico, perciò in base ai dati non possiamo rifiutare l'ipotesi nulla di adattamento, e concludiamo che, al livello di significatività del 5%, c'è un buon adattamento dei dati alla distribuzione normale.

Sebbene sia frequente l'uso del test chi-quadro per saggiare l'eventuale distribuzione normale, esso in realtà non è il più idoneo quando la distribuzione ipotizzata è continua.
Esistono altri test più indicati per distribuzioni continue, ad esempio il test di Kolmogorov-Smirnov, che non sarà trattato in queste lezioni.

9.3 Test chi-quadro di indipendenza

Il test chi-quadro può essere utilizzato anche per verificare l'indipendenza o meno di due variabili: questa è forse la più frequente fra le applicazioni della distribuzione χ^2.
In questo test si vuole sottoporre a test l'ipotesi nulla che due criteri di classificazione, quando applicati al medesimo insieme di dati, siano indipendenti.
Si dice che due criteri di classificazione sono indipendenti se la distribuzione rispetto a un criterio non viene influenzata dalla classificazione rispetto all'altro criterio. Se l'ipotesi nulla viene rifiutata, concludiamo che i due criteri di classificazione sono indipendenti.
Vediamo alcuni esempi illustrativi.

Esempio 11
Un corso universitario è impartito dallo stesso insegnante a studenti del secondo anno di tre indirizzi di laurea diversi; gli esami superati e non superati sono registrati nella seguente tabella

	Laurea A	*Laurea B*	*Laurea C*
esame superato	80	40	110
esame non superato	50	20	40

Il rendimento degli studenti dei tre corsi, rispetto a questo esame, si può ritenere sostanzialmente equivalente, oppure le differenze sono statisticamente significative?
Questo equivale a chiedersi se le due variabili (qualitative) "indirizzo di laurea" e "superamento dell'esame" sono indipendenti.

Esempio 12
Per stabilire l'efficacia di un vaccino anti-influenzale è stata condotta una ricerca, somministrando il vaccino a 500 persone e controllando il loro stato di salute in un anno; lo stesso controllo è stato fatto per un gruppo di altre 500 persone non vaccinate; in base ai risultati dell'esperimento si è ottenuta la seguente tabella

	nessuna influenza	*una influenza*	*più di una influenza*
vaccinati	252	145	103
non vaccinati	224	136	140

Si può ritenere che il vaccino sia efficace, ossia sottoponendosi alla vaccinazione si ha un minor rischio di contrarre la malattia, oppure il vaccino non è efficace?
Questo equivale a chiedersi se le due variabili (qualitative) "vaccinazione" e "minor numero di influenze" sono indipendenti oppure no.

Esempio 13
Per verificare la qualità della produzione in una fabbrica, un ingegnere controlla il numero di pezzi difettosi prodotti da tre macchine diverse, e ottiene la seguente tabella di dati

	macchina 1	*macchina 2*	*macchina 3*
buoni	150	140	200
difettosi	25	40	20

Si può ritenere che la quantità di pezzi difettosi non dipenda dalla macchina che si utilizza?

In tutti gli esempi considerati disponiamo di *n* osservazioni congiunte di due variabili e ci chiediamo se esiste una forma di dipendenza fra le due variabili.

L'ipotesi nulla sarà che le due variabili siano indipendenti; se si rifiuta l'ipotesi nulla, la conclusione sarà che vi sia qualche interazione fra i due criteri di classificazione.
Tabelle come quelle riprodotte negli esempi si chiamano **tabelle di contingenza**. In una tabella di questo tipo *n* osservazioni sono classificate secondo un certo criterio *X*, ossia secondo il valore di una certa variabile, in *r* classi e, contemporaneamente, sono classificate secondo un altro criterio *Y* , ossia secondo i valori assunti da un'altra variabile, in *c* classi; la tabella riporta all'incrocio di ogni riga con ogni colonna la frequenza assoluta osservata O_{ij}

		Classi				
		1	2	3	…	*c*
Classi	1	O_{11}	O_{12}	O_{13}	…	O_{1c}
	2	O_{21}	O_{22}	O_{23}	…	O_{2c}
	3	O_{31}	O_{32}	O_{33}	…	O_{3c}
	…	…	…	…	…	…
	r	O_{r1}	O_{r2}	O_{r3}		O_{rc}

Tabella 20

Partendo da questa tabella si costruisce la **tabella delle frequenze attese**, ossia delle frequenze che si avrebbero nell'ipotesi di indipendenza; ogni frequenza attesa A_{ij} si ottiene con la seguente formula

$$A_{ij} = \frac{(\text{totale riga } i)\cdot(\text{totale colonna } j)}{\text{totale generale}}$$

Si dimostra che, per *n* sufficientemente grande, la statistica

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - A_{ij})^2}{A_{ij}} \tag{9.3}$$

detta il **chi-quadro** calcolato dal campione, ha approssimativamente la distribuzione χ^2 con grado di libertà $\nu = (r-1)\cdot(c-1)$.
Se l'ipotesi nulla H_0 è che le due variabili siano indipendenti, la regola di decisione sarà: si rifiuti l'ipotesi nulla, se il valore della statistica χ^2 calcolato dai dati è maggiore del valore critico χ^2_α

$$\chi^2 > \chi^2_\alpha \tag{9.4}$$

α è il livello di significatività stabilito e il grado di libertà della distribuzione χ^2 è $\nu = (r-1)\cdot(c-1)$.
Questa proceduta, detta **test chi-quadro di indipendenza**, è valida purché le frequenze assolute attese siano tutte maggiori o uguali a 5.

Esempio 11 – parte 2
Riprendiamo l'esempio 11.
Un corso universitario è impartito dallo stesso insegnante a studenti del secondo anno di tre indirizzi di laurea diversi; gli esami superati e non superati sono registrati nella tabella 21

Tabella di contingenza – Frequenze osservate				
	Laurea A	*Laurea B*	*Laurea C*	*Totale esami*
esame superato	80	40	110	230
esame non superato	50	20	40	110
Totale studenti iscritti	130	60	150	340

Tabella 21

Il rendimento degli studenti dei tre corsi, rispetto a questo esame, si può ritenere sostanzialmente equivalente, oppure le differenze sono statisticamente significative?

Costruiamo la tabella delle frequenze attese, ricordando che ogni casella contiene il prodotto del totale di riga per il totale di colonna, diviso per il totale generale. Nella tabella 21 l'ultima colonna e l'ultima riga contengono i totali parziali delle righe e delle colonne, che servono per calcolare le frequenze attese, l'ultima casella in basso a destra contiene il totale generale.

Frequenze attese			
	Laurea A	*Laurea B*	*Laurea C*
esame superato	87.94	40.59	101.47
esame non superato	42.06	19.41	48.53

Tabella 22

Servendosi delle tabelle delle frequenze osservate e delle frequenze attese si calcola il valore della statistica chi-quadro con la formula (9.3)

$$\chi^2 = \frac{(80-87.94)^2}{87.94} + \frac{(40-40.59)^2}{40.59} + \frac{(110-101.47)^2}{101.47} +$$
$$+ \frac{(50-42.06)^2}{42.06} + \frac{(20-19.41)^2}{19.41} + \frac{(40-48.53)^2}{48.53} = 4.46$$

Il grado di libertà è $\nu = (2-1)\cdot(3-1) = 2$; il valore critico al livello di significatività del 5% è $\chi^2_{0.05} = 5.991$. Poiché il valore della statistica chi-quadro è minore del valore critico, i dati non consentono di rifiutare l'ipotesi nulla e si conclude che il risultato dell'esame è indipendente dall'indirizzo di laurea, ossia il rendimento è equivalente.

Esempio 12 – parte 2

Riprendiamo l'esempio 12.

Per stabilire l'efficacia di un vaccino anti-influenzale è stata condotta una ricerca, somministrando il vaccino a 500 persone e controllando il loro stato di salute in un anno; lo stesso controllo è stato fatto per un gruppo di altre 500 persone non vaccinate; in base ai risultati dell'esperimento si è ottenuta la seguente tabella

Frequenze osservate				
	nessuna influenza	*una influenza*	*più di una influenza*	*Totale*
vaccinati	252	145	103	500
non vaccinati	224	136	140	500
Totale	476	281	243	1000

Tabella 23

Si può ritenere che il vaccino sia efficace, ossia sottoponendosi alla vaccinazione si ha un minor rischio di contrarre la malattia, oppure il vaccino non è efficace?

Costruiamo la tabella delle frequenze attese

Frequenze attese			
	nessuna influenza	*una influenza*	*più di una influenza*
vaccinati	238	140.5	121.5
non vaccinati	238	140.5	121.5

Tabella 24

Servendosi delle tabelle delle frequenze osservate e delle frequenze attese, si calcola il valore della statistica chi-quadro

$$\chi^2 = \frac{(252-238)^2}{238} + \frac{(145-140.5)^2}{140.5} + \frac{(103-121.5)^2}{121.5} +$$
$$+ \frac{(224-238)^2}{238} + \frac{(136-140.5)^2}{140.5} + \frac{(140-121.5)^2}{121.5} = 7.57$$

Il grado di libertà è $\nu = (2-1)\cdot(3-1) = 2$; il valore critico al livello di significatività del 5% è $\chi^2_{0.05} = 5.991$. Poiché il valore della statistica chi-quadro è maggiore del valore critico, i dati consentono di rifiutare l'ipotesi nulla: c'è evidenza statistica di efficacia del vaccino.

Per il livello di significatività dell'1% il valore critico è $\chi^2_{0.01} = 9.210$; in questo caso il valore della statistica chi-quadro è minore del valore critico, perciò non si rifiuta l'ipotesi nulla e si conclude che non c'è evidenza significativa di efficacia del vaccino: si tratta evidentemente di un caso che richiede ulteriori indagini.

Esempio 13 – parte 2

Per verificare la qualità della produzione in una fabbrica, un ingegnere controlla il numero di pezzi difettosi prodotti da tre macchine diverse, e ottiene la seguente tabella di dati

Frequenze osservate				
	macchina 1	*macchina 2*	*macchina 3*	*Totale buoni*
buoni	150	140	200	490
difettosi	25	40	20	85
Totale macchina	175	180	220	575

Tabella 25

Si può ritenere che la quantità di pezzi difettosi non dipenda dalla macchina che si utilizza?

Costruiamo la tabella delle frequenze attese

Frequenze attese			
	macchina 1	*macchina 2*	*macchina 3*
buoni	149.13	153.39	187.48
difettosi	25.87	26.61	32.52

Tabella 26

Servendosi delle tabelle delle frequenze osservate e delle frequenze attese, si calcola il valore della statistica chi-quadro

$$\chi^2 = \frac{(150-149.13)^2}{149.13} + \frac{(140-153.39)^2}{153.39} + \frac{(200-187.48)^2}{187.48} +$$
$$+ \frac{(25-25.87)^2}{25.87} + \frac{(40-26.61)^2}{26.61} + \frac{(20-32.52)^2}{32.52} = 13.60$$

Il grado di libertà è $\nu = (2-1)\cdot(3-1) = 2$; il valore critico al livello di significatività del 5% è $\chi^2_{0.05} = 5.991$. Poiché il valore della statistica chi-quadro è maggiore del valore critico, i dati consentono di rifiutare l'ipotesi nulle e si conclude che c'è evidenza statistica di una dipendenza del numero dei pezzi difettosi dalla macchina che si utilizza.

Esempio 14
Dall'esame del colore dei capelli dei bambini di una certa regione, si sono ricavati i seguenti dati

Frequenze osservate						
	biondo	*rosso*	*castano*	*bruno*	*nero*	*Totale*
maschi	592	119	849	504	36	2100
femmine	544	97	677	451	14	1783
Totale	1136	216	1526	955	50	3883

Tabella 27

Il colore dei capelli è indipendente dal sesso?

Costruiamo la tabella delle frequenze attese

Frequenze attese					
	biondo	*rosso*	*castano*	*bruno*	*nero*
maschi	614.37	116.82	825.29	516.48	27.04
femmine	521.63	99.18	700.71	438.52	22.96

Tabella 28

Servendosi delle tabelle delle frequenze osservate e delle frequenze attese si calcola il valore della statistica chi-quadro

$$\chi^2 = \frac{(592-614.37)^2}{614.37} + \frac{(119-116.82)^2}{116.82} + \frac{(849-825.29)^2}{825.29} +$$
$$+ \frac{(504-516.48)^2}{516.48} + \frac{(36-27.04)^2}{27.04} + \frac{(544-521.63)^2}{521.63} +$$
$$+ \frac{(97-99.18)^2}{99.18} + \frac{(677-700.71)^2}{700.71} + \frac{(451-438.52)^2}{521.63} +$$
$$+ \frac{(14-22.96)^2}{22.96} = 10.47$$

Il grado di libertà è $\nu = (2-1)\cdot(5-1) = 4$; il valore critico al livello di significatività dell'1% è $\chi^2_{0.05} = 13.277$. Poiché il valore della statistica chi-quadro è minore del valore critico, i dati non consentono di rifiutare l'ipotesi nulla e si conclude che c'è evidenza statistica di indipendenza del colore dei capelli dal sesso.

Il valore critico al livello di significatività del 5% è invece $\chi^2_{0.05} = 9.488$. Poiché il valore della statistica chi-quadro è in questo caso maggiore del valore critico, i dati consentono di rifiutare l'ipotesi nulla e si conclude che non c'è evidenza statistica di indipendenza del colore dei capelli dal sesso.
I risultati trovati ai due livelli di significatività non sono in accordo e questo fatto suggerisce la necessità di indagini più approfondite.

Appendice A. Tavole statistiche

Tavola 1. Distribuzione binomiale

La tavola fornisce i valori della funzione di distribuzione binomiale

$$B(x;n,p)=\sum_{k=0}^{x}\binom{n}{k}p^k(1-p)^{n-k}$$

per i valori $n=2:1:20$ e $p=0.05:0.05:0.95$.

Tavola 2. Distribuzione di Poisson

La tavola fornisce i valori della funzione di distribuzione di Poisson

$$F(x;\lambda)=\sum_{k=0}^{x}\frac{e^{-\lambda}\lambda^k}{k!}$$

per valori scelti di λ compresi fra 0.01 e 25 .

Tavola 3. Distribuzione normale standardizzata

La tavola fornisce il valore della funzione di distribuzione della variabile aleatoria standardizzata Z

$$F(z)=P(Z\le z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{t^2}{2}}dt$$

ossia l'area sottesa dalla curva $f(z)$, tra $-\infty$ e z.

Tavola 4. Percentili per la distribuzione normale standardizzata

La tavola fornisce i valori di z_α per i quali $P(z>z_\alpha)=\alpha\cdot 100\%=q\%$, per alcuni valori notevoli di q.

Tavola 5. Distribuzione *t* di Student

La tavola fornisce i valori di t_α per i quali $P(t>t_\alpha)=\alpha$, per i valori notevoli α = 0.10, 0.05, 0.025, 0.01 0.005 e per i valori del grado di libertà ν = 1:1:29.

Tavola 6. Distribuzione χ^2

La tavola fornisce i valori di χ^2_α per i quali $P(\chi^2>\chi^2_\alpha)=\alpha$, per i valori notevoli α = 0.995, 0.99, 0.975, 0.95, 0.05, 0.025, 0.01, 0.005 e per i valori del grado di libertà ν = 1:1:29.

Tavola 7. Distribuzione *F*

La tavola fornisce i valori di F_α per i quali $P(F>F_\alpha)=\alpha$, per i valori notevoli α = 0.25, 0.10, 0.05, 0.025, 0.01, 0.005 e per varie combinazioni di valori dei gradi di libertà ν_1 e ν_2.

Tavola 1 – Funzione di distribuzione binomiale $B(x;n,p)$

		p																		
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0-85	0.90	0.95
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500	0.2025	0.1600	0.1225	0.0900	0.0625	0.0400	0.0225	0.0100	0.0025
	1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500	0.6975	0.6400	0.5775	0.5100	0.4375	0.3600	0.2775	0.1900	0.0975
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250	0.0911	0.0640	0.0429	0.0270	0.0156	0.0080	0.0034	0.0010	0.0001
	1	0.9927	0.9720	0.9392	0.8960	0.8438	0.7840	0.7183	0.6480	0.5748	0.5000	0.4252	0.3520	0.2817	0.2160	0.1563	0.1040	0.0608	0.0280	0.0073
	2	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750	0.8336	0.7840	0.7254	0.6570	0.5781	0.4880	0.3859	0.2710	0.1426
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625	0.0410	0.0256	0.0150	0.0081	0.0039	0.0016	0.0005	0.0001	0.0000
	1	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125	0.2415	0.1792	0.1265	0.0837	0.0508	0.0272	0.0120	0.0037	0.0005
	2	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875	0.6090	0.5248	0.4370	0.3483	0.2617	0.1808	0.1095	0.0523	0.0140
	3	1.0000	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375	0.9085	0.8704	0.8215	0.7599	0.6836	0.5904	0.4780	0.3439	0.1855
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313	0.0185	0.0102	0.0053	0.0024	0.0010	0.0003	0.0001	0.0000	0.0000
	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875	0.1312	0.0870	0.0540	0.0308	0.0156	0.0067	0.0022	0.0005	0.0000
	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000	0.4069	0.3174	0.2352	0.1631	0.1035	0.0579	0.0266	0.0086	0.0012
	3	1.0000	0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125	0.7438	0.6630	0.5716	0.4718	0.3672	0.2627	0.1648	0.0815	0.0226
	4	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9687	0.9497	0.9222	0.8840	0.8319	0.7627	0.6723	0.5563	0.4095	0.2262
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156	0.0083	0.0041	0.0018	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000
	1	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094	0.0692	0.0410	0.0223	0.0109	0.0046	0.0016	0.0004	0.0001	0.0000
	2	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438	0.2553	0.1792	0.1174	0.0705	0.0376	0.0170	0.0059	0.0013	0.0001
	3	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6563	0.5585	0.4557	0.3529	0.2557	0.1694	0.0989	0.0473	0.0158	0.0022
	4	1.0000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906	0.8364	0.7667	0.6809	0.5798	0.4661	0.3446	0.2235	0.1143	0.0328
	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844	0.9723	0.9533	0.9246	0.8824	0.8220	0.7379	0.6229	0.4686	0.2649
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078	0.0037	0.0016	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625	0.0357	0.0188	0.0090	0.0038	0.0013	0.0004	0.0001	0.0000	0.0000
	2	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266	0.1529	0.0963	0.0556	0.0288	0.0129	0.0047	0.0012	0.0002	0.0000
	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000	0.3917	0.2898	0.1998	0.1260	0.0706	0.0333	0.0121	0.0027	0.0002
	4	1.0000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734	0.6836	0.5801	0.4677	0.3529	0.2436	0.1480	0.0738	0.0257	0.0038
	5	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375	0.8976	0.8414	0.7662	0.6706	0.5551	0.4233	0.2834	0.1497	0.0444
	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922	0.9848	0.9720	0.9510	0.9176	0.8665	0.7903	0.6794	0.5217	0.3017
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039	0.0017	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352	0.0181	0.0085	0.0036	0.0013	0.0004	0.0001	0.0000	0.0000	0.0000
	2	0.9942	0.9619	0.8948	0.7969	0.6785	05518	0.4278	0.3154	0.2201	0.1445	0.0885	0.0498	0.0253	0.0113	0.0042	0.0012	0.0002	0.0000	0.0000
	3	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633	0.2604	0.1737	0.1061	0.0580	0.0273	0.0104	0.0029	0.0004	0.0000
	4	1.0000	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367	0.5230	0.4059	0.2936	0.1941	0.1138	0.0563	0.0214	0.0050	0.0004
	5	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555	0.7799	0.6846	0.5722	0.4482	0.3215	0.2031	0.1052	0.0381	0.0058
	6	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648	0.9368	0.8936	0.8309	0.7447	0.6329	0.4967	0.3428	0.1869	0.0572
	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961	0.9916	0.9832	0.9681	0.9424	0.8999	0.8322	0.7275	0.5695	0.3366

		p																		
n	***x***	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0-85	0.90	0.95
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195	0.0091	0.0038	0.0014	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898	0.0498	0.0250	0.0112	0.0043	0.0013	0.0003	0.0000	0.0000	0.0000
	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539	0.1658	0.0994	0.0536	0.0253	0.0100	0.0031	0.0006	0.0001	0.0000
	4	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000	0.3786	0.2666	0.1717	0.0988	0.0489	0.0196	0.0056	0.0009	0.0000
	5	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461	0.6386	0.5174	0.3911	0.2703	0.1657	0.0856	0.0339	0.0083	0.0006
	6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102	0.8505	0.7682	0.6627	0.5372	0.3993	0.2618	0.1409	0.0530	0.0084
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805	0.9615	0.9295	0.8789	0.8040	0.6997	0.5638	0.4005	0.2252	0.0712
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980	0.9954	0.9899	0.9793	0.9596	0.9249	0.8658	0.7684	0.6126	0.3698
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107	0.0045	0.0017	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547	0.0274	0.0123	0.0048	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000
	3	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719	0.1020	0.0548	0.0260	0.0106	0.0035	0.0009	0.0001	0.0000	0.0000
	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770	0.2616	0.1662	0.0949	0.0473	0.0197	0.0064	0.0014	0.0001	0.0000
	5	1.0000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230	0.4956	0.3669	0.2485	0.1503	0.0781	0.0328	0.0099	0.0016	0.0001
	6	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281	0.7340	0.6177	0.4862	0.3504	0.2241	0.1209	0.0500	0.0128	0.0010
	7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453	0.9004	0.8327	0.7384	0.6172	0.4744	0.3222	0.1798	0.0702	0.0115
	8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893	0.9767	0.9536	0.9140	0.8507	0.7560	0.6242	0.4557	0.2639	0.0861
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9975	0.9940	0.9865	0.9718	0.9437	0.8926	0.8031	0.6513	0.4013
11	0	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.8981	0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059	0.0022	0.0007	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327	0.0148	0.0059	0.0020	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	3	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133	0.0610	0.0293	0.0122	0.0043	0.0012	0.0002	0.0000	0.0000	0.0000
	4	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744	0.1738	0.0994	0.0501	0.0216	0.0076	0.0020	0.0003	0.0000	0.0000
	5	1.0000	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000	0.3669	0.2465	0.1487	0.0782	0.0343	0.0117	0.0027	0.0003	0.0000
	6	1.0000	1.0000	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256	0.6029	0.4672	0.3317	0.2103	0.1146	0.0504	0.0159	0.0028	0.0001
	7	1.0000	1.0000	1.0000	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867	0.8089	0.7037	0.5744	0.4304	0.2867	0.1611	0.0694	0.0185	0.0016
	8	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673	0.9348	0.8811	0.7999	0.6873	0.5448	0.3826	0.2212	0.0896	0.0152
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9978	0.9941	0.9861	0.9698	0.9394	0.8870	0.8029	0.6779	0.5078	0.3026	0.1019
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9995	0.9986	0.9964	0.9912	0.9802	0.9578	0.9141	0.8327	0.6862	0.4312
12	0	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.8816	0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032	0.0011	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9804	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193	0.0079	0.0028	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9978	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730	0.0356	0.0153	0.0056	0.0017	0.0004	0.0001	0.0000	0.0000	0.0000
	4	0.9998	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938	0.1117	0.0573	0.0255	0.0095	0.0028	0.0006	0.0001	0.0000	0.0000
	5	1.0000	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872	0.2607	0.1582	0.0846	0.0386	0.0143	0.0039	0.0007	0.0001	0.0000
	6	1.0000	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128	0.4731	0.3348	0.2127	0.1178	0.0544	0.0194	0.0046	0.0005	0.0000
	7	1.0000	1.0000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062	0.6956	0.5618	0.4167	0.2763	0.1576	0.0726	0.0239	0.0043	0.0002
	8	1.0000	1.0000	1.0000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270	0.8655	0.7747	0.6533	0.5075	0.3512	0.2054	0.0922	0.0256	0.0022
	9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9972	0.9921	0.9807	0.9579	0.9166	0.8487	0.7472	0.6093	0.4417	0.2642	0.1109	0.0196
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968	0.9917	0.9804	0.9576	0.9150	0.8416	0.7251	0.5565	0.3410	0.1184
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9992	0.9978	0.9943	0.9862	0.9683	0.9313	0.8578	0.7176	0.4596

		p																
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0-85	0.90	0.95
13	0	0.5133	0.2542	0.1209	0.0550	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.8646	0.6213	0.3983	0.2336	0.1267	0.0637	0.0296	0.0126	0.0049	0.0017	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9755	0.8661	0.6920	0.5017	0.3326	0.2025	0.1132	0.0579	0.0269	0.0112	0.0041	0.0013	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9969	0.9658	0.8820	0.7473	0.5843	0.4206	0.2783	0.1686	0.0929	0.0461	0.0203	0.0078	0.0025	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	4	0.9997	0.9935	0.9658	0.9009	0.7940	0.6543	0.5005	0.3530	0.2279	0.1334	0.0698	0.0321	0.0126	0.0040	0.0010	0.0002	0.0000	0.0000	0.0000
	5	1.0000	0.9991	0.9925	0.9700	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905	0.1788	0.0977	0.0462	0.0182	0.0056	0.0012	0.0002	0.0000	0.0000
	6	1.0000	0.9999	0.9987	0.9930	0.9757	0.9376	0.8705	0.7712	0.6437	0.5000	0.3563	0.2288	0.1295	0.0624	0.0243	0.0070	0.0013	0.0001	0.0000
	7	1.0000	1.0000	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095	0.5732	0.4256	0.2841	0.1654	0.0802	0.0300	0.0075	0.0009	0.0000
	8	1.0000	1.0000	1.0000	0.9998	0.9990	0.9960	0.9874	0.9679	0.9302	0.8666	0.7721	0.6470	0.4995	0.3457	0.2060	0.0991	0.0342	0.0065	0.0003
	9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539	0.9071	0.8314	0.7217	0.5794	0.4157	0.2527	0.1180	0.0342	0.0031
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9987	0.9959	0.9888	0.9731	0.9421	0.8868	0.7975	0.6674	0.4983	0.3080	0.1339	0.0245
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9951	0.9874	0.9704	0.9363	0.8733	0.7664	0.6017	0.3787	0.1354
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9963	0.9903	0.9762	0.9450	0.8791	0.7458	0.4867
14	0	0.4877	0.2288	0.1028	0.0440	0.0178	0.0068	0.0024	0.0008	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.8470	0.5846	0.3567	0.1979	0.1010	0.0475	0.0205	0.0081	0.0029	0.0009	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9699	0.8416	0.6479	0.4481	0.2811	0.1608	0.0839	0.0398	0.0170	0.0065	0.0022	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9958	0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.0632	0.0287	0.0114	0.0039	0.0011	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9996	0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.0898	0.0426	0.0175	0.0060	0.0017	0.0003	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.2120	0.1189	0.0583	0.0243	0.0083	0.0022	0.0004	0.0000	0.0000	0.0000
	6	1.0000	0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953	0.2586	0.1501	0.0753	0.0315	0.0103	0.0024	0.0003	0.0000	0.0000
	7	1.0000	1.0000	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047	0.4539	0.3075	0.1836	0.0933	0.0383	0.0116	0.0022	0.0002	0.0000
	8	1.0000	1.0000	1.0000	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.7880	0.6627	0.5141	0.3595	0.2195	0.1117	0.0439	0.0115	0.0015	0.0000
	9	1.0000	1.0000	1.0000	1.0000	0.9997	0.9983	0.9940	0.9825	0.9574	0.9102	0.8328	0.7207	0.5773	0.4158	0.2585	0.1298	0.0467	0.0092	0.0004
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9989	0.9961	0.9886	0.9713	0.9368	0.8757	0.7795	0.6448	0.4787	0.3018	0.1465	0.0441	0.0042
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9978	0.9935	0.9830	0.9602	0.9161	0.8392	0.7189	0.5519	0.3521	0.1584	0.0301
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991	0.9971	0.9919	0.9795	0.9525	0.8990	0.8021	0.6433	0.4154	0.1530
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9992	0.9976	0.9932	0.9822	0.9560	0.8972	0.7712	0.5123
15	0	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037	0.0011	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176	0.0063	0.0019	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592	0.0255	0.0093	0.0028	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	5	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509	0.0769	0.0338	0.0124	0.0037	0.0008	0.0001	0.0000	0.0000	0.0000
	6	1.0000	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036	0.1818	0.0950	0.0422	0.0152	0.0042	0.0008	0.0001	0.0000	0.0000
	7	1.0000	1.0000	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000	0.3465	0.2131	0.1132	0.0500	0.0173	0.0042	0.0006	0.0000	0.0000
	8	1.0000	1.0000	0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964	0.5478	0.3902	0.2452	0.1311	0.0566	0.0181	0.0036	0.0003	0.0000
	9	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491	0.7392	0.5968	0.4357	0.2784	0.1484	0.0611	0.0168	0.0022	0.0001
	10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408	0.8796	0.7827	0.6481	0.4845	0.3135	0.1642	0.0617	0.0127	0.0006
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9937	0.9824	0.9576	0.9095	0.8273	0.7031	0.5387	0.3518	0.1773	0.0556	0.0055
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9963	0.9893	0.9729	0.9383	0.8732	0.7639	0.6020	0.3958	0.1841	0.0362
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9948	0.9858	0.9647	0.9198	0.8329	0.6814	0.4510	0.1710
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9984	0.9953	0.9866	0.9648	0.9126	0.7941	0.5367

		p																		
n	***x***	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0-85	0.90	0.95
16	0	0.4401	0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.8108	0.5147	0.2839	0.1407	0.0635	0.0261	0.0098	0.0033	0.0010	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9571	0.7892	0.5614	0.3518	0.1971	0.0994	0.0451	0.0183	0.0066	0.0021	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9930	0.9316	0.7899	0.5981	0.4050	0.2459	0.1339	0.0651	0.0281	0.0106	0.0035	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9991	0.9830	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.0853	0.0384	0.0149	0.0049	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9999	0.9967	0.9765	0.9183	0.8103	0.6598	0.4900	0.3288	0.1976	0.1051	0.0486	0.0191	0.0062	0.0016	0.0003	0.0000	0.0000	0.0000	0.0000
	6	1.0000	0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.3660	0.2272	0.1241	0.0583	0.0229	0.0071	0.0016	0.0002	0.0000	0.0000	0.0000
	7	1.0000	0.9999	0.9989	0.9930	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018	0.2559	0.1423	0.0671	0.0257	0.0075	0.0015	0.0002	0.0000	0.0000
	8	1.0000	1.0000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982	0.4371	0.2839	0.1594	0.0744	0.0271	0.0070	0.0011	0.0001	0.0000
	9	1.0000	1.0000	1.0000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728	0.6340	0.4728	0.3119	0.1753	0.0796	0.0267	0.0056	0.0005	0.0000
	10	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949	0.8024	0.6712	0.5100	0.3402	0.1897	0.0817	0.0235	0.0033	0.0001
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9851	0.9616	0.9147	0.8334	0.7108	0.5501	0.3698	0.2018	0.0791	0.0170	0.0009
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9991	0.9965	0.9894	0.9719	0.9349	0.8661	0.7541	0.5950	0.4019	0.2101	0.0684	0.0070
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9979	0.9934	0.9817	0.9549	0.9006	0.8029	0.6482	0.4386	0.2108	0.0429
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9967	0.9902	0.9739	0.9365	0.8593	0.7161	0.4853	0.1892
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9967	0.9900	0.9719	0.9257	0.8147	0.5599
17	0	0.4181	0.1668	0.0631	0.0225	0.0075	0.0023	0.0007	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.7922	0.4818	0.2525	0.1182	0.0501	0.0193	0.0067	0.0021	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9497	0.7618	0.5198	0.3096	0.1637	0.0774	0.0327	0.0123	0.0041	0.0012	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9912	0.9174	0.7556	0.5489	0.3530	0.2019	0.1028	0.0464	0.0184	0.0064	0.0019	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9988	0.9779	0.9013	0.7582	0.5739	0.3887	0.2348	0.1260	0.0596	0.0245	0.0086	0.0025	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9999	0.9953	0.9681	0.8943	0.7653	0.5968	0.4197	0.2639	0.1471	0.0717	0.0301	0.0106	0.0030	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	6	1.0000	0.9992	0.9917	0.9623	0.8929	0.7752	0.6188	0.4478	0.2902	0.1662	0.0826	0.0348	0.0120	0.0032	0.0006	0.0001	0.0000	0.0000	0.0000
	7	1.0000	0.9999	0.9983	0.9891	0.9598	0.8954	0.7872	0.6405	0.4743	0.3145	0.1834	0.0919	0.0383	0.0127	0.0031	0.0005	0.0000	0.0000	0.0000
	8	1.0000	1.0000	0.9997	0.9974	0.9876	0.9597	0.9006	0.8011	0.6626	0.5000	0.3374	0.1989	0.0994	0.0403	0.0124	0.0026	0.0003	0.0000	0.0000
	9	1.0000	1.0000	1.0000	0.9995	0.9969	0.9873	0.9617	0.9081	0.8166	0.6855	0.5257	0.3595	0.2128	0.1046	0.0402	0.0109	0.0017	0.0001	0.0000
	10	1.0000	1.0000	1.0000	0.9999	0.9994	0.9968	0.9880	0.9652	0.9174	0.8338	0.7098	0.5522	0.3812	0.2248	0.1071	0.0377	0.0083	0.0008	0.0000
	11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9970	0.9894	0.9699	0.9283	0.8529	0.7361	0.5803	0.4032	0.2347	0.1057	0.0319	0.0047	0.0001
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9975	0.9914	0.9755	0.9404	0.8740	0.7652	0.6113	0.4261	0.2418	0.0987	0.0221	0.0012
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9936	0.9816	0.9536	0.8972	0.7981	0.6470	0.4511	0.2444	0.0826	0.0088
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9988	0.9959	0.9877	0.9673	0.9226	0.8363	0.6904	0.4802	0.2382	0.0503
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9979	0.9933	0.9807	0.9499	0.8818	0.7475	0.5182	0.2078
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9977	0.9925	0.9775	0.9369	0.8332	0.5819
18	0	0.3972	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.7735	0.4503	0.2241	0.0991	0.0395	0.0142	0.0046	0.0013	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9419	0.7338	0.4797	0.2713	0.1353	0.0600	0.0236	0.0082	0.0025	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9891	0.9018	0.7202	0.5010	0.3057	0.1646	0.0783	0.0328	0.0120	0.0038	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9985	0.9718	0.8794	0.7164	0.5187	0.3327	0.1886	0.0942	0.0411	0.0154	0.0049	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9998	0.9936	0.9581	0.8671	0.7175	0.5344	0.3550	0.2088	0.1077	0.0481	0.0183	0.0058	0.0014	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	0.9988	0.9882	0.9487	0.8610	0.7217	0.5491	0.3743	0.2258	0.1189	0.0537	0.0203	0.0062	0.0014	0.0002	0.0000	0.0000	0.0000	0.0000
	7	1.0000	0.9998	0.9973	0.9837	0.9431	0.8593	0.7283	0.5634	0.3915	0.2403	0.1280	0.0576	0.0212	0.0061	0.0012	0.0002	0.0000	0.0000	0.0000
	8	1.0000	1.0000	0.9995	0.9957	0.9807	0.9404	0.8609	0.7368	0.5778	0.4073	0.2527	0.1347	0.0597	0.0210	0.0054	0.0009	0.0001	0.0000	0.0000
	9	1.0000	1.0000	0.9999	0.9991	0.9946	0.9790	0.9403	0.8653	0.7473	0.5927	0.4222	0.2632	0.1391	0.0596	0.0193	0.0043	0.0005	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9998	0.9988	0.9939	0.9788	0.9424	0.8720	0.7597	0.6085	0.4366	0.2717	0.1407	0.0569	0.0163	0.0027	0.0002	0.0000
	11	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986	0.9938	0.9797	0.9463	0.8811	0.7742	0.6257	0.4509	0.2783	0.1390	0.0513	0.0118	0.0012	0.0000

		p																		
n	*x*	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0-85	0.90	0.95
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9986	0.9942	0.9817	0.9519	0.8923	0.7912	0.6450	0.4656	0.2825	0.1329	0.0419	0.0064	0.0002
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9846	0.9589	0.9058	0.8114	0.6673	0.4813	0.2836	0.1206	0.0282	0.0015
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9962	0.9880	0.9672	0.9217	0.8354	0.6943	0.4990	0.2798	0.0982	0.0109
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9975	0.9918	0.9764	0.9400	0.8647	0.7287	0.5203	0.2662	0.0581
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9987	0.9954	0.9858	0.9605	0.9009	0.7759	0.5497	0.2265
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9944	0.9820	0.9464	0.8499	0.6028
19	0	0.3774	0.1351	0.0456	0.0144	0.0042	0.0011	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.7547	0.4203	0.1985	0.0829	0.0310	0.0104	0.0031	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9335	0.7054	0.4413	0.2369	0.1113	0.0462	0.0170	0.0055	0.0015	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9868	0.8850	0.6841	0.4551	0.2631	0.1332	0.0591	0.0230	0.0077	0.0022	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9980	0.9648	0.8556	0.6733	0.4654	0.2822	0.1500	0.0696	0.0280	0.0096	0.0028	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9998	0.9914	0.9463	0.8369	0.6678	0.4739	0.2968	0.1629	0.0777	0.0318	0.0109	0.0031	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	0.9983	0.9837	0.9324	0.8251	0.6655	0.4812	0.3081	0.1727	0.0835	0.0342	0.0116	0.0031	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	7	1.0000	0.9997	0.9959	0.9767	0.9225	0.8180	0.6656	0.4878	0.3169	0.1796	0.0871	0.0352	0.0114	0.0028	0.0005	0.0000	0.0000	0.0000	0.0000
	8	1.0000	1.0000	0.9992	0.9933	0.9713	0.9161	0.8145	0.6675	0.4940	0.3238	0.1841	0.0885	0.0347	0.0105	0.0023	0.0003	0.0000	0.0000	0.0000
	9	1.0000	1.0000	0.9999	0.9984	0.9911	0.9674	0.9125	0.8139	0.6710	0.5000	0.3290	0.1861	0.0875	0.0326	0.0089	0.0016	0.0001	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9997	0.9977	0.9895	0.9653	0.9115	0.8159	0.6762	0.5060	0.3325	0.1855	0.0839	0.0287	0.0067	0.0008	0.0000	0.0000
	11	1.0000	1.0000	1.0000	1.0000	0.9995	0.9972	0.9886	0.9648	0.9129	0.8204	0.6831	0.5122	0.3344	0.1820	0.0775	0.0233	0.0041	0.0003	0.0000
	12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9969	0.9884	0.9658	0.9165	0.8273	0.6919	0.5188	0.3345	0.1749	0.0676	0.0163	0.0017	0.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9969	0.9891	0.9682	0.9223	0.8371	0.7032	0.5261	0.3322	0.1631	0.0537	0.0086	0.0002
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9972	0.9904	0.9720	0.9304	0.8500	0.7178	0.5346	0.3267	0.1444	0.0352	0.0020
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9978	0.9923	0.9770	0.9409	0.8668	0.7369	0.5449	0.3159	0.1150	0.0132
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9985	0.9945	0.9830	0.9538	0.8887	0.7631	0.5587	0.2946	0.0665
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9969	0.9896	0.9690	0.9171	0.8015	0.5797	0.2453
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9958	0.9856	0.9544	0.8649	0.6226
20	0	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.0510	0.0189	0.0059	0.0015	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207	0.0064	0.0016	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577	0.0214	0.0065	0.0015	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	7	1.0000	0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316	0.0580	0.0210	0.0060	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000
	8	1.0000	0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517	0.1308	0.0565	0.0196	0.0051	0.0009	0.0001	0.0000	0.0000	0.0000
	9	1.0000	1.0000	0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119	0.2493	0.1275	0.0532	0.0171	0.0039	0.0006	0.0000	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881	0.4086	0.2447	0.1218	0.0480	0.0139	0.0026	0.0002	0.0000	0.0000
	11	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483	0.5857	0.4044	0.2376	0.1133	0.0409	0.0100	0.0013	0.0001	0.0000
	12	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9940	0.9790	0.9420	0.8684	0.7480	0.5841	0.3990	0.2277	0.1018	0.0321	0.0059	0.0004	0.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9935	0.9786	0.9423	0.8701	0.7500	0.5834	0.3920	0.2142	0.0867	0.0219	0.0024	0.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9936	0.9793	0.9447	0.8744	0.7546	0.5836	0.3828	0.1958	0.0673	0.0113	0.0003
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9941	0.9811	0.9490	0.8818	0.7625	0.5852	0.3704	0.1702	0.0432	0.0026
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9951	0.9840	0.9556	0.8929	0.7748	0.5886	0.3523	0.1330	0.0159
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9991	0.9964	0.9879	0.9645	0.9087	0.7939	0.5951	0.3231	0.0755
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9979	0.9924	0.9757	0.9308	0.8244	0.6083	0.2642
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9968	0.9885	0.9612	0.8784	0.6415

Tavola 2 – Funzione di distribuzione di Poisson $F(x;\lambda)$

$$F(x;\lambda)=\sum_{k=0}^{x}\frac{e^{-\lambda}\lambda^{k}}{k!}$$

λ \ x	0	1	2	3	4	5	6	7	8	9
0.01	0.9900	1.0000								
0.02	0.9802	0.9998								
0.03	0.9704	0.9996								
0.04	0.9608	0.9992								
0.05	0.9512	0.9988	1.0000							
0.06	0.9418	0.9983	1.0000							
0.07	0.9324	0.9977	0.9999							
0.08	0.9231	0.9970	0.9999							
0.09	0.9139	0.9962	0.9999							
0.10	0.9048	0.9953	0.9998	1.0000						
0.15	0.8607	0.9898	0.9995	1.0000						
0.20	0.8187	0.9825	0.9989	0.9999						
0.25	0.7788	0.9735	0.9978	0.9999						
0.30	0.7408	0.9631	0.9964	0.9997	1.0000					
0.35	0.7047	0.9513	0.9945	0.9995	1.0000					
0.40	0.6703	0.9384	0.9921	0.9992	0.9999					
0.45	0.6376	0.9246	0.9891	0.9988	0.9999					
0.50	0.6065	0.9098	0.9856	0.9982	0.9998					
0.55	0.5769	0.8943	0.9815	0.9975	0.9997	1.0000				
0.60	0.5488	0.8781	0.9769	0.9966	0.9996	1.0000				
0.65	0.5220	0.8614	0.9717	0.9956	0.9994	0.9999				
0.70	0.4966	0.8442	0.9659	0.9942	0.9992	0.9999				
0.75	0.4724	0.8266	0.9595	0.9927	0.9989	0.9999				
0.80	0.4493	0.8088	0.9526	0.9909	0.9986	0.9998				
0.85	0.4274	0.7907	0.9451	0.9889	0.9982	0.9997	1.0000			
0.90	0.4066	0.7725	0.9371	0.9865	0.9977	0.9997	1.0000			
0.95	0.3867	0.7541	0.9287	0.9839	0.9971	0.9995	0.9999			
1.00	0.3679	0.7358	0.9197	0.9810	0.9963	0.9994	0.9999			
1.1	0.3329	0.6990	0.9004	0.9743	0.9946	0.9990	0.9999	1.0000		
1.2	0.3012	0.6626	0.8795	0.9662	0.9923	0.9985	0.9997	1.0000		
1.3	0.2725	0.6268	0.8571	0.9569	0.9893	0.9978	0.9996	0.9999		
1.4	0.2466	0.5918	0.8335	0.9463	0.9857	0.9968	0.9994	0.9999		
1.5	0.2231	0.5578	0.8088	0.9344	0.9814	0.9955	0.9991	0.9998	1.0000	
1.6	0.2019	0.5249	0.7834	0.9212	0.9763	0.9940	0.9987	0.9997	1.0000	
1.7	0.1827	0.4932	0.7572	0.9068	0.9704	0.9920	0.9981	0.9996	0.9999	
1.8	0.1653	0.4628	0.7306	0.8913	0.9636	0.9896	0.9974	0.9994	0.9999	
1.9	0.1496	0.4337	0.7037	0.8747	0.9559	0.9868	0.9966	0.9992	0.9998	1.0000
2.0	0.1353	0.4060	0.6767	0.8571	0.9473	0.9834	0.9955	0.9989	0.9998	1.0000

λ \ x	0	1	2	3	4	5	6	7	8	9
2.1	0.1225	0.3796	0.6496	0.8386	0.9379	0.9796	0.9941	0.9985	0.9997	0.9999
2.2	0.1108	0.3546	0.6227	0.8194	0.9275	0.9751	0.9925	0.9980	0.9995	0.9999
2.3	0.1003	0.3309	0.5960	0.7993	0.9162	0.9700	0.9906	0.9974	0.9994	0.9999
2.4	0.0907	0.3084	0.5697	0.7787	0.9041	0.9643	0.9884	0.9967	0.9991	0.9998
2.5	0.0821	0.2873	0.5438	0.7576	0.8912	0.9580	0.9858	0.9958	0.9989	0.9997
2.6	0.0743	0.2674	0.5184	0.7360	0.8774	0.9510	0.9828	0.9947	0.9985	0.9996
2.7	0.0672	0.2487	0.4936	0.7141	0.8629	0.9433	0.9794	0.9934	0.9981	0.9995
2.8	0.0608	0.2311	0.4695	0.6919	0.8477	0.9349	0.9756	0.9919	0.9976	0.9993
2.9	0.0550	0.2146	0.4460	0.6696	0.8318	0.9258	0.9713	0.9901	0.9969	0.9991
3.0	0.0498	0.1991	0.4232	0.6472	0.8153	0.9161	0.9665	0.9881	0.9962	0.9989
3.2	0.0408	0.1712	0.3799	0.6025	0.7806	0.8946	0.9554	0.9832	0.9943	0.9982
3.4	0.0334	0.1468	0.3397	0.5584	0.7442	0.8705	0.9421	0.9769	0.9917	0.9973
3.6	0.0273	0.1257	0.3027	0.5152	0.7064	0.8441	0.9267	0.9692	0.9883	0.9960
3.8	0.0224	0.1074	0.2689	0.4735	0.6678	0.8156	0.9091	0.9599	0.9840	0.9942
4.0	0.0183	0.0916	0.2381	0.4335	0.6288	0.7851	0.8893	0.9489	0.9786	0.9919
4.2	0.0150	0.0780	0.2102	0.3954	0.5898	0.7531	0.8675	0.9361	0.9721	0.9889
4.4	0.0123	0.0663	0.1851	0.3594	0.5512	0.7199	0.8436	0.9214	0.9642	0.9851
4.6	0.0101	0.0563	0.1626	0.3257	0.5132	0.6858	0.8180	0.9049	0.9549	0.9805
4.8	0.0082	0.0477	0.1425	0.2942	0.4763	0.6510	0.7908	0.8867	0.9442	0.9749
5.0	0.0067	0.0404	0.1247	0.2650	0.4405	0.6160	0.7622	0.8666	0.9319	0.9682
5.2	0.0055	0.0342	0.1088	0.2381	0.4061	0.5809	0.7324	0.8449	0.9181	0.9603
5.4	0.0045	0.0289	0.0948	0.2133	0.3733	0.5461	0.7017	0.8217	0.9027	0.9512
5.6	0.0037	0.0244	0.0824	0.1906	0.3422	0.5119	0.6703	0.7970	0.8857	0.9409
5.8	0.0030	0.0206	0.0715	0.1700	0.3127	0.4783	0.6384	0.7710	0.8672	0.9292
6.0	0.0025	0.0174	0.0620	0.1512	0.2851	0.4457	0.6063	0.7440	0.8472	0.9161

λ \ x	10	11	12	13	14	15	16	17
2.6	0.9999	1.0000						
2.8	0.9998	1.0000						
2.9	0.9998	0.9999						
3.0	0.9997	0.9999	1.0000					
3.2	0.9995	0.9999	1.0000	1.0000				
3.4	0.9992	0.9998	0.9999	1.0000				
3.6	0.9987	0.9996	0.9999	1.0000				
3.8	0.9981	0.9994	0.9998	1.0000				
4.0	0.9972	0.9991	0.9997	0.9999	1.0000			
4.2	0.9959	0.9986	0.9996	0.9999	1.0000			
4.4	0.9943	0.9980	0.9993	0.9998	0.9999			
4.6	0.9922	0.9971	0.9990	0.9997	0.9999	1.0000		
4.8	0.9896	0.9960	0.9986	0.9995	0.9999	1.0000		
5.0	0.9863	0.9945	0.9980	0.9993	0.9998	0.9999	1.0000	
5.2	0.9823	0.9927	0.9972	0.9990	0.9997	0.9999	1.0000	
5.4	0.9775	0.9904	0.9962	0.9986	0.9995	0.9998	0.9999	
5.6	0.9718	0.9875	0.9949	0.9980	0.9993	0.9998	0.9999	1.0000
5.8	0.9651	0.9841	0.9932	0.9973	0.9990	0.9996	0.9999	1.0000
6.0	0.9574	0.9799	0.9912	0.9964	0.9986	0.9995	0.9998	0.9999

λ \ x	0	1	2	3	4	5	6	7	8	9
6.2	0.0020	0.0146	0.0536	0.1342	0.2592	0.4141	0.5742	0.7160	0.8259	0.9016
6.4	0.0017	0.0123	0.0463	0.1189	0.2351	0.3837	0.5423	0.6873	0.8033	0.8858
6.6	0.0014	0.0103	0.0400	0.1052	0.2127	0.3547	0.5108	0.6581	0.7796	0.8686
6.8	0.0011	0.0087	0.0344	0.0928	0.1920	0.3270	0.4799	0.6285	0.7548	0.8502
7.0	0.0009	0.0073	0.0296	0.0818	0.1730	0.3007	0.4497	0.5987	0.7291	0.8305
7.2	0.0007	0.0061	0.0255	0.0719	0.1555	0.2759	0.4204	0.5689	0.7027	0.8096
7.4	0.0006	0.0051	0.0219	0.0632	0.1395	0.2526	0.3920	0.5393	0.6757	0.7877
7.6	0.0005	0.0043	0.0188	0.0554	0.1249	0.2307	0.3646	0.5100	0.6482	0.7649
7.8	0.0004	0.0036	0.0161	0.0485	0.1117	0.2103	0.3384	0.4812	0.6204	0.7411
8.0	0.0003	0.0030	0.0138	0.0424	0.0996	0.1912	0.3134	0.4530	0.5925	0.7166
8.2	0.0003	0.0025	0.0118	0.0370	0.0887	0.1736	0.2896	0.4254	0.5647	0.6915
8.4	0.0002	0.0021	0.0100	0.0323	0.0789	0.1573	0.2670	0.3987	0.5369	0.6659
8.6	0.0002	0.0018	0.0086	0.0281	0.0701	0.1422	0.2457	0.3728	0.5094	0.6400
8.8	0.0002	0.0015	0.0073	0.0244	0.0621	0.1284	0.2256	0.3478	0.4823	0.6137
9.0	0.0001	0.0012	0.0062	0.0212	0.0550	0.1157	0.2068	0.3239	0.4557	0.5874
9.2	0.0001	0.0010	0.0053	0.0184	0.0486	0.1041	0.1892	0.3010	0.4296	0.5611
9.4	0.0001	0.0009	0.0045	0.0160	0.0429	0.0935	0.1727	0.2792	0.4042	0.5349
9.6	0.0001	0.0007	0.0038	0.0138	0.0378	0.0838	0.1574	0.2584	0.3796	0.5089
9.8	0.0001	0.0006	0.0033	0.0120	0.0333	0.0750	0.1433	0.2388	0.3558	0.4832
10.0	0.0000	0.0005	0.0028	0.0103	0.0293	0.0671	0.1301	0.2202	0.3328	0.4579

λ \ x	10	11	12	13	14	15	16	17	18	19
6.2	0.9486	0.9750	0.9887	0.9952	0.9981	0.9993	0.9997	0.9999	1.0000	
6.4	0.9386	0.9693	0.9857	0.9937	0.9974	0.9990	0.9996	0.9999	1.0000	
6.6	0.9274	0.9627	0.9821	0.9920	0.9966	0.9986	0.9995	0.9998	0.9999	
6.8	0.9151	0.9552	0.9779	0.9898	0.9956	0.9982	0.9993	0.9997	0.9999	1.0000
7.0	0.9015	0.9467	0.9730	0.9872	0.9943	0.9976	0.9990	0.9996	0.9999	1.0000
7.2	0.8867	0.9371	0.9673	0.9841	0.9927	0.9969	0.9987	0.9995	0.9998	0.9999
7.4	0.8707	0.9265	0.9609	0.9805	0.9908	0.9959	0.9983	0.9993	0.9997	0.9999
7.6	0.8535	0.9148	0.9536	0.9762	0.9886	0.9948	0.9978	0.9991	0.9996	0.9999
7.8	0.8352	0.9020	0.9454	0.9714	0.9859	0.9934	0.9971	0.9988	0.9995	0.9998
8.0	0.8159	0.8881	0.9362	0.9658	0.9827	0.9918	0.9963	0.9984	0.9993	0.9997
8.2	0.7955	0.8731	0.9261	0.9595	0.9791	0.9898	0.9953	0.9979	0.9991	0.9997
8.4	0.7743	0.8571	0.9150	0.9524	0.9749	0.9875	0.9941	0.9973	0.9989	0.9995
8.6	0.7522	0.8400	0.9029	0.9445	0.9701	0.9848	0.9926	0.9966	0.9985	0.9994
8.8	0.7294	0.8220	0.8898	0.9358	0.9647	0.9816	0.9909	0.9957	0.9981	0.9992
9.0	0.7060	0.8030	0.8758	0.9261	0.9585	0.9780	0.9889	0.9947	0.9976	0.9989
9.2	0.6820	0.7832	0.8607	0.9156	0.9517	0.9738	0.9865	0.9934	0.9969	0.9986
9.4	0.6576	0.7626	0.8448	0.9042	0.9441	0.9691	0.9838	0.9919	0.9962	0.9983
9.6	0.6329	0.7412	0.8279	0.8919	0.9357	0.9638	0.9806	0.9902	0.9952	0.9978
9.8	0.6080	0.7193	0.8101	0.8786	0.9265	0.9579	0.9770	0.9881	0.9941	0.9972
10.0	0.5830	0.6968	0.7916	0.8645	0.9165	0.9513	0.9730	0.9857	0.9928	0.9965

λ \ x	20	21	22	23	24
7.4	1.0000				
7.6	1.0000				
7.8	0.9999				
8.0	0.9999	1.0000			
8.2	0.9999	1.0000			
8.4	0.9998	0.9999			
8.6	0.9998	0.9999	1.0000		
8.8	0.9997	0.9999	1.0000		
9.0	0.9996	0.9998	0.9999		
9.2	0.9994	0.9998	0.9999	1.0000	
9.4	0.9992	0.9997	0.9999	1.0000	
9.6	0.9990	0.9996	0.9998	0.9999	
9.8	0.9987	0.9995	0.9998	0.9999	1.0000
10.0	0.9984	0.9993	0.9997	0.9999	1.0000

λ \ x	0	1	2	3	4	5	6	7	8	9
10.5	0.0000	0.0003	0.0018	0.0071	0.0211	0.0504	0.1016	0.1785	0.2794	0.3971
11.0	0.0000	0.0002	0.0012	0.0049	0.0151	0.0375	0.0786	0.1432	0.2320	0.3405
11.5	0.0000	0.0001	0.0008	0.0034	0.0107	0.0277	0.0603	0.1137	0.1906	0.2888
12.0	0.0000	0.0001	0.0005	0.0023	0.0076	0.0203	0.0458	0.0895	0.1550	0.2424
12.5	0.0000	0.0001	0.0003	0.0016	0.0053	0.0148	0.0346	0.0698	0.1249	0.2014
13.0	0.0000	0.0000	0.0002	0.0011	0.0037	0.0107	0.0259	0.0540	0.0998	0.1658
13.5	0.0000	0.0000	0.0001	0.0007	0.0026	0.0077	0.0193	0.0415	0.0790	0.1353
14.0	0.0000	0.0000	0.0001	0.0005	0.0018	0.0055	0.0142	0.0316	0.0621	0.1094
14.5	0.0000	0.0000	0.0001	0.0003	0.0012	0.0039	0.0105	0.0239	0.0484	0.0878
15.0	0.0000	0.0000	0.0000	0.0002	0.0009	0.0028	0.0076	0.0180	0.0374	0.0699

λ \ x	10	11	12	13	14	15	16	17	18	19
10.5	0.5207	0.6387	0.7420	0.8253	0.8879	0.9317	0.9604	0.9781	0.9885	0.9942
11.0	0.4599	0.5793	0.6887	0.7813	0.8540	0.9074	0.9441	0.9678	0.9823	0.9907
11.5	0.4017	0.5198	0.6329	0.7330	0.8153	0.8783	0.9236	0.9542	0.9738	0.9857
12.0	0.3472	0.4616	0.5760	0.6815	0.7720	0.8444	0.8987	0.9370	0.9626	0.9787
12.5	0.2971	0.4058	0.5190	0.6278	0.7250	0.8060	0.8693	0.9158	0.9481	0.9694
13.0	0.2517	0.3532	0.4631	0.5730	0.6751	0.7636	0.8355	0.8905	0.9302	0.9573
13.5	0.2112	0.3045	0.4093	0.5182	0.6233	0.7178	0.7975	0.8609	0.9084	0.9421
14.0	0.1757	0.2600	0.3585	0.4644	0.5704	0.6694	0.7559	0.8272	0.8826	0.9235
14.5	0.1449	0.2201	0.3111	0.4125	0.5176	0.6192	0.7112	0.7897	0.8530	0.9012
15.0	0.1185	0.1848	0.2676	0.3632	0.4657	0.5681	0.6641	0.7489	0.8195	0.8752

λ \ x	20	21	22	23	24	25	26	27	28	29
10.5	0.9972	0.9987	0.9994	0.9998	0.9999	1.0000	1.0000			
11.0	0.9953	0.9977	0.9990	0.9995	0.9998	0.9999	1.0000	1.0000		
11.5	0.9925	0.9962	0.9982	0.9992	0.9996	0.9998	0.9999	1.0000		
12.0	0.9884	0.9939	0.9970	0.9985	0.9993	0.9997	0.9999	0.9999	1.0000	
12.5	0.9827	0.9906	0.9951	0.9975	0.9988	0.9994	0.9997	0.9999	1.0000	1.0000
13.0	0.9750	0.9859	0.9924	0.9960	0.9980	0.9990	0.9995	0.9998	0.9999	1.0000
13.5	0.9649	0.9796	0.9885	0.9938	0.9968	0.9984	0.9992	0.9996	0.9998	0.9999
14.0	0.9521	0.9712	0.9833	0.9907	0.9950	0.9974	0.9987	0.9994	0.9997	0.9999
14.5	0.9362	0.9604	0.9763	0.9863	0.9924	0.9959	0.9979	0.9989	0.9995	0.9998
15.0	0.9170	0.9469	0.9673	0.9805	0.9888	0.9938	0.9967	0.9983	0.9991	0.9996

λ \ x	29	30	31
13.0	1.0000		
13.5	1.0000		
14.0	0.9999	1.0000	
14.5	0.9999	1.0000	1.0000
15.0	0.9998	0.9999	1.0000

λ \ x	2	3	4	5	6	7	8	9	10	11
16	0.0000	0.0001	0.0004	0.0014	0.0040	0.0100	0.0220	0.0433	0.0774	0.1270
17	0.0000	0.0000	0.0004	0.0007	0.0021	0.0054	0.0126	0.0261	0.0491	0.0847
18	0.0000	0.0000	0.0001	0.0003	0.0010	0.0029	0.0071	0.0154	0.0304	0.0549
19	0.0000	0.0000	0.0001	0.0002	0.0005	0.0015	0.0039	0.0089	0.0183	0.0347
20	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0021	0.0050	0.0108	0.0214
21	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0011	0.0028	0.0063	0.0129
22	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0015	0.0035	0.0076
23	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020	0.0044
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0004	0.0011	0.0025
25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0014

λ \ x	12	13	14	15	16	17	18	19	20	21
16	0.1931	0.2745	0.3675	0.4667	0.5660	0.6593	0.7423	0.8122	0.8682	0.9108
17	0.1350	0.2009	0.2808	0.3715	0.4677	0.5640	0.6550	0.7363	0.8055	0.8615
18	0.0917	0.1426	0.2081	0.2867	0.3751	0.4686	0.5622	0.6509	0.7307	0.7991
19	0.0606	0.0984	0.1497	0.2148	0.2920	0.3784	0.4695	0.5606	0.6472	0.7255
20	0.0390	0.0661	0.1049	0.1565	0.2211	0.2970	0.3814	0.4703	0.5591	0.6437
21	0.0245	0.0434	0.0716	0.1111	0.1629	0.2270	0.3017	0.3843	0.4710	0.5577
22	0.0151	0.0278	0.0477	0.0769	0.1170	0.1690	0.2325	0.3060	0.3869	0.4716
23	0.0091	0.0174	0.0311	0.0520	0.0821	0.1228	0.1748	0.2377	0.3101	0.3894
24	0.0054	0.0107	0.0198	0.0344	0.0563	0.0871	0.1283	0.1803	0.2426	0.3139
25	0.0031	0.0065	0.0124	0.0223	0.0377	0.0605	0.0920	0.1336	0.1855	0.2473

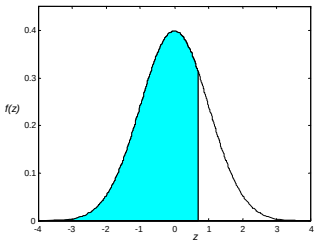
λ \ x	22	23	24	25	26	27	28	29	30	31
16	0.9418	0.9633	0.9777	0.9869	0.9925	0.9959	0.9978	0.9989	0.9994	0.9997
17	0.9047	0.9367	0.9594	0.9748	0.9848	0.9912	0.9950	0.9973	0.9986	0.9993
18	0.8551	0.8989	0.9317	0.9554	0.9718	0.9827	0.9897	0.9941	0.9967	0.9982
19	0.7931	0.8490	0.8933	0.9269	0.9514	0.9687	0.9805	0.9882	0.9930	0.9960
20	0.7206	0.7875	0.8432	0.8878	0.9221	0.9475	0.9657	0.9782	0.9865	0.9919
21	0.6405	0.7160	0.7822	0.8377	0.8826	0.9175	0.9436	0.9626	0.9758	0.9848
22	0.5564	0.6374	0.7117	0.7771	0.8324	0.8775	0.9129	0.9398	0.9595	0.9735
23	0.4723	0.5551	0.6346	0.7077	0.7723	0.8274	0.8726	0.9085	0.9360	0.9564
24	0.3917	0.4728	0.5540	0.6319	0.7038	0.7677	0.8225	0.8679	0.9042	0.9322
25	0.3175	0.3939	0.4734	0.5529	0.6294	0.7002	0.7634	0.8179	0.8633	0.8999

λ \ x	32	33	34	35	36	37	38	39	40	41
16	0.9999	0.9999	1.0000	1.0000	1.0000					
17	0.9996	0.9998	0.9999	1.0000	1.0000	1.0000				
18	0.9990	0.9995	0.9998	0.9999	0.9999	1.0000	1.0000	1.0000		
19	0.9978	0.9988	0.9994	0.9997	0.9998	0.9999	1.0000	1.0000	1.0000	
20	0.9953	0.9973	0.9985	0.9992	0.9996	0.9998	0.9999	0.9999	1.0000	1.0000
21	0.9907	0.9945	0.9968	0.9982	0.9990	0.9995	0.9997	0.9999	0.9999	1.0000
22	0.9831	0.9895	0.9936	0.9962	0.9978	0.9988	0.9993	0.9996	0.9998	0.9999
23	0.9711	0.9813	0.9882	0.9927	0.9956	0.9974	0.9985	0.9992	0.9996	0.9998
24	0.9533	0.9686	0.9794	0.9868	0.9918	0.9950	0.9970	0.9983	0.9990	0.9995
25	0.9285	0.9502	0.9662	0.9775	0.9854	0.9908	0.9943	0.9966	0.9980	0.9988

λ \ x	42	43	44	45	46	47
21	1.0000	1.0000				
22	1.0000	1.0000	1.0000			
23	0.9999	0.9999	1.0000	1.0000	1.0000	
24	0.9997	0.9998	0.9999	1.0000	1.0000	1.0000
25	0.9993	0.9996	0.9998	0.9999	0.9999	1.0000

Tavola 3 – Distribuzione normale standardizzata

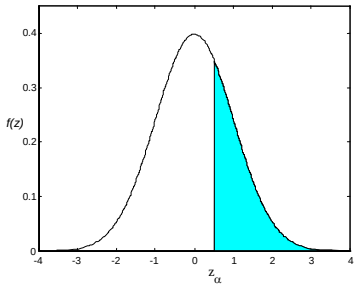
La tavola fornisce il valore dell'area sottesa dalla distribuzione normale standardizzata $f(z)$, tra $-\infty$ e z



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	0.99995	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	0.99997									
5.0	0.9999997									
6.0	0.9999999990									

Tavola 4 – Percentili per la distribuzione normale standardizzata

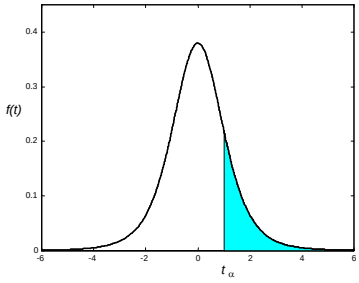
La tavola fornisce i valori di z_α per i quali $P(z > z_\alpha) = \alpha \cdot 100\% = q\%$, per alcuni valori notevoli di q.



$q\%$	z	$q\%$	z	$q\%$	z	$q\%$	z
50	0.000	9	1.341	2.9	1.896	0.4	2.652
45	0.126	8	1.405	2.8	1.911	0.3	2.748
40	0.253	7	1.476	2.7	1.927	0.2	2.878
35	0.385	6	1.555	2.6	1.943	0.1	3.090
30	0.524	5	1.645	2.5	1.960		
29	0.553	4.9	1.655	2.4	1.977	0.09	3.121
28	0.583	4.8	1.665	2.3	1.995	0.08	3.156
27	0.613	4.7	1.675	2.2	2.014	0.07	3.195
26	0.643	4.6	1.685	2.1	2.034	0.06	3.239
25	0.674	4.5	1.695	2.0	2.054	0.05	3.291
24	0.706	4.4	1.706	1.9	2.075	0.04	3.353
23	0.739	4.3	1.717	1.8	2.097	0.03	3.432
22	0.772	4.2	1.728	1.7	2.120	0.02	3.540
21	0.806	4.1	1.739	1.6	2.144	0.01	3.719
20	0.842	4.0	1.751	1.5	2.170	0.005	3.891
19	0.878	3.9	1.762	1.4	2.197	0.001	4.265
18	0.915	3.8	1.774	1.3	2.226	0.0005	4.417
17	0.954	3.7	1.787	1.2	2.257	0.0001	4.753
16	0.994	3.6	1.799	1.1	2.290	0.00005	4.892
15	1.036	3.5	1.812	1.0	2.326	0.00001	5.199
14	1.080	3.4	1.825	0.9	2.366	0.000005	5.327
13	1.126	3.3	1.838	0.8	2.409	0.000001	5.612
12	1.175	3.2	1.852	0.7	2.457	0.0000005	5.731
11	1.227	3.1	1.866	0.6	2.512	0.0000001	5.998
10	1.282	3.0	1.881	0.5	2.576	0.00000005	6.109

Tavola 5 – Distribuzione *t* di Student

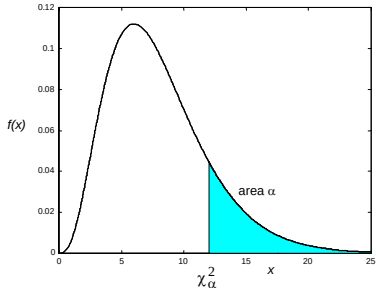
La tavola fornisce i valori di t_α per i quali $P(t > t_\alpha) = \alpha$, per alcuni valori notevoli di α e per il grado di libertà ν.



ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	ν
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
∞	1.282	1.645	1.960	2.326	2.576	∞

Tavola 6 – Distribuzione χ^2

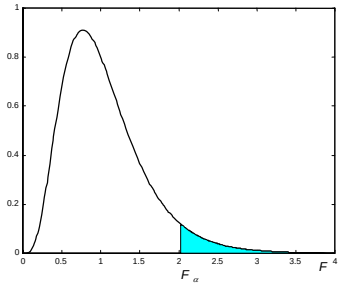
La tavola fornisce i valori di χ^2_α per i quali $P\left(\chi^2 > \chi^2_\alpha\right) = \alpha$, per alcuni valori notevoli di α e per il grado di libertà ν.



ν	$\alpha = 0.995$	$\alpha = 0.99$	$\alpha = 0.975$	$\alpha = 0.95$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	ν
1	0.0000393	0.000157	0.000982	0.00393	3.841	5.024	6.635	7.879	1
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597	2
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750	5
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558	24
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30
40	20.706	22.164	24.433	26.509	55.758	59.342	63.691	66.766	40
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490	50
60	35.535	37.485	40.482	43.188	79.082	83.298	88.379	91.952	60
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215	70
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321	80
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299	90
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169	100

Tavola 7 – Distribuzione *F*

La tavola fornisce i valori di F_α per i quali $P\left(F > F_\alpha\right) = \alpha$, per alcuni valori notevoli di α e per i gradi di libertà ν_1 e ν_2 del numeratore e del denominatore.



$F_{0.25}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1										
	1	2	3	4	5	6	7	8	9	10	12
1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.68
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39
21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38
22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37
23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37
24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36
25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36
26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35
27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35
28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34
29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34
30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34
40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31
50	1.35	1.43	1.41	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30
60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29
80	1.34	1.41	1.40	1.38	1.36	1.34	1.32	1.31	1.30	1.29	1.27
120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26
∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24

Gradi di libertà del denominatore ν_2

$F_{0.25}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1									
Gradi di libertà del denominatore ν_2	15	20	22	24	30	40	50	60	120	∞
1	9.49	9.58	9.61	9.63	9.67	9.71	9.74	9.76	9.80	9.85
2	3.41	3.43	3.43	3.43	3.44	3.45	3.46	3.46	3.47	3.48
3	2.46	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47
4	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.89	1.88	1.88	1.88	1.88	1.88	1.88	1.87	1.87	1.87
6	1.76	1.76	1.76	1.75	1.75	1.75	1.75	1.74	1.74	1.74
7	1.68	1.67	1.67	1.67	1.66	1.66	1.66	1.65	1.65	1.65
8	1.62	1.61	1.61	1.60	1.60	1.59	1.59	1.59	1.58	1.58
9	1.57	1.56	1.56	1.56	1.55	1.54	1.54	1.54	1.53	1.53
10	1.53	1.52	1.52	1.52	1.51	1.51	1.50	1.50	1.49	1.48
11	1.50	1.49	1.49	1.49	1.48	1.47	1.47	1.47	1.46	1.45
12	1.48	1.47	1.46	1.46	1.45	1.45	1.44	1.44	1.43	1.42
13	1.46	1.45	1.44	1.44	1.43	1.42	1.42	1.42	1.41	1.40
14	1.44	1.43	1.42	1.42	1.41	1.41	1.40	1.40	1.39	1.38
15	1.43	1.41	1.41	1.41	1.40	1.39	1.38	1.38	1.37	1.36
16	1.41	1.40	1.39	1.39	1.38	1.37	1.37	1.36	1.35	1.34
17	1.40	1.39	1.38	1.38	1.37	1.36	1.36	1.35	1.34	1.33
18	1.39	1.38	1.37	1.37	1.36	1.35	1.34	1.34	1.33	1.32
19	1.38	1.37	1.36	1.36	1.35	1.34	1.33	1.33	1.32	1.30
20	1.37	1.36	1.35	1.35	1.34	1.33	1.32	1.32	1.31	1.29
21	1.37	1.35	1.35	1.34	1.33	1.32	1.32	1.31	1.30	1.28
22	1.36	1.34	1.34	1.33	1.32	1.31	1.31	1.30	1.29	1.28
23	1.35	1.34	1.33	1.33	1.32	1.31	1.30	1.30	1.28	1.27
24	1.35	1.33	1.33	1.32	1.31	1.30	1.29	1.29	1.28	1.26
25	1.34	1.33	1.32	1.32	1.31	1.29	1.29	1.28	1.27	1.25
26	1.34	1.32	1.32	1.31	1.30	1.29	1.28	1.28	1.26	1.25
27	1.33	1.32	1.31	1.31	1.30	1.28	1.28	1.27	1.26	1.24
28	1.33	1.31	1.31	1.30	1.29	1.28	1.27	1.27	1.25	1.24
29	1.32	1.31	1.30	1.30	1.29	1.27	1.27	1.26	1.25	1.23
30	1.32	1.30	1.30	1.29	1.28	1.27	1.26	1.26	1.24	1.23
40	1.30	1.28	1.27	1.26	1.25	1.24	1.23	1.22	1.21	1.19
50	1.28	1.26	1.25	1.25	1.23	1.22	1.21	1.20	1.19	1.16
60	1.27	1.25	1.24	1.24	1.22	1.21	1.20	1.19	1.17	1.15
80	1.26	1.23	1.23	1.22	1.21	1.19	1.18	1.17	1.15	1.12
120	1.24	1.22	1.21	1.21	1.19	1.18	1.16	1.16	1.13	1.10
∞	1.22	1.19	1.18	1.18	1.16	1.14	1.13	1.12	1.08	1.00

$F_{0.10}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1										
Gradi di libertà del denominatore ν_2	1	2	3	4	5	6	7	8	9	10	12
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.68
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66
80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68	1.63
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55

$F_{0.10}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1									
Gradi di libertà del denominatore ν_2	15	20	22	24	30	40	50	60	120	∞
1	61.22	61.74	61.88	62.00	62.26	62.53	62.69	62.79	63.06	63.33
2	9.42	9.44	9.45	9.45	9.46	9.47	9.47	9.47	9.48	9.49
3	5.20	5.18	5.18	5.18	5.17	5.16	5.15	5.15	5.14	5.13
4	3.87	3.84	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
5	3.24	3.21	3.20	3.19	3.17	3.16	3.15	3.14	3.12	3.10
6	2.87	2.84	2.83	2.82	2.80	2.78	2.77	2.76	2.74	2.72
7	2.63	2.59	2.58	2.58	2.56	2.54	2.52	2.51	2.49	2.47
8	2.46	2.42	2.41	2.40	2.38	2.36	2.35	2.34	2.32	2.29
9	2.34	2.30	2.29	2.28	2.25	2.23	2.22	2.21	2.18	2.16
10	2.24	2.20	2.19	2.18	2.16	2.13	2.12	2.11	2.08	2.06
11	2.17	2.12	2.11	2.10	2.08	2.05	2.04	2.03	2.00	1.97
12	2.10	2.06	2.05	2.04	2.01	1.99	1.97	1.96	1.93	1.90
13	2.05	2.01	1.99	1.98	1.96	1.93	1.92	1.90	1.88	1.85
14	2.01	1.96	1.95	1.94	1.91	1.89	1.87	1.86	1.83	1.80
15	1.97	1.92	1.91	1.90	1.87	1.85	1.83	1.82	1.79	1.76
16	1.94	1.89	1.88	1.87	1.84	1.81	1.79	1.78	1.75	1.72
17	1.91	1.86	1.85	1.84	1.81	1.78	1.76	1.75	1.72	1.69
18	1.89	1.84	1.82	1.81	1.78	1.75	1.74	1.72	1.69	1.66
19	1.86	1.81	1.80	1.79	1.76	1.73	1.71	1.70	1.67	1.63
20	1.84	1.79	1.78	1.77	1.74	1.71	1.69	1.68	1.64	1.61
21	1.83	1.78	1.76	1.75	1.72	1.69	1.67	1.66	1.62	1.59
22	1.81	1.76	1.74	1.73	1.70	1.67	1.65	1.64	1.60	1.57
23	1.80	1.74	1.73	1.72	1.69	1.66	1.64	1.62	1.59	1.55
24	1.78	1.73	1.71	1.70	1.67	1.64	1.62	1.61	1.57	1.53
25	1.77	1.72	1.70	1.69	1.66	1.63	1.61	1.59	1.56	1.52
26	1.76	1.71	1.69	1.68	1.65	1.61	1.59	1.58	1.54	1.50
27	1.75	1.70	1.68	1.67	1.64	1.60	1.58	1.57	1.53	1.49
28	1.74	1.69	1.67	1.66	1.63	1.59	1.57	1.56	1.52	1.48
29	1.73	1.68	1.66	1.65	1.62	1.58	1.56	1.55	1.51	1.47
30	1.72	1.67	1.65	1.64	1.61	1.57	1.55	1.54	1.50	1.46
40	1.66	1.61	1.59	1.57	1.54	1.51	1.48	1.47	1.42	1.38
50	1.63	1.57	1.55	1.54	1.50	1.46	1.44	1.42	1.38	1.33
60	1.60	1.54	1.53	1.51	1.48	1.44	1.41	1.40	1.35	1.29
80	1.57	1.51	1.49	1.48	1.44	1.40	1.38	1.36	1.31	1.24
120	1.55	1.48	1.46	1.45	1.41	1.37	1.34	1.32	1.26	1.19
∞	1.49	1.42	1.40	1.38	1.34	1.30	1.26	1.24	1.17	1.00

$F_{0.05}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1										
Gradi di libertà del denominatore ν_2	1	2	3	4	5	6	7	8	9	10	12
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75

$F_{0.05}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1									
Gradi di libertà del denominatore ν_2	15	20	22	24	30	40	50	60	120	∞
1	245.95	248.01	248.58	249.05	250.10	251.14	251.77	252.20	253.25	254.30
2	19.43	19.45	19.45	19.45	19.46	19.47	19.48	19.48	19.49	19.50
3	8.70	8.66	8.65	8.64	8.62	8.59	8.58	8.57	8.55	8.53
4	5.86	5.80	5.79	5.77	5.75	5.72	5.70	5.69	5.66	5.63
5	4.62	4.56	4.54	4.53	4.50	4.46	4.44	4.43	4.40	4.37
6	3.94	3.87	3.86	3.84	3.81	3.77	3.75	3.74	3.70	3.67
7	3.51	3.44	3.43	3.41	3.38	3.34	3.32	3.30	3.27	3.23
8	3.22	3.15	3.13	3.12	3.08	3.04	3.02	3.01	2.97	2.93
9	3.01	2.94	2.92	2.90	2.86	2.83	2.80	2.79	2.75	2.71
10	2.85	2.77	2.75	2.74	2.70	2.66	2.64	2.62	2.58	2.54
11	2.72	2.65	2.63	2.61	2.57	2.53	2.51	2.49	2.45	2.40
12	2.62	2.54	2.52	2.51	2.47	2.43	2.40	2.38	2.34	2.30
13	2.53	2.46	2.44	2.42	2.38	2.34	2.31	2.30	2.25	2.21
14	2.46	2.39	2.37	2.35	2.31	2.27	2.24	2.22	2.18	2.13
15	2.40	2.33	2.31	2.29	2.25	2.20	2.18	2.16	2.11	2.07
16	2.35	2.28	2.25	2.24	2.19	2.15	2.12	2.11	2.06	2.01
17	2.31	2.23	2.21	2.19	2.15	2.10	2.08	2.06	2.01	1.96
18	2.27	2.19	2.17	2.15	2.11	2.06	2.04	2.02	1.97	1.92
19	2.23	2.16	2.13	2.11	2.07	2.03	2.00	1.98	1.93	1.88
20	2.20	2.12	2.10	2.08	2.04	1.99	1.97	1.95	1.90	1.84
21	2.18	2.10	2.07	2.05	2.01	1.96	1.94	1.92	1.87	1.81
22	2.15	2.07	2.05	2.03	1.98	1.94	1.91	1.89	1.84	1.78
23	2.13	2.05	2.02	2.01	1.96	1.91	1.88	1.86	1.81	1.76
24	2.11	2.03	2.00	1.98	1.94	1.89	1.86	1.84	1.79	1.73
25	2.09	2.01	1.98	1.96	1.92	1.87	1.84	1.82	1.77	1.71
26	2.07	1.99	1.97	1.95	1.90	1.85	1.82	1.80	1.75	1.69
27	2.06	1.97	1.95	1.93	1.88	1.84	1.81	1.79	1.73	1.67
28	2.04	1.96	1.93	1.91	1.87	1.82	1.79	1.77	1.71	1.65
29	2.03	1.94	1.92	1.90	1.85	1.81	1.77	1.75	1.70	1.64
30	2.01	1.93	1.91	1.89	1.84	1.79	1.76	1.74	1.68	1.62
40	1.92	1.84	1.81	1.79	1.74	1.69	1.66	1.64	1.58	1.51
50	1.87	1.78	1.76	1.74	1.69	1.63	1.60	1.58	1.51	1.44
60	1.84	1.75	1.72	1.70	1.65	1.59	1.56	1.53	1.47	1.39
80	1.79	1.70	1.68	1.65	1.60	1.54	1.51	1.48	1.41	1.33
120	1.75	1.66	1.63	1.61	1.55	1.50	1.46	1.43	1.35	1.26
∞	1.67	1.57	1.54	1.52	1.46	1.39	1.35	1.32	1.22	1.00

$F_{0.025}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1										
Gradi di libertà del denominatore ν_2	1	2	3	4	5	6	7	8	9	10	12
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	976.71
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.22
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.11
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94

$F_{0.025}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1									
	15	20	22	24	30	40	50	60	120	∞
1	984.87	993.10	995.36	997.25	1001.41	1005.60	1008.12	1009.80	1014.02	1018.25
2	39.43	39.45	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50
3	14.25	14.17	14.14	14.12	14.08	14.04	14.01	13.99	13.95	13.90
4	8.66	8.56	8.53	8.51	8.46	8.41	8.38	8.36	8.31	8.26
5	6.43	6.33	6.30	6.28	6.23	6.18	6.14	6.12	6.07	6.02
6	5.27	5.17	5.14	5.12	5.07	5.01	4.98	4.96	4.90	4.85
7	4.57	4.47	4.44	4.41	4.36	4.31	4.28	4.25	4.20	4.14
8	4.10	4.00	3.97	3.95	3.89	3.84	3.81	3.78	3.73	3.67
9	3.77	3.67	3.64	3.61	3.56	3.51	3.47	3.45	3.39	3.33
10	3.52	3.42	3.39	3.37	3.31	3.26	3.22	3.20	3.14	3.08
11	3.33	3.23	3.20	3.17	3.12	3.06	3.03	3.00	2.94	2.88
12	3.18	3.07	3.04	3.02	2.96	2.91	2.87	2.85	2.79	2.72
13	3.05	2.95	2.92	2.89	2.84	2.78	2.74	2.72	2.66	2.60
14	2.95	2.84	2.81	2.79	2.73	2.67	2.64	2.61	2.55	2.49
15	2.86	2.76	2.73	2.70	2.64	2.59	2.55	2.52	2.46	2.40
16	2.79	2.68	2.65	2.63	2.57	2.51	2.47	2.45	2.38	2.32
17	2.72	2.62	2.59	2.56	2.50	2.44	2.41	2.38	2.32	2.25
18	2.67	2.56	2.53	2.50	2.44	2.38	2.35	2.32	2.26	2.19
19	2.62	2.51	2.48	2.45	2.39	2.33	2.30	2.27	2.20	2.13
20	2.57	2.46	2.43	2.41	2.35	2.29	2.25	2.22	2.16	2.09
21	2.53	2.42	2.39	2.37	2.31	2.25	2.21	2.18	2.11	2.04
22	2.50	2.39	2.36	2.33	2.27	2.21	2.17	2.14	2.08	2.00
23	2.47	2.36	2.33	2.30	2.24	2.18	2.14	2.11	2.04	1.97
24	2.44	2.33	2.30	2.27	2.21	2.15	2.11	2.08	2.01	1.94
25	2.41	2.30	2.27	2.24	2.18	2.12	2.08	2.05	1.98	1.91
26	2.39	2.28	2.24	2.22	2.16	2.09	2.05	2.03	1.95	1.88
27	2.36	2.25	2.22	2.19	2.13	2.07	2.03	2.00	1.93	1.85
28	2.34	2.23	2.20	2.17	2.11	2.05	2.01	1.98	1.91	1.83
29	2.32	2.21	2.18	2.15	2.09	2.03	1.99	1.96	1.89	1.81
30	2.31	2.20	2.16	2.14	2.07	2.01	1.97	1.94	1.87	1.79
40	2.18	2.07	2.03	2.01	1.94	1.88	1.83	1.80	1.72	1.64
50	2.11	1.99	1.96	1.93	1.87	1.80	1.75	1.72	1.64	1.55
60	2.06	1.94	1.91	1.88	1.82	1.74	1.70	1.67	1.58	1.48
80	2.00	1.88	1.85	1.82	1.75	1.68	1.63	1.60	1.51	1.40
120	1.94	1.82	1.79	1.76	1.69	1.61	1.56	1.53	1.43	1.31
∞	1.83	1.71	1.67	1.64	1.57	1.48	1.43	1.39	1.27	1.00

Gradi di libertà del denominatore ν_2

$F_{0.01}(\nu_1,\nu_2)$

ν_2 \ ν_1		Gradi di libertà del numeratore ν_1										
		1	2	3	4	5	6	7	8	9	10	12
Gradi di libertà del denominatore ν_2	1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66
	50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50
	80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34
	∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18

$F_{0.01}(\nu_1,\nu_2)$

ν_2 \ ν_1		Gradi di libertà del numeratore ν_1									
		15	20	22	24	30	40	50	60	120	∞
Gradi di libertà del denominatore ν_2	1	6157.28	6208.73	6222.84	6234.63	6260.65	6286.78	6302.52	6313.03	6339.39	6365.84
	2	99.43	99.45	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
	3	26.87	26.69	26.64	26.60	26.50	26.41	26.35	26.32	26.22	26.13
	4	14.20	14.02	13.97	13.93	13.84	13.75	13.69	13.65	13.56	13.46
	5	9.72	9.55	9.51	9.47	9.38	9.29	9.24	9.20	9.11	9.02
	6	7.56	7.40	7.35	7.31	7.23	7.14	7.09	7.06	6.97	6.88
	7	6.31	6.16	6.11	6.07	5.99	5.91	5.86	5.82	5.74	5.65
	8	5.52	5.36	5.32	5.28	5.20	5.12	5.07	5.03	4.95	4.86
	9	4.96	4.81	4.77	4.73	4.65	4.57	4.52	4.48	4.40	4.31
	10	4.56	4.41	4.36	4.33	4.25	4.17	4.12	4.08	4.00	3.91
	11	4.25	4.10	4.06	4.02	3.94	3.86	3.81	3.78	3.69	3.60
	12	4.01	3.86	3.82	3.78	3.70	3.62	3.57	3.54	3.45	3.36
	13	3.82	3.66	3.62	3.59	3.51	3.43	3.38	3.34	3.25	3.17
	14	3.66	3.51	3.46	3.43	3.35	3.27	3.22	3.18	3.09	3.00
	15	3.52	3.37	3.33	3.29	3.21	3.13	3.08	3.05	2.96	2.87
	16	3.41	3.26	3.22	3.18	3.10	3.02	2.97	2.93	2.84	2.75
	17	3.31	3.16	3.12	3.08	3.00	2.92	2.87	2.83	2.75	2.65
	18	3.23	3.08	3.03	3.00	2.92	2.84	2.78	2.75	2.66	2.57
	19	3.15	3.00	2.96	2.92	2.84	2.76	2.71	2.67	2.58	2.49
	20	3.09	2.94	2.90	2.86	2.78	2.69	2.64	2.61	2.52	2.42
	21	3.03	2.88	2.84	2.80	2.72	2.64	2.58	2.55	2.46	2.36
	22	2.98	2.83	2.78	2.75	2.67	2.58	2.53	2.50	2.40	2.31
	23	2.93	2.78	2.74	2.70	2.62	2.54	2.48	2.45	2.35	2.26
	24	2.89	2.74	2.70	2.66	2.58	2.49	2.44	2.40	2.31	2.21
	25	2.85	2.70	2.66	2.62	2.54	2.45	2.40	2.36	2.27	2.17
	26	2.81	2.66	2.62	2.58	2.50	2.42	2.36	2.33	2.23	2.13
	27	2.78	2.63	2.59	2.55	2.47	2.38	2.33	2.29	2.20	2.10
	28	2.75	2.60	2.56	2.52	2.44	2.35	2.30	2.26	2.17	2.06
	29	2.73	2.57	2.53	2.49	2.41	2.33	2.27	2.23	2.14	2.03
	30	2.70	2.55	2.51	2.47	2.39	2.30	2.25	2.21	2.11	2.01
	40	2.52	2.37	2.33	2.29	2.20	2.11	2.06	2.02	1.92	1.80
	50	2.42	2.27	2.22	2.18	2.10	2.01	1.95	1.91	1.80	1.68
	60	2.35	2.20	2.15	2.12	2.03	1.94	1.88	1.84	1.73	1.60
	80	2.27	2.12	2.07	2.03	1.94	1.85	1.79	1.75	1.63	1.49
	120	2.19	2.03	1.99	1.95	1.86	1.76	1.70	1.66	1.53	1.38
	∞	2.04	1.88	1.83	1.79	1.70	1.59	1.53	1.47	1.32	1.00

$F_{0.005}(\nu_1,\nu_2)$

ν_2 \ ν_1		Gradi di libertà del numeratore ν_1										
		1	2	3	4	5	6	7	8	9	10	12
Gradi di libertà del denominatore ν_2	1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426
	2	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40	199.42
	3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39
	4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70
	5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38
	6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03
	7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18
	8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01
	9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23
	10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66
	11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24
	12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91
	13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64
	14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43
	15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25
	16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10
	17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97
	18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86
	19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76
	20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68
	21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60
	22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54
	23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47
	24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42
	25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37
	26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33
	27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28
	28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25
	29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21
	30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18
	40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95
	50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.82
	60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74
	80	8.33	5.67	4.61	4.03	3.65	3.39	3.19	3.03	2.91	2.80	2.64
	120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54
	∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.75	2.62	2.52	2.36

$F_{0.005}(\nu_1,\nu_2)$

ν_2 \ ν_1	Gradi di libertà del numeratore ν_1									
	15	20	22	24	30	40	50	60	120	∞
1	24630	24836	24892	24940	25044	25148	25211	25253	25358.57	25464
2	199.43	199.45	199.45	199.46	199.47	199.47	199.48	199.48	199.49	199.50
3	43.08	42.78	42.69	42.62	42.47	42.31	42.21	42.15	41.99	41.83
4	20.44	20.17	20.09	20.03	19.89	19.75	19.67	19.61	19.47	19.33
5	13.15	12.90	12.84	12.78	12.66	12.53	12.45	12.40	12.27	12.14
6	9.81	9.59	9.53	9.47	9.36	9.24	9.17	9.12	9.00	8.88
7	7.97	7.75	7.69	7.64	7.53	7.42	7.35	7.31	7.19	7.08
8	6.81	6.61	6.55	6.50	6.40	6.29	6.22	6.18	6.06	5.95
9	6.03	5.83	5.78	5.73	5.62	5.52	5.45	5.41	5.30	5.19
10	5.47	5.27	5.22	5.17	5.07	4.97	4.90	4.86	4.75	4.64
11	5.05	4.86	4.80	4.76	4.65	4.55	4.49	4.45	4.34	4.23
12	4.72	4.53	4.48	4.43	4.33	4.23	4.17	4.12	4.01	3.90
13	4.46	4.27	4.22	4.17	4.07	3.97	3.91	3.87	3.76	3.65
14	4.25	4.06	4.01	3.96	3.86	3.76	3.70	3.66	3.55	3.44
15	4.07	3.88	3.83	3.79	3.69	3.58	3.52	3.48	3.37	3.26
16	3.92	3.73	3.68	3.64	3.54	3.44	3.37	3.33	3.22	3.11
17	3.79	3.61	3.56	3.51	3.41	3.31	3.25	3.21	3.10	2.98
18	3.68	3.50	3.45	3.40	3.30	3.20	3.14	3.10	2.99	2.87
19	3.59	3.40	3.35	3.31	3.21	3.11	3.04	3.00	2.89	2.78
20	3.50	3.32	3.27	3.22	3.12	3.02	2.96	2.92	2.81	2.69
21	3.43	3.24	3.19	3.15	3.05	2.95	2.88	2.84	2.73	2.61
22	3.36	3.18	3.12	3.08	2.98	2.88	2.82	2.77	2.66	2.55
23	3.30	3.12	3.06	3.02	2.92	2.82	2.76	2.71	2.60	2.48
24	3.25	3.06	3.01	2.97	2.87	2.77	2.70	2.66	2.55	2.43
25	3.20	3.01	2.96	2.92	2.82	2.72	2.65	2.61	2.50	2.38
26	3.15	2.97	2.92	2.87	2.77	2.67	2.61	2.56	2.45	2.33
27	3.11	2.93	2.88	2.83	2.73	2.63	2.57	2.52	2.41	2.29
28	3.07	2.89	2.84	2.79	2.69	2.59	2.53	2.48	2.37	2.25
29	3.04	2.86	2.80	2.76	2.66	2.56	2.49	2.45	2.33	2.21
30	3.01	2.82	2.77	2.73	2.63	2.52	2.46	2.42	2.30	2.18
40	2.78	2.60	2.55	2.50	2.40	2.30	2.23	2.18	2.06	1.93
50	2.65	2.47	2.42	2.37	2.27	2.16	2.10	2.05	1.93	1.79
60	2.57	2.39	2.33	2.29	2.19	2.08	2.01	1.96	1.83	1.69
80	2.47	2.29	2.23	2.19	2.08	1.97	1.90	1.85	1.72	1.56
120	2.37	2.19	2.13	2.09	1.98	1.87	1.80	1.75	1.61	1.43
∞	2.19	2.00	1.95	1.90	1.79	1.67	1.59	1.53	1.37	1.04

Gradi di libertà del denominatore ν_2

Appendice B. Formulario

Valor medio campionario

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Varianza campionaria

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i{}^2 - n\bar{x}^2\right]$$

Scarto quadratico medio campionario (deviazione standard)

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Dati raggruppati

n = n° dati k = n° classi m_i = valori centrali f_i = frequenze assolute

Valor medio campionario

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i f_i$$

Varianza campionaria

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(m_i - \bar{x}\right)^2 f_i = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i m_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i m_i\right)^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i m_i{}^2 - n\bar{x}^2\right]$$

Covarianza - Coefficiente correlazione lineare

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}\right) \qquad r = \frac{S_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$$

Retta di regressione

$$y = Ax + B \qquad E = \sum_{i=1}^{n}\left(Ax_i + B - y_i\right)^2$$

$$\begin{cases} A\sum_{i=1}^{n} x_i^2 + B\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ A\sum_{i=1}^{n} x_i + nB = \sum_{i=1}^{n} y_i \end{cases} \qquad \text{oppure} \qquad \begin{cases} A = \dfrac{S_{xy}}{s_x^2} \\ B = \bar{y} - A\bar{x} \end{cases}$$

Parabola dei minimi quadrati

$$y = Ax^2 + Bx + C \qquad E = \sum_{i=1}^{n}\left(Ax_i^2 + Bx_i + C - y_i\right)^2$$

$$\begin{cases} A\sum_{i=1}^{n} x_i^4 + B\sum_{i=1}^{n} x_i^3 + C\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i \\ A\sum_{i=1}^{n} x_i^3 + B\sum_{i=1}^{n} x_i^2 + C\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i\, y_i \\ A\sum_{i=1}^{n} x_i^2 + B\sum_{i=1}^{n} x_i + nC = \sum_{i=1}^{n} y_i \end{cases}$$

Linearizzazione

Funzione $y = f(x)$	Forma linearizzata $Y = AX + B$	Cambiamenti di variabili e costanti
$y = C \cdot x^A$	$\ln y = A \ln x + \ln C$	$X = \ln x \quad Y = \ln y$ $C = e^B$
$y = C \cdot e^{Ax}$	$\ln y = A \cdot x + \ln C$	$X = x \quad Y = \ln y$ $C = e^B$
$y = A \ln x + B$	$y = A \ln x + B$	$X = \ln x \quad Y = y$
$y = \frac{A}{x} + B$	$y = A\frac{1}{x} + B$	$X = \frac{1}{x} \quad Y = y$
$y = \frac{1}{Ax + B}$	$\frac{1}{y} = Ax + B$	$X = x \quad Y = \frac{1}{y}$
$y = \frac{x}{A + Bx}$	$\frac{1}{y} = A\frac{1}{x} + B$	$X = \frac{1}{x} \quad Y = \frac{1}{y}$
$y = \frac{D}{x + C}$	$y = -\frac{1}{C}(xy) + \frac{D}{C}$	$X = xy \quad Y = y$ $C = -\frac{1}{A} \quad D = -\frac{B}{A}$
$y = \frac{L}{1 + Ce^{Ax}}$	$\ln\left(\frac{L}{y} - 1\right) = Ax + \ln C$	$X = x \quad Y = \ln\left(\frac{L}{y} - 1\right)$ $C = e^B \quad L =$ costante assegnata
$y = \frac{1}{B + Ae^{-x}}$	$\frac{1}{y} = Ae^{-x} + B$	$X = e^{-x} \quad Y = \frac{1}{y}$

Disposizioni con ripetizione

$$D_{n,k}^{(r)} = n^k$$

Disposizioni semplici

$$D_{n,k} = \frac{n!}{(n-k)!}$$

Permutazioni

$$P_n = n! \qquad P_{n,n_1,n_2,\ldots,n_k} = \frac{n!}{n_1! n_2! \ldots . n_k!}$$

Combinazioni

$$\binom{n}{k} = C_{n,k} = \frac{D_{n,k}}{k!} = \frac{n!}{k!(n-k)!}$$

Regola additiva della probabilità

$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A} \cap \mathrm{B})$

Probabilità condizionata

$$P(\mathrm{B} \mid \mathrm{A}) = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{A})} \qquad P(\mathrm{A}) \neq 0$$

$$P(\mathrm{A} \mid \mathrm{B}) = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{B})} \qquad P(\mathrm{B}) \neq 0$$

Eventi indipendenti - Regola di moltiplicazione

$P(\mathrm{B} \mid \mathrm{A}) = P(\mathrm{B}) \qquad P(\mathrm{A} \mid \mathrm{B}) = P(\mathrm{A})$

$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}) \cdot P(\mathrm{B})$

Probabilità totale

$$P(\mathrm{A}) = P(\mathrm{A} \mid \mathrm{B}_1) \cdot P(\mathrm{B}_1) + P(\mathrm{A} \mid \mathrm{B}_2) \cdot P(\mathrm{B}_2) + \ldots + P(\mathrm{A} \mid \mathrm{B}_n) \cdot P(\mathrm{B}_n) = \sum_{i=1}^{n} P(\mathrm{A} \mid \mathrm{B}_i) \cdot P(\mathrm{B}_i)$$

Teorema di Bayes

$$P(\mathrm{B}_k \mid \mathrm{A}) = \frac{P(\mathrm{A} \mid \mathrm{B}_k) \cdot P(\mathrm{B}_k)}{\sum_{i=1}^{n} P(\mathrm{A} \mid \mathrm{B}_i) \cdot P(\mathrm{B}_i)} \qquad \text{per ogni } k$$

Parametri di una distribuzione – Valor medio e varianza

Caso discreto

$$\begin{cases} \mu = \mathrm{E}(X) = \sum_{i=1}^{n} x_i \; P(X = x_i) = \sum_{i=1}^{n} x_i f(x_i) \\ \sigma^2 = \mathrm{var}(X) = \mathrm{E}\left[(X-\mu)^2\right] = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i) = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2 \end{cases}$$

Caso continuo

$$\begin{cases} \mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x \, f(x)\, dx \\ \sigma^2 = \mathrm{var}(X) = \mathrm{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2 \end{cases}$$

Proprietà di valor medio e varianza $(a,b \in \mathbf{R})$

$$E(aX+b) = aE(X) + b$$

$$\mathrm{var}(aX+b) = a^2 \mathrm{var}(X)$$

$$E(aX+bY) = aE(X) + bE(\mathrm{Y})$$

$$\mathrm{var}\,(aX+bY) = a^2 \mathrm{var}\,(X) + b^2 \mathrm{var}\,(Y)$$

Variabile standardizzata

$$Z = \frac{X-\mu}{\sigma} \qquad \mu = E(Z) = 0 \qquad \sigma^2 = \mathrm{var}\,(Z) = 1$$

Disuguaglianza di Chebishev

$$P\left(|X-\mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2} \qquad P\left(|X-\mu| \leq \varepsilon\right) \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

Distribuzione binomiale o di Bernoulli

$$f(x) = \mathrm{P}(X = x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n \qquad \binom{n}{x} = \frac{n!}{x!\;(n-x)!}$$

$$F(x) = \mathrm{P}(X \leq x) = B(x; n, p) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mu = n\,p \qquad \sigma^2 = np\,(1-p)$$

Proprietà distribuzione binomiale

$$P(X < x) = P(X \leq x-1)$$

$$P(X > x) = 1 - P(X \leq x)$$

$$P(X \geq x) = 1 - P(X \leq x-1)$$

$$P(X = x) = P(X \leq x) - P(X \leq x-1)$$

Relazione di ricorrenza

$$P(X = x+1) = \frac{n-x}{x+1} \cdot \frac{p}{1-p} \cdot P(X = x)$$

Distribuzione di Poisson

$$f(x;\lambda) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots.$$

$$F(x;\lambda) = P(X \le x) = \sum_{k=0}^{x} f(k;\lambda) = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!}$$

$$\mu = \lambda \qquad \sigma^2 = \lambda$$

Proprietà distribuzione di Poisson

$$f(x;\lambda) = F(x;\lambda) - F(x-1;\lambda)$$

Relazione di ricorrenza

$$P(X = x+1) = \frac{\lambda}{x+1} P(X = x)$$

Distribuzione normale o di Gauss

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

$$F(x) = \mathrm{P}(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

Distribuzione normale standardizzata

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad -\infty < z < \infty$$

$$F(z) = \mathrm{P}(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

Proprietà distribuzione normale

$$P(-\infty < Z < \infty) = 1$$

$$P(-\infty < Z < 0) = P(0 < Z < \infty) = F(0) = \frac{1}{2}$$

$$P(Z \le -z) = F(-z) = 1 - F(z)$$

$$P(z_1 \le Z \le z_2) = F(z_2) - F(z_1)$$

$$P(-z_1 \le Z \le 0) = P(0 \le Z \le z_1)$$

Approssimazione distribuzione binomiale con distribuzione normale

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \qquad np \ge 5 \qquad n(1-p) \ge 5$$

Approssimazione distribuzione di Poisson con distribuzione normale

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \qquad \lambda \ge 10$$

Distribuzione uniforme

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{altrimenti} \end{cases}$$

$$F(x) = P(X \le x) = \begin{cases} 0 & x \le a \\ \dfrac{x-a}{b-a} & a < x < b \\ 1 & x \ge b \end{cases}$$

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

Distribuzione *t* di Student

$$T = \frac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}}$$ grado di libertà $\nu = n - 1$

Distribuzione χ^2

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2}$$ grado di libertà $\nu = n - 1$

Distribuzione *F*

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$ gradi di libertà $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}$$

Intervallo di confidenza per la media, con grado di fiducia (1 – α) 100% (varianza nota)

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

grado di fiducia 90% $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

grado di fiducia 95% $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

grado di fiducia 99% $z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$

$$E = \max\left|\overline{X} - \mu\right| = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$ grado di fiducia $(1-\alpha)100\%$ $$n \ge \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2$$

Intervallo di confidenza per la media, con grado di fiducia (1 – α) 100% (varianza incognita)

$$\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

grado di libertà $\nu = n - 1$

grado di fiducia 90% $t_{\frac{\alpha}{2}} = t_{0.05}$

grado di fiducia 95% $t_{\frac{\alpha}{2}} = t_{0.025}$

grado di fiducia 99% $t_{\frac{\alpha}{2}} = t_{0.005}$

Intervallo di confidenza per la proporzione, con grado di fiducia (1 – α) 100%

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$E = \max\left|\hat{P} - p\right| = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \qquad \text{grado di fiducia } (1-\alpha)100\%$$

$$n \geq p(1-p)\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \qquad n \geq \frac{1}{4}\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2$$

Intervallo di confidenza per la differenza fra due medie, con grado di fiducia (1 – α) 100% (varianze note)

$$\bar{x}_1 - \bar{x}_2 - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Intervallo di confidenza per la differenza fra due medie, con grado di fiducia (1 – α) 100% (varianze incognite)

$$\bar{x}_1 - \bar{x}_2 - t_{\frac{\alpha}{2}} \cdot \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + t_{\frac{\alpha}{2}} \cdot \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

grado di libertà $\nu = n_1 + n_2 - 2$

stima congiunta della varianza $s^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$

Intervallo di confidenza per la differenza fra due proporzioni, con grado di fiducia (1 – α) 100%

$$(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Intervallo di confidenza per la varianza, con grado di fiducia (1 – α) 100%

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

grado di libertà $\nu = n - 1$

grado di fiducia 90% $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.05}$ $\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.95}$

grado di fiducia 95% $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025}$ $\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975}$

grado di fiducia 99% $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005}$ $\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995}$

Intervallo di confidenza per lo scarto quadratico medio, con grado di fiducia (1 – α) 100% , $n \geq 30$

$$\frac{s}{1 + \frac{z_{\frac{\alpha}{2}}}{\sqrt{2n}}} < \sigma < \frac{s}{1 - \frac{z_{\frac{\alpha}{2}}}{\sqrt{2n}}}$$

Intervallo di confidenza per il rapporto di due varianze, con grado di fiducia $(1-\alpha)$ 100%

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}} \qquad F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2) = \frac{1}{F_{\frac{\alpha}{2}}(\nu_2,\nu_1)}$$

gradi di libertà $\nu_1 = n_1 - 1$ $\nu_2 = n_2 - 1$

Test di ipotesi sulla media (varianza nota)

Statistica test $$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Test	Ipot. nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu \le \mu_0$	$\mu > \mu_0$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$\mu \ge \mu_0$	$\mu < \mu_0$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$\mu = \mu_0$	$\mu \ne \mu_0$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Test di ipotesi sulla media (varianza incognita)

Statistica test $$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$ grado di libertà $\nu = n - 1$

Test	Ipot. nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu \le \mu_0$	$\mu > \mu_0$	0.01	$t_\alpha = t_{0.01}$	$T > t_{0.01}$
			0.05	$t_\alpha = t_{0.05}$	$T > t_{0.05}$
una coda	$\mu \ge \mu_0$	$\mu < \mu_0$	0.01	$t_\alpha = -t_{0.01}$	$T < -t_{0.01}$
			0.05	$t_\alpha = -t_{0.05}$	$T < -t_{0.05}$
due code	$\mu = \mu_0$	$\mu \ne \mu_0$	0.01	$t_{\frac{\alpha}{2}} = t_{0.005}$ $t_{\frac{\alpha}{2}} = -t_{0.005}$	$T > t_{0.005}$ $T < -t_{0.005}$
			0.05	$t_{\frac{\alpha}{2}} = t_{0.025}$ $t_{\frac{\alpha}{2}} = -t_{0.025}$	$T > t_{0.025}$ $T < -t_{0.025}$

Test di ipotesi sulla proporzione

Statistica test $$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$p \le p_0$	$p > p_0$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$p \ge p_0$	$p < p_0$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$p = p_0$	$p \ne p_0$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Test di ipotesi sulla differenza fra due medie (varianze note)

Statistica test $$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu_1 - \mu_2 \le d$	$\mu_1 - \mu_2 > d$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$\mu_1 - \mu_2 \ge d$	$\mu_1 - \mu_2 < d$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \ne d$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Test di ipotesi sulla differenza fra due medie (varianze incognite)

Statistica test $$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ grado di libertà $$\nu = n_1 + n_2 - 2$$

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\mu_1 - \mu_2 \le d$	$\mu_1 - \mu_2 > d$	0.01	$t_\alpha = t_{0.01}$	$T > t_{0.01}$
			0.05	$t_\alpha = t_{0.05}$	$T > t_{0.05}$
una coda	$\mu_1 - \mu_2 \ge d$	$\mu_1 - \mu_2 < d$	0.01	$t_\alpha = -t_{0.01}$	$T < -t_{0.01}$
			0.05	$t_\alpha = -t_{0.05}$	$T < -t_{0.05}$
due code	$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \ne d$	0.01	$t_{\frac{\alpha}{2}} = t_{0.005}$ $t_{\frac{\alpha}{2}} = -t_{0.005}$	$T > t_{0.005}$ $T < -t_{0.005}$
			0.05	$t_{\frac{\alpha}{2}} = t_{0.025}$ $t_{\frac{\alpha}{2}} = -t_{0.025}$	$T > t_{0.025}$ $T < -t_{0.025}$

Test di ipotesi sulla differenza fra due proporzioni

Statistica test $$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$p_1 \le p_2$	$p_1 > p_2$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$p_1 \ge p_2$	$p_1 < p_2$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$p_1 = p_2$	$p_1 \ne p_2$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Test di ipotesi sulla varianza

Statistica test $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$ grado di libertà $\nu = n - 1$

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\sigma^2 \le \sigma_0^2$	$\sigma^2 > \sigma_0^2$	0.01	$\chi^2_\alpha = \chi^2_{0.01}$	$\chi^2 > \chi^2_{0.01}$
			0.05	$\chi^2_\alpha = \chi^2_{0.05}$	$\chi^2 > \chi^2_{0.05}$
una coda	$\sigma^2 \ge \sigma_0^2$	$\sigma^2 < \sigma_0^2$	0.01	$\chi^2_{1-\alpha} = \chi^2_{0.99}$	$\chi^2 < \chi^2_{0.99}$
			0.05	$\chi^2_{1-\alpha} = \chi^2_{0.95}$	$\chi^2 < \chi^2_{0.95}$
due code	$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	0.01	$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995}$ $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005}$	$\chi^2 < \chi^2_{0.995}$ $\chi^2 > \chi^2_{0.005}$
			0.05	$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975}$ $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025}$	$\chi^2 < \chi^2_{0.975}$ $\chi^2 > \chi^2_{0.025}$

Test di ipotesi sulla varianza, $n \ge 30$

Statistica test $Z = \dfrac{S - \sigma_0}{\dfrac{\sigma_0}{\sqrt{2n}}}$

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\sigma^2 \le \sigma_0^2$	$\sigma^2 > \sigma_0^2$	0.01	2.326	$Z > 2.326$
			0.05	1.645	$Z > 1.645$
una coda	$\sigma^2 \ge \sigma_0^2$	$\sigma^2 < \sigma_0^2$	0.01	–2.326	$Z < -2.326$
			0.05	–1.645	$Z < -1.645$
due code	$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	0.01	–2.576 e 2.576	$Z < -2.576$ $Z > 2.576$
			0.05	–1.96 e 1.96	$Z < -1.96$ $Z > 1.96$

Test di ipotesi sul rapporto di due varianze

Statistica test $F = \dfrac{S_1^2}{S_2^2}$

gradi di libertà $\nu_1 = n_1 - 1 \qquad \nu_2 = n_2 - 1$

$$F_{1-\frac{\alpha}{2}}(\nu_1, \nu_2) = \frac{1}{F_{\frac{\alpha}{2}}(\nu_2, \nu_1)}$$

Test	Ipotesi nulla H_0	Ipot. altern. H_1	Liv. signif. α	Valori critici	Reg. rifiuto
una coda	$\sigma_1^2 \le \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	0.01	$F_\alpha = F_{0.01}$	$F > F_{0.01}$
			0.05	$F_\alpha = F_{0.05}$	$F > F_{0.05}$
una coda	$\sigma_1^2 \ge \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	0.01	$F_{1-\alpha} = F_{0.99}$	$F < F_{0.99}$
			0.05	$F_{1-\alpha} = F_{0.95}$	$F < F_{0.95}$
due code	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \ne \sigma_2^2$	0.01	$F_{1-\frac{\alpha}{2}} = F_{0.995}$ $F_{\frac{\alpha}{2}} = F_{0.005}$	$F < F_{0.995}$ $F > F_{0.005}$
			0.05	$F_{1-\frac{\alpha}{2}} = F_{0.975}$ $F_{\frac{\alpha}{2}} = F_{0.025}$	$F < F_{0.975}$ $F > F_{0.025}$

Test chi-quadro di adattamento

Ipotesi nulla H_0 : i dati si adattano alla distribuzione teorica

Ipotesi alternativa H_1 : i dati non si adattano alla distribuzione teorica

Statistica test $\chi^2 = \sum_{i=1}^{k} \dfrac{(O_i - A_i)^2}{A_i}$

Regione di rifiuto $\chi^2 > \chi_\alpha^2$ Grado di libertà $\nu = k - 1 - m$

Test chi-quadro di indipendenza

Ipotesi nulla H_0 : indipendenza

Ipotesi alternativa H_1 : dipendenza

Statistica test $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \dfrac{(O_{ij} - A_{ij})^2}{A_{ij}}$

Regione di rifiuto $\chi^2 > \chi_\alpha^2$ Grado di libertà $\nu = (r-1)\cdot(c-1)$

Appendice C. Bibliografia

1. Bramanti M., *Calcolo delle Probabilità e Statistica per il Corso di Diploma in Ingegneria, Teoria ed esercizi*, Progetto Leonardo, 1997
2. Cavalli-Sforza L. L., *Analisi statistica per medici e biologi. Una introduzione elementare*, Bollati Boringhieri, 1992
3. Cerasoli M., Tomassetti G., *Elementi di Statistica. Introduzione alla matematica dell'incerto*, Zanichelli, 1987
4. Cerasoli A. M., Cerasoli M., *Elementi di Calcolo delle Probabilità. Introduzione alla matematica dell'incerto*, Zanichelli, 1987
5. Daniel W.W., *Biostatistica. Concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria*, EdiSES, 1996
6. Freund J. E., Simon G.A., *Modern Elementary Statistics*, Prentice-Hall Int. Ed., 1992
7. Freund J. E., Walpole R. E., *Mathematical Statistics*, Prentice-Hall Int. Inc., 1987.
8. Freund R.J., Wilson W.J., *Metodi statistici*, Piccin, 2001
9. Johnson R.A., *Miller and Freund's Probability and Statistics for Engineer*, Prentice-Hall Int. Inc., 1994.
10. Levine D. M., Krehbiel T.C., Berenson M. L., *Statistica*, Apogeo, 2002
11. Montgomery D.C., Runger G.C., *Applied Statistics and Probability for Engineers*, John Wiley & Sons, 1999
12. Rosner B., *Fundamentals of Biostatistics*, Wadsworth Publishing Company, ITP, 1995
13. Rosner B., *Study Guide for Fundamentals of Biostatistics*, Wadsworth Publishing Company, ITP, 1995
14. Ross S.M., *Probabilità e Statistica per l'ingegneria e le scienze*, Apogeo, 2003
15. Rossi C., Serio G., *La metodologia statistica nelle applicazioni biomediche*, Springer-Verlag, 1990
16. Sokal R. R., Rohlf F. J., *Introduction to Biostatistics*, W. H. Freeman & C., 1987
17. Spiegel M.R., *Statistica*, McGraw-Hill Libri Italia, 1994
18. Spiegel M. R., *Probabilità e Statistica*, McGraw-Hill Libri Italia, 1994
19. Upton G., Cook I., *Introducing Statistics*, Oxford University Press, 1998
20. Wonnacott T.H., Wonnacott R.J., *Introduzione alla Statistica*, Franco Angeli, 1995