Indici statistici di variabilità

Rendimento azioni Amazon e Microsoft a confronto

Anno	Microsoft	Amazon
1	1,6%	2,5%
2	5,4%	4,5%

Il rendimento medio e mediano sono uguali, ma Amazon ha rendimenti che vanno dal 2,5% al 4%, mentre Microsoft ha variazioni più ampie.

Indici di variabilità

L'utilizzo di indici di tendenza centrale non è sufficiente a discriminare situazioni molto differenti.

La **variabilità** rappresenta l'attitudine della variabile ad assumere diverse modalità.

In questo caso l'indicatore dovrà essere capace di graduare la variabile in termini di dispersione delle modalità rispetto ad un unico valore di sintesi (ad esempio una misura di posizione).

La concentrazione è in qualche modo una misura simmetrica della variabilità; quando un fenomeno è molto concentrato su un valore, si dice ragionevolmente che c'è poca variabilità.

Indici di variabilità

Un indice di variabilità:
- deve essere un numero positivo
- deve valere zero se calcolato su una distribuzione costante
- deve essere invariante se aggiungo una costante ad X

Gli indici di dispersione sono:
- l'**Intervallo di variazione (range)**
- lo **scarto interquartile**
- la **Varianza**
- la **Deviazione standard**
- il **Coefficiente di variazione**

Range

Il **range** (intervallo o campo di variazione) è la misura più semplice di variabilità ed è dato dalla differenza tra il valore più grande meno il valore più piccolo della distribuzione. Fornisce un'idea dello spazio all'interno del quale si muove il fenomeno, ma non dice nulla sulla variabilità all'interno dell'intervallo.

Range = Valore più grande – valore più piccolo
Esempio: calcolo il range dei seguenti stipendi

3310	3355	3450	3650	3730	3925

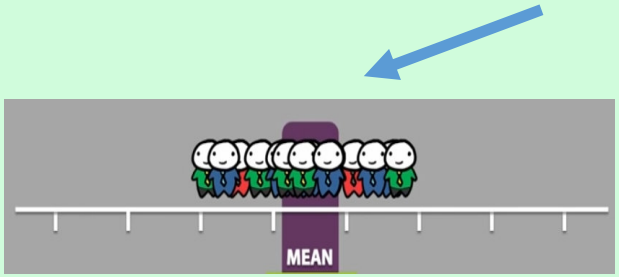
Range = 3925-3310 = 615

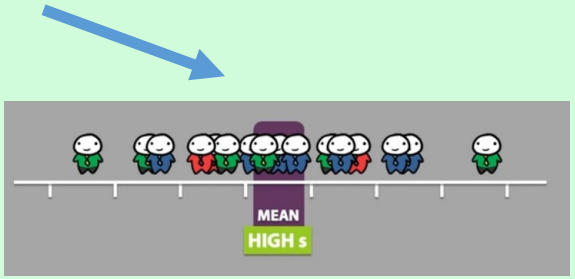
Limiti del range

Si base solo su due osservazioni e quindi è fortemente influenzato dai valori estremi.

È una misura che non indica come, ma solo quanto sono dispersi i dati.

Non ci dice se i dati sono concentrati attorno al valore centrale o attorno ad un altro valore o se sono distribuiti in modo omogeneo.





Esercizio

Calcolare il range e lo scarto interquartile della seguente distribuzione: 3, 3, 4, 5, 5, 6, 6, 6, 7, 8, 24.

Range = 24-3 = 19
Q1 = 0,25*11(N) = 2,75 quindi la posizione 3, corrisponde al numero 4.
Q3 = 0,75*11(N) = 8,25 quindi la posizione 9, che corrisponde al numero 7.
IQR = 7-4 = 3

Varianza

La **varianza** è una misura della variabilità che utilizza tutti i dati.

Si basa sulla differenza tra il valore di ciascuna osservazione xi e la media. Tale differenza viene definita scarto dalla media.

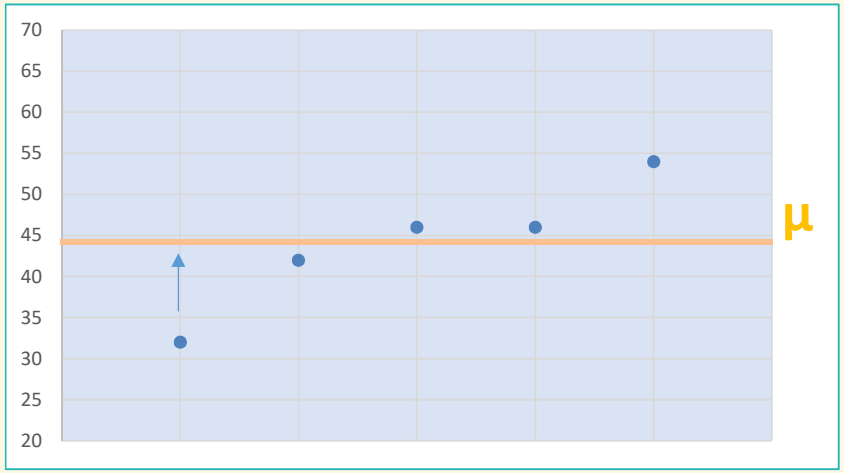
Nel calcolo gli scarti sono elevati al quadrato, quindi la varianza ha sempre valore positivo.

$$\sigma^2 = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2$$

Calcolo della varianza

La **varianza** descrive e quantifica la dispersione dei dati intorno al valore centrale della popolazione.

Si calcola la distanza esistente tra ogni singola osservazione e la media, che per definizione è nulla, per cui bisogna elevare al quadrato le distanze rilevate.



Esercizio

N. di studenti	xi - μ	xi - μ	$(xi - \mu)^2$
46	46-44	2	4
54	54-44	10	100
42	42-44	-2	4
46	46-44	2	4
32	32-44	-12	144
			$\sum$ 256

$\sigma^2 = 256/5 = 51,2$

Proprietà della varianza

Il valore zero equivale alla non dispersione e può essere ottenuto solo se tutte le osservazioni sono identiche.

Es. la seria 3, 3, 3, 3 avrà media 3 e varianza 0.

Il caso di massima variabilità si ha quando una unità possiede tutto il fenomeno e le altre n-1 unità hanno la modalità pari a zero.

Se moltiplico X per qualsiasi costante b, Y=bX allora la varianza di Y sarà moltiplicata per b^2 ovvero **$varY=b^2\,\sigma^2$**

Varianza per le distribuzioni di frequenza

$$\sigma^2 = \frac{(x_1 - M)^2 n_1 + (x_2 - M)^2 n_2 + \ldots + (x_k - M)^2 n_k}{n_1 + n_2 + \ldots + n_k} = \frac{1}{n}\sum_{i=1}^{k}(x_i - M)^2 n_i = \sum_{i=1}^{k}(x_i - M)^2 f_i$$

xi (voti)	ni (studenti)	xi - μ	$(xi - \mu)^2$	$(xi - \mu)^2 * ni$
24	18	-2,5	6,25	112,5
25	12	-1,5	2,25	27
26	16	-0,5	0,25	4
27	17	0,5	0,25	4,25
28	10	1,5	2,25	22,5
29	22	2,5	6,25	137,5
	∑ 95	0		∑ 307,75

μ=26,5

σ^2 = 307,75/95= 3,239

Deviazione standard

La **deviazione standard** (detta anche **errore standard, scarto quadratico medio, scarto tipo**) è definita come la radice quadrata della varianza.
Oltre a fornire informazioni su come il fenomeno sia disperso intorno al valore medio, dà un risultato nella stessa unità di misura della media.

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

Indici di variabilità

La **varianza** e la **deviazione standard** appartengono al gruppo degli scostamenti medi, indicatori che misurano la tendenza delle varie modalità del carattere a disperdersi attorno a un valore medio (dispersione), solitamente la media aritmetica.
In assenza di dispersione (distribuzione costante), gli indici assumono valore nullo.

Varianza

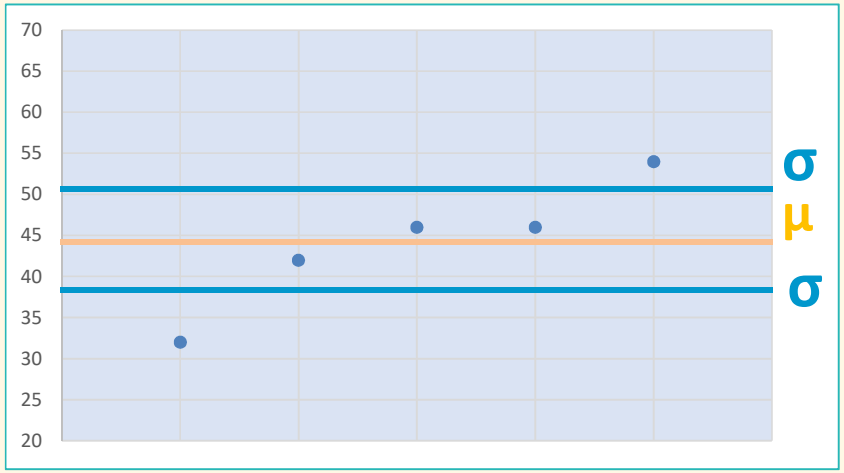
espressa al quadrato dell'unità di misura del carattere

Deviazione standard

espressa su scala lineare, di migliore interpretazione

Calcolo

$\sigma^2 = 256/5 = 51{,}2 \quad \sigma = 51{,}2^{(1/2)} = 7{,}1$



Esercizio

Calcolare la varianza e la deviazione standard per la seguente distribuzione: 4, 9, 9, 5, 8, 5, 6, 1.

$\mu = 5{,}875$

xi	xi - μ	$(xi - \mu)^2$
4	-1,875	3,52
9	3,125	9,77
9	3,125	9,77
5	-0,875	0,77
8	2,125	4,52
5	-0,875	0,77
6	0,125	0,02
1	-4,875	23,77
		$\sum$ 52,87

$\sigma^2 = 52{,}87/8 = 6{,}61$

$\sigma = \sqrt{6{,}61} = 2{,}57$

Esempio

Abbiamo 2 farmaci per il trattamento della pressione arteriosa e vediamo che effetto hanno dopo la somministrazione:

Farmaco	Prima		Dopo	
	Media	DV	Media	DV
1	90	15	70	10
2	90	15	70	15

Esempio

I 2 farmaci hanno prodotto gli stessi risultati medi? V

I pazienti del farmaco 1 hanno avuto una reazione più omogenea? V

Ogni paziente del farmaco 1 dopo il trattamento ha registrato livello di pressione inferiore rispetto ai pazienti del trattamento 2? F

I pazienti del farmaco 2 hanno avuto reazioni molto diverse al trattamento? V

Esempio

La deviazione standard è una misura comunemente utilizzata del rischio associato all'investimento in azioni o in fondi azionari.
Essa fornisce una misura di come i rendimenti mensili fluttuano attorno la rendimento medio di lungo periodo.

Azioni	**Microsoft**	**Amazon**
Rendimento	5% negli ultimi 5 anni	15% negli ultimi 5 anni
Deviazione standard	10	20

Minore è la deviazione standard, maggiore è la probabilità di ottenere un rendimento vicino a quello medio.

Esempio

Andamento del petrolio fine 2018 - inizi 2020

σ

In rosso i 3 momenti in cui il mercato ha mostrato un cambiamento di trend.
In questi casi la deviazione standard indica che vi è un'elevata volatilità e può dare indicazioni utili per anticipare un cambio di trend.

Coefficiente di variazione

Permette di confrontare fenomeni riferiti a unità di misura differenti o un ordine di misura diversa.
È una misura della variabilità relativa.

Ad esempio non possiamo confrontare la varianza di due diverse valute oppure il peso dei bambini con quello degli adulti.
In questi casi, per confrontare la variabilità di due distribuzioni per il carattere X con media M positiva, può essere utilizzato il **coefficiente di variazione**:

CV ha un campo di variazione positivo.

$$CV = \frac{\sigma}{|\mu|} * 100$$

Valore assoluto della media con $\mu \neq 0$

Esempio

	Altezza (m)	Peso (kg)
Media	165,8	60,64
Varianza	123,49	23,13

DEV	11,11	4,81
CV	6,70%	7,93%

CV del peso > CV dell’altezza

La variabilità per i caratteri qualitativi

Con riferimento ai caratteri qualitativi si parla di **mutabilità** o **eterogeneità**.
Con mutabilità si intende l'attitudine di un carattere qualitativo ad assumere differenti modalità.
Esempio: mettere a confronto le seguenti distribuzioni D1 e D2

Colore della macchina	Frequenze relative D1	Frequenze relative D2
Bianco	0,25	0
Nero	0,25	0
Rosso	0,25	0
Verde	0,25	1
	1	1

La variabilità per i caratteri qualitativi

Massima eterogeneità (omogeneità nulla) = le modalità del fenomeno qualitativo presentano uguale frequenza.

D1 = equidistribuzione

Massima omogeneità (eterogeneità nulla) = tutte le frequenze (100%) sono concentrate su un'unica modalità (la moda).

D2= tutti hanno lo stesso colore di macchina (il fenomeno presenta una sola modalità con frequenza non nulla)

Indice di eterogeneità di Gini

L'indice di Gini viene usato anche per lo studio della concentrazione industriale o di mercato.

Somma dei quadrati delle frequenze relative

$$G = 1 - \sum_{i=1}^{k} f_i^2$$

$$0 \leq G \leq \frac{k-1}{k}$$

Massima omogeneità: G = 0 se il collettivo è omogeneo: si osserverà solo una delle k modalità del carattere, che avrà frequenza assoluta pari a N. Le frequenze relative delle k–1 restanti modalità saranno nulle, tranne quella della modalità osservata, che varrà uno.

Minima omogeneità: G = (k-1)/k Nel caso di massima eterogeneità, i dati sono distribuiti equamente su tutte le k modalità, che hanno pari frequenza relativa.

Esercizio

Calcolare l'indice di Gini per la seguente distribuzione:

Stato civile	Frequenze assolute	Frequenze relative	fi^ 2
Celibe	36	0,177	0,031
Coniugato	74	0,365	0,133
Divorziato	60	0,296	0,087
Vedovo	33	0,163	0,026
	203	1	∑ 0,278

G = 1-0,278 = 0,722

I dati sono distribuiti in modo altamente eterogeneo sulle 4 scelte.

Indice di Gini normalizzato

Quando si hanno 2 distribuzioni dei dati per avere informazioni sul grado di >< elevatezza dell'eterogeneità bisogna normalizzare l'indice di Gini. Si ottiene moltiplicando l'indice ottenuto per k (che indica il numero delle modalità) e dividendolo per k-1.
La formula per calcolare l'indice di Gini normalizzato è la seguente:

$$G_N = G^* \frac{k}{k-1}$$

Quest'ultimo indice è chiaramente compreso tra 0 e 1.

Gn = 0,722*4/3 = 0,963

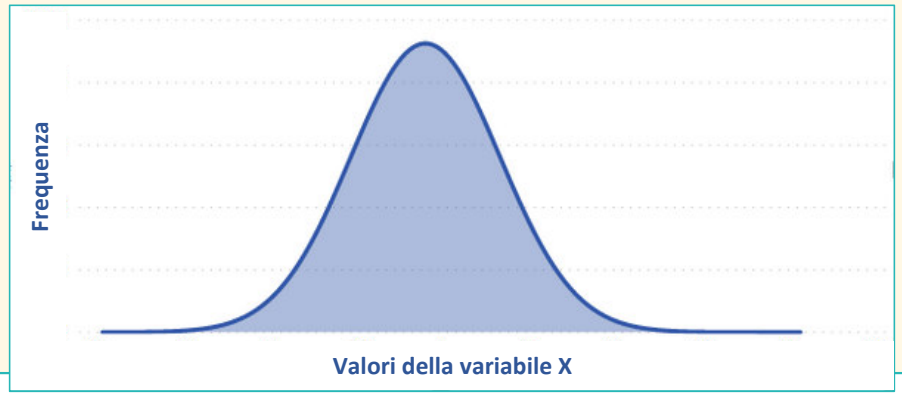
Riassunto

Carattere	Indice di dispersione
Qualitativo	Indici di mutabilità (eterogeneità di Gini)
Quantitativo	Range, IQR, Varianza, Deviazione standard, Coefficiente di variazione

Curva di Gauss o normale

La **curva di Gauss (curva normale)**, dalla classica forma a campana, viene utilizzata per descrivere la distribuzione di una variabile statistica continua.

L'evidenza ha dimostrato come si adatti bene a rappresentare il comportamento di una serie di fenomeni collettivi in svariati settori, es. medicina, biologia, economia etc. in cui le modalità intorno al valore medio sono più frequenti, mentre quelle alle estremità più rare.



Esempi

Nelle discipline scientifiche la curva di Gauss approssima la distribuzione degli errori commessi nel misurare ripetutamente una stessa grandezza (curva degli errori accidentali).

Molte variabili biologiche (peso, statura, pressione arteriosa, glicemia etc,) seguono la distribuzione normale.

Caratteristiche

Quando la popolazione è molto numerosa e le modalità sono tante, i rettangoli degli istogrammi di frequenza sono così fitti che danno luogo a una linea che sembra una vera e propria curva.

La curva continua approssima l'istogramma sottostante, ossia la distribuzione di frequenza.

Variabile aleatoria

Una **variabile aleatoria** (detta anche variabile casuale) è l'insieme dei possibili risultati numerici di un esperimento, cioè di una prova i cui esiti non sono prevedibili con certezza perché cambiano a seconda del verificarsi di eventi aleatori.

Sono grandezze che nel corso di un esperimento possono assumere diversi valori non prevedibili a priori in modo deterministico.

Esempio; nel lancio dei dadi non si può conoscere a priori il valore della faccia che si presenterà.

Variabile aleatoria

La relazione tra variabile aleatoria e probabilità si può ricavare assegnando a ogni esito possibile la probabilità che si verifichi. L'insieme di tutti i valori delle probabilità si chiama **distribuzione di probabilità**.

Si definisce distribuzione di probabilità (o funzione di probabilità) di una variabile aleatoria X l'insieme dei valori xi e delle relative probabilità pi.

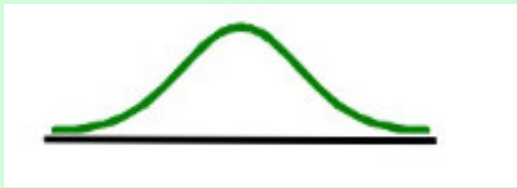
Si possono distinguere:
- **variabili aleatorie discrete** se l'insieme delle possibili realizzazioni è finito o numerabile (es. il numero dei giorni di pioggia in un anno).
- **variabili aleatorie continue**, se l'insieme delle possibili realizzazioni è un intervallo (es. la velocità di una macchina)

Funzione di probabilità e di densità

Per una variabile casuale discreta, la **funzione di probabilità** fornisce la probabilità che la variabile casuale assuma un particolare valore.

Con le variabili casuali continue, la controparte della funzione di probabilità è la **funzione di densità**. La differenza è che la funzione di densità non fornisce direttamente le probabilità.

Per descrivere la distribuzione di una variabile aleatoria continua, non si può più assegnare una probabilità positiva ad ogni valore possibile.



Funzione di probabilità

Esempio: negli ultimi 300 giorni di esercizio una concessionaria di macchine riporta i seguenti risultati di vendita

Numero di macchine vendute	Giorni	F(x)
0	54	0,18
1	117	0,39
2	72	0,24
3	42	0,14
4	12	0,04
5	3	0,01
	∑ 300	1

Funzione di probabilità

Dai dati storici sappiamo che x è una variabile casuale discreta che può assumere i valori 0, 1, 2, 3, 4, 5.

f(0) fornisce la probabilità di vendere 0 automobili, f(1) la probabilità di vendere 1 automobile e così via.

f(0) sta ad indicare che la probabilità di vendere 0 automobili nell'arco di un giorno è 0,18.

Nella costruzione di una funzione di probabilità discreta devono essere soddisfatte 2 condizioni:

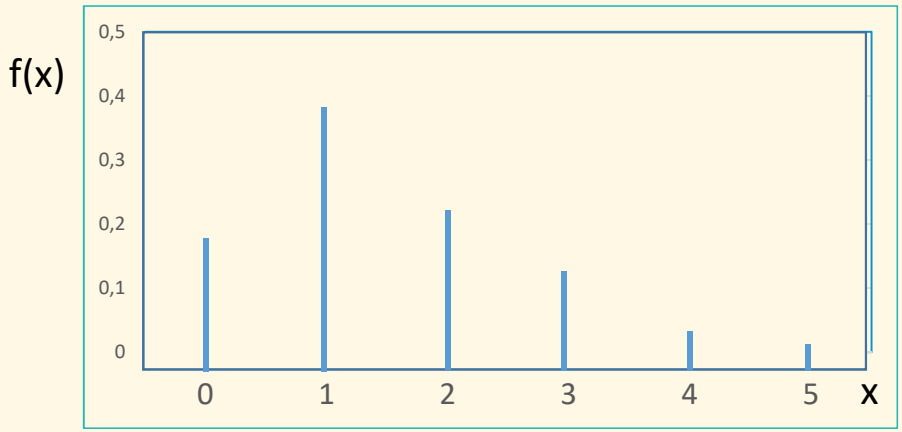
$$f(x) \geq 0$$

$$\sum f(x) = 1$$

Nella tabella precedente f(x) è ≥ 0 per tutti i valori delle x e la somma della probabilità è pari a 1

Rappresentazione grafica

Distribuzione di probabilità del numero di automobili vendute nell'arco di un giorno



Esercizio

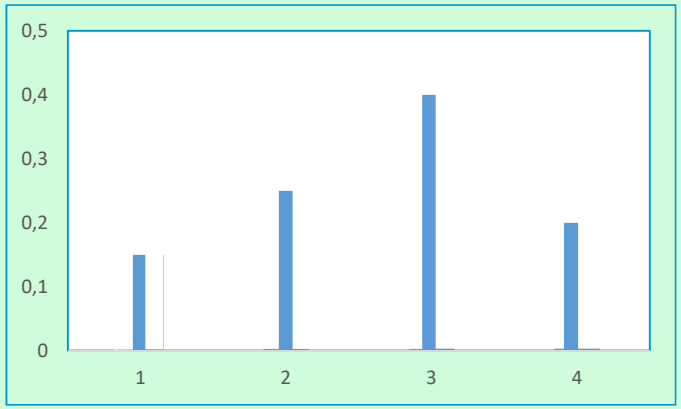
Contando il numero di sale operatorie in uso presso il Gemelli, nel corso di un periodo di 20 giorni sono stati raccolti i seguenti dati:
in 3 dei 20 gg è stata utilizzata solo 1 sala operatoria, in 5 dei venti giorni ne sono state utilizzate 2, in 8 dei 20 giorni ne sono state utilizzate 3 e in 4 dei 20 giorni sono state utilizzate tutte e 4 le sale.

1- costruire la distribuzione di probabilità del numero di sale operatorie utilizzate in un giorno
2- disegnare il grafico della distribuzione
3- mostrare che la distribuzione di probabilità ottenuta soddisfi le condizioni necessarie affinché una distribuzione di probabilità discreta sia valida.

Soluzione

x	f(x)
1	3/20=0,15
2	5/20=0,25
3	8/20=0,40
4	4/20=0,20
	1

Soddisfa entrambe le condizioni



Funzione di probabilità uniforme discreta

$$f(x) = 1/n$$

n = numero di valori che la variabile casuale può assumere

L'esperimento consiste nel lancio di un dado, la variabile casuale x è il numero che appare.
Sono possibili n = 6 valori per la variabile casuale x = 1, 2, 3, 4, 5, 6.

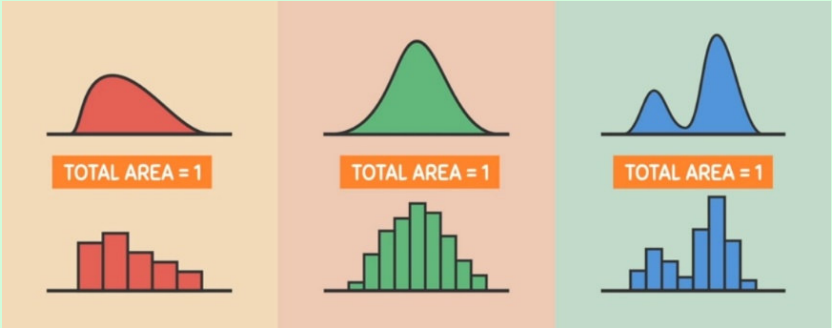
x	f(x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Funzione di densità di probabilità

Quando la variabile aleatoria X è continua la distribuzione delle probabilità P(X) si chiama **funzione di densità di probabilità**.

La sua rappresentazione grafica è una curva continua identificata da un'equazione y= P(x) e l'area totale tra l'asse orizzontale e la curva stessa è uguale a 1 se parliamo di frequenze relative o al 100% se parliamo di frequenza percentuale.

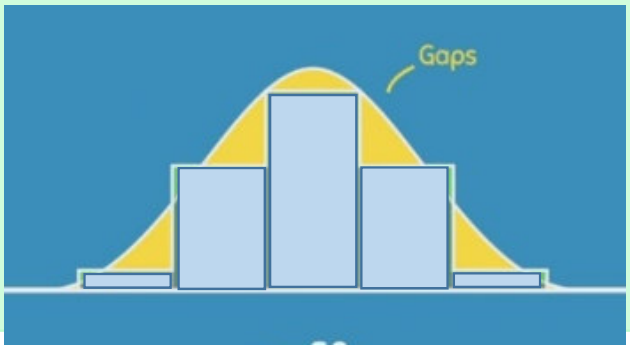
La curva di densità ci aiuta a visualizzare la forma della distribuzione e si può ricavare da ogni tipo di istogramma.



Vantaggi rispetto all'istogramma

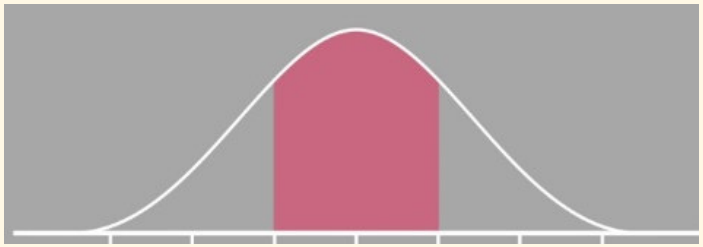
1- In un istogramma più intervalli si hanno a disposizione, migliore è la rappresentazione della distribuzione. Con la curva di densità non si è limitati dal numero di intervalli che si hanno.

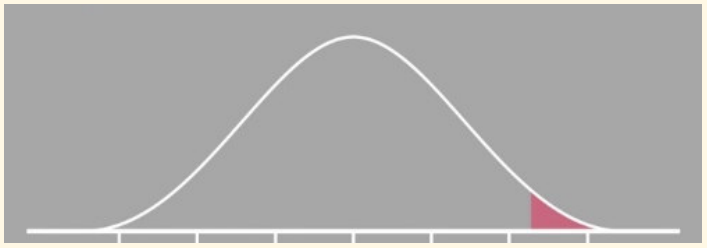
2- E' più facile lavorare con una curva soprattutto quando si ha a che fare con una popolazione di grandi dimensioni, mentre se si ha un numero esiguo di osservazioni la rappresentazione tramite una curva non è accurata, ci sono dei gap maggiori.



Funzione di densità di probabilità

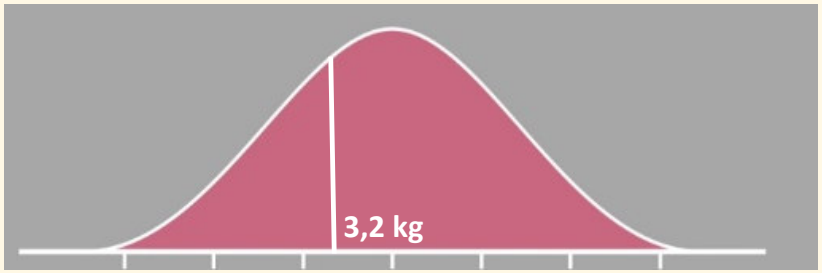
Il termine densità richiama il concetto di area sottostante la curva.
Se tutta l'area rappresenta il 100% delle probabilità, una piccola parte di essa quantifica la probabilità che la variabile casuale assuma un valore compreso nell'intervallo.





Esempio di utilizzo

Qual è la probabilità che un neonato pesi 3,2 kg alla nascita?



La formula della probabilità si calcola come casi favorevoli sul numero dei casi possibili, in questo caso infinito. Il risultato quindi è zero.

Non ha senso calcolare la probabilità di una singola modalità, ossia la probabilità che una variabile continua assuma un determinato valore è sempre 0.

Funzione di densità di probabilità

La funzione di densità non fornisce direttamente la probabilità, tuttavia l'area al di sotto della curva corrispondente a un dato intervallo fornisce la probabilità che la variabile casuale continua assuma un valore in quell'intervallo.

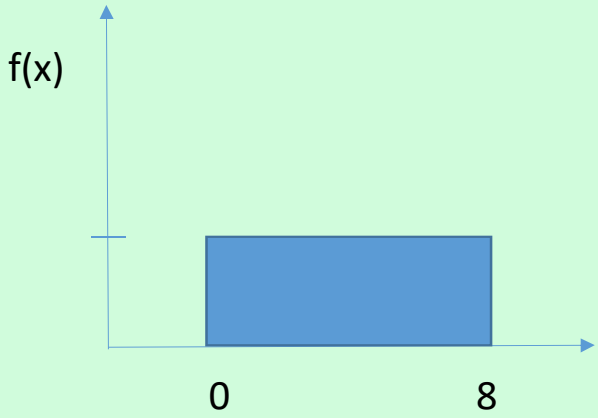
Quindi per variabili casuali continue si calcola la probabilità che la variabile assuma qualunque valore in un determinato intervallo.

Presi 2 valori a e b, l'area identificata dalle loro proiezioni sull'asse orizzontale rappresenta quante probabilità abbiamo che la variabile aleatoria assuma uno qualunque dei valori compresi tra i 2 estremi.

Funzione di densità uniforme

Ogni volta che la probabilità è proporzionale alla lunghezza dell'intervallo, la variabile casuale si distribuisce uniformemente.

$$f(x) = \frac{1}{b-a} \qquad a \leq x \leq b$$



Area compresa nel rettangolo = 1
Si calcola come base per altezza

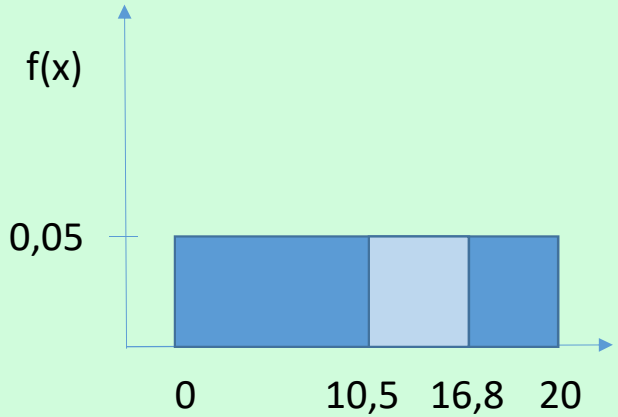
Base = 8
Altezza = 1:8 = 0,125

Esempio

Qual è la proporzione dei valori compresi tra 10,5 e 16,8?

f(x) = 1/20 = 0,05

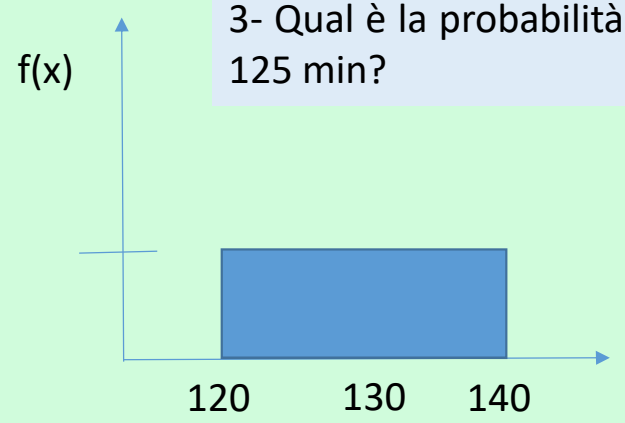
Base = 6,3 Area = 0,05*6,3 = 0,315 o 31,15%



Esempio

Tempo di volo compreso in un intervallo di 120-140min con ogni intervallo di 1 min ugualmente probabile

1- Qual è la probabilità che il tempo di volo sia compreso tra 120 e 130 min? Quanto vale la $P(120 \leq X \leq 130)$?
2- Qual è la probabilità che il tempo di volo sia compreso tra 128 e 136 min?
3- Qual è la probabilità che il tempo di volo sia 125 min?



Caratteristiche della curva normale

1- Simmetrica rispetto alla retta $X = \mu$, ossia speculare rispetto al suo valore centrale. La forma della curva a sinistra è speculare rispetto a quella a destra.

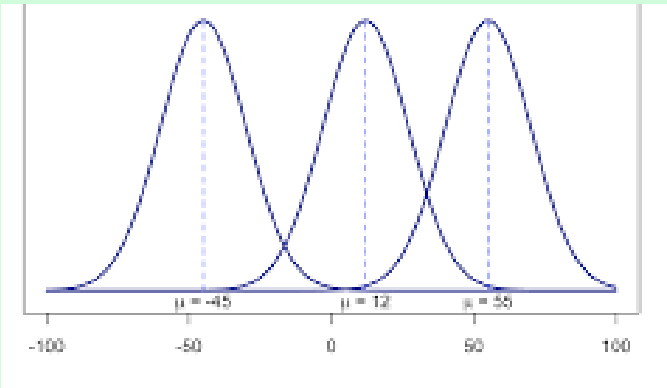
2- La moda si trova nel punto di **massima frequenza** e quindi il punto di massimo della curva.
Moda, mediana e media coincidono.

3- Bastano 2 valori per disegnare la curva normale, la **μ** e la **σ**. Tali valori determinano la posizione e la forma della distribuzione normale e la differenziano da un'altra curva normale.

Caratteristiche

Il valore della media che indica il centro della distribuzione (della campana), caratterizza la posizione della curva rispetto all'asse delle ordinate.

Al variare della media la curva si sposta lungo l'asse x, mentre la sua forma non si modifica.
La media può assumere anche valore negativo.



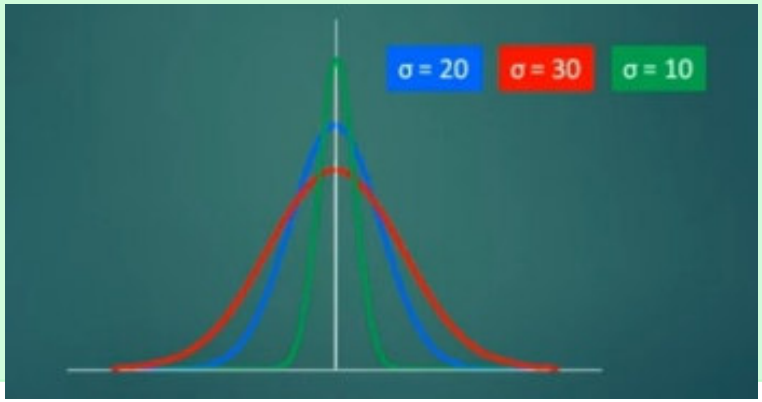
Uguale σ ma μ diversa

Fonte; *Hartmann, K., Krois, J., Waske, B. (2018)*

Caratteristiche

La deviazione standard caratterizza la forma della curva, in quanto è una misura della dispersione dei valori attorno al valore medio: al variare di σ la curva cambia forma.
Al crescere di σ la curva si appiattisce e si allarga, mentre al diminuire di σ la curva si restringe e si alza.

In corrispondenza dei punti di flesso si manifesta un cambiamento di curvatura.
I punti **μ−σ** e **μ+σ** sono punti di flesso, ossia si trovano una σ sopra la media e una σ sotto la media.



Caratteristiche

4- Trattandosi di variabili continue l'intera area sotto la curva a campana è sempre uguale a 1. L'area alla sinistra della media è pari a 0,5 così come quella a destra.

5- Asintotica rispetto all'asse delle x (al tendere di *x* verso $-\infty$ o $+\infty$). Allontanandosi dal valore centrale si avvicina sempre più all'asse orizzontale senza toccarlo.
Non potendo sapere con certezza quale sarà l'ultimo valore con almeno una osservazione e tenendo presente che le frequenze verso gli estremi diminuiscono sempre di più, si disegna una curva che non si interrompe in un punto specifico.

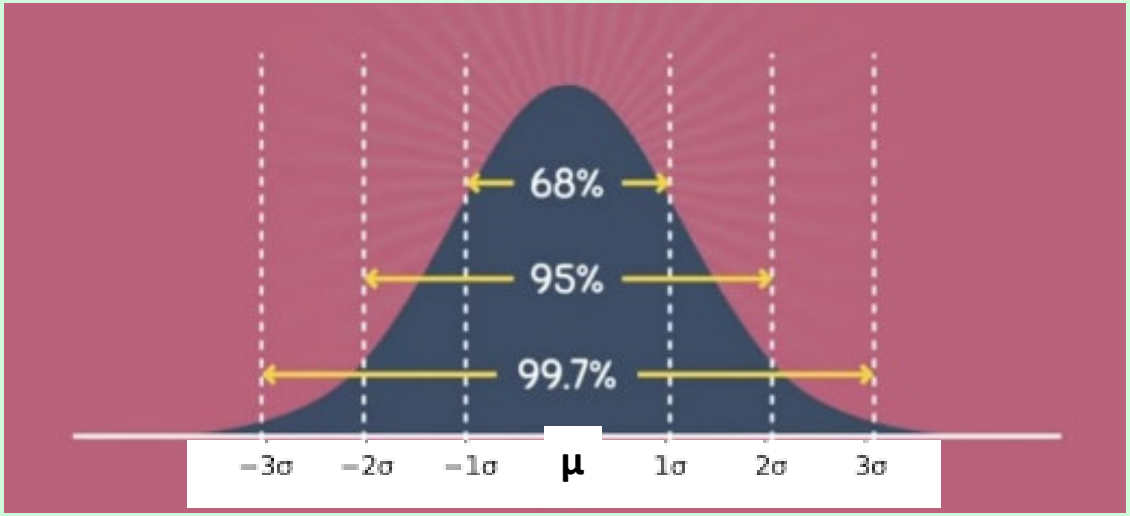
Funzione di densità normale

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{(x-\mu)}{\sigma}\right]^2}$$

μ = media
σ = dev std
π = 3.14159
e = 2.71828 base dei logaritmi naturali

Regola empirica

Regola empirica o del 68-95-99,7 è una regola statistica che afferma che per una distribuzione normale, quasi tutti i dati osservati cadranno entro tre deviazioni standard



Regola empirica

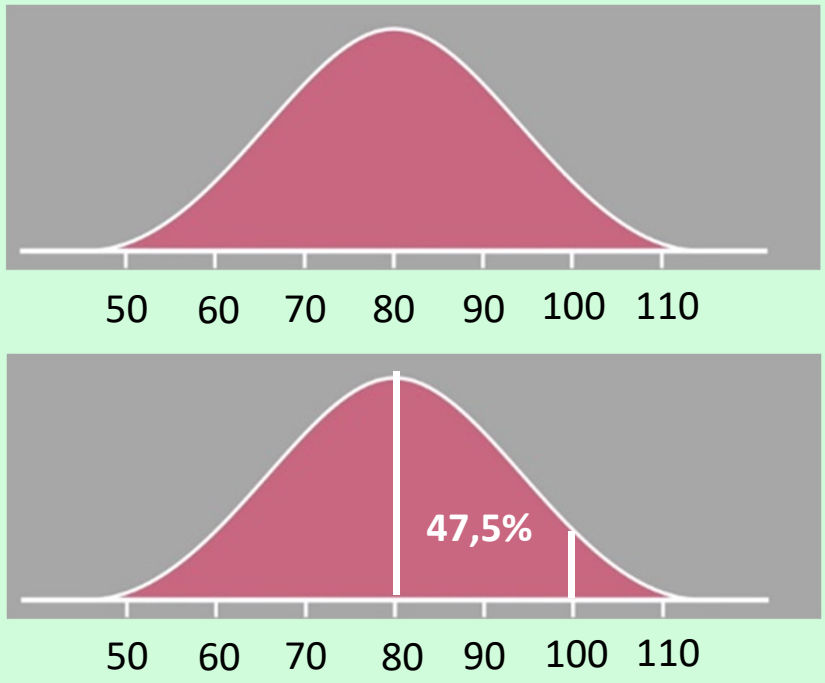
Nell'intervallo compreso fra $\mu - \sigma$ e $\mu + \sigma$ si trovano circa il 68% dei valori osservati della distribuzione (rientrano in una deviazione standard).
Per cui il 32% dei valori si trovano equamente presenti all'esterno dell'intervallo.

Nell'intervallo compreso fra $\mu - 2\sigma$ e $\mu + 2\sigma$ si trovano circa il 95% dei valori osservati. Per cui il 5% dei valori si trovano equamente presenti all'esterno dell'intervallo.

Il 99,7% dei dati osservati si trova entro 3 deviazioni standard dalla media.

Esempio

La distribuzione normale nel grafico ha una deviazione standard di 10. Qual è l'area contenuta tra 80 e 100?



Esempio

Per la distribuzione normale nel grafico qual è l'area contenuta tra -2 e 1? Media = 0 e Deviazione standard 1