UNIVERSITÁ DEGLI STUDI DI TERAMO

SCIENZE DELLA COMUNICAZIONE

ANALISI STATISTICA DEI DATI AZIENDALI

LA STATISTICA DESCRITTIVA

FABRIZIO ANTOLINI
fantolini@unite.it

I CARATTERI STATISTICI E LE TECNICHE DI RILEVAZIONE

La **Statistica** può essere definita come un insieme di tecniche che hanno come scopo la conoscenza quantitativa dei fenomeni collettivi, ossia i fenomeni il cui studio richiede l'osservazione di un insieme di manifestazioni individuali.

Operazioni tipiche delle analisi statistiche sono:

- il conteggio
- la classificazione
- la misurazione
- la sintesi tramite modelli esplicativi dei fenomeni reali

Qualche breve definizione:

- **unità statistica:** è l'unità elementare su cui vengono osservati i caratteri oggetto di studio. Le unità statistiche possono essere unità semplici come autovetture, persone singole, lanci di monete, ecc.; oppure unità composte, cioè aggregati di unità semplici, come le famiglie, le aziende, ecc.
- **carattere:** il carattere rappresenta ciò che si intende osservare dell'unità statistica oggetto di osservazione. Attraverso l'osservazione dei caratteri si inizia a quantificare o a rendere misurabile fenomeno collettivo.
- **collettivo o popolazione:** Per popolazione (o collettivo statistico o aggregato) si intende l'insieme degli elementi che sono oggetto di studio, ovvero l'insieme delle unità statistiche sulle quali viene effettuata la rilevazione delle modalità con le quali il fenomeno studiato si presenta.
- **modalità del carattere:** La modalità rappresenta le diverse manifestazioni del carattere.

I CARATTERI STATISTICI E LE TECNICHE DI RILEVAZIONE

I caratteri possono essere di due tipi:

- Quantitativi
- Qualitativi

Un carattere è detto **QUANTITATIVO** quando le modalità sono espresse da numeri (ad esempio l'età, il peso, l'altezza, etc..).

Un carattere è detto **QUALITATIVO (o MUTABILE)** quando le modalità sono espresse da parole, aggettivi, avverbi e qualsiasi altra modalità non numerica (ad esempio colore occhi, capelli, professione, titolo di studio, etc..).

I caratteri quantitativi possono essere:

- Su **scala di intervalli** (es. temperatura) o su **scala di rapporti** (es. altezza). La differenza è che nelle variabili ad intervallo il valore zero è arbitrario mentre in quelle a rapporto lo zero è dotato di un significato.
- **Continui** (possono assumere qualsiasi valore all'interno di un intervallo predefinito) o **discreti** (possono assumere solo numeri interi).
- **Trasferibili** (l'ammontare del carattere posseduto da una unità statistica può essere trasferito da una ad un'altra) e **non trasferibili** (es. pesò ed età).

I caratteri qualitativi possono essere:

- **Sconnessi** (le modalità non sono ordinabili mediante criteri oggettivi) o **ordinabili** (rettilineo o ciclico).
- Su **scala nominale (**Le variabili nominali possono invece essere costituite anche da più di due modalità che non hanno un criterio di ordinamento logico, ad esempio il tipo di sport praticato. L'ordine è arbitrario e cambiandolo non si perde o guadagna nulla in termini informativi).

I CARATTERI STATISTICI E LE TECNICHE DI RILEVAZIONE

L'indagine statistica

Lo studio di un fenomeno collettivo con il metodo statistico si può articolare nelle seguenti fasi:

Definizione degli obiettivi

- definizione delle unità e delle variabili da rilevare
- scelta del periodo di riferimento

Individuazione della popolazione e della lista delle unità statistiche

Definizione del piano di campionamento

Raccolta dei dati

- scelta della tecnica di rilevazione
- formulazione del questionario
- rilevazione sul campo

Registrazione dei dati

- registrazione su supporto magnetico
- controllo e correzione

Elaborazione e analisi dei dati

I CARATTERI STATISTICI E LE TECNICHE DI RILEVAZIONE

La rivelazione dei dati

È quel complesso di operazioni con le quali si perviene alla conoscenza dei dati, ossia delle modalità di uno o più caratteri di un collettivo statistico.

La raccolta delle informazioni può essere **completa** oppure **parziale**.

È completa quando si esaminano tutte le unità statistiche che compongono la popolazione oggetto di studio.

Pregi:

- Accuratezza delle stime anche a livelli territoriali molto spinti
- Ricchezza delle informazioni raccolte
- Esaustività

Difetti:

- Costo elevato
- Tempi di elaborazione dei dati molto lunghi
- Qualità dei dati non elevata

È parziale quando ci si limita a studiare un sottoinsieme, detto "campione" dell'insieme di riferimento.

Pregi:

- Continuità della rilevazione
- Economicità
- Indagini più mirate e approfondite

Difetti:

- Riferimento territoriale non spinto
- Variabilità campionaria

I CARATTERI STATISTICI E LE TECNICHE DI RILEVAZIONE

La **statistica descrittiva** fornisce gli strumenti per sintetizzare ed esplicitare in forma corretta il modo in cui il fenomeno si è manifestato nel collettivo osservato.

Mediante **l'inferenza statistica** è possibile misurare e controllare l'attendibilità delle informazioni provenienti da un campione -> Estrazione del campione.

Estrazione del campione:

- *Campionamento casuale*: insieme di tutte quelle tecniche di formazione del campione in cui la selezione delle unità è affidata a regole probabilistiche.
- *Campionamento casuale semplice*: i campioni della stessa dimensione estraibili da una popolazione hanno uguale probabilità di essere estratti.
- *Campionamento casuale stratificato*: la popolazione viene suddivisa in un certo numero di strati. Da ogni strato in maniera indipendente viene poi estratto un campione casuale semplice. Se gli strati sono stati ben scelti la stima subisce un miglioramento e si ottiene la possibilità anche di ottenere la stima per le singole sottopopolazioni (o strati).

Vedremo più approfonditamente le strategie e tecniche campionarie nelle prossime unità didattiche

Progettazione del questionario

Il questionario è lo strumento realizzato per raccogliere tutte le informazioni che interessano la ricerca a cui lo stesso si riferisce.

La sua funzione è di consentire la rilevazione delle informazioni in modo univoco, allo scopo di permettere la classificazione e la misurazione dei dati raccolti. Il questionario è uno dei due pilastri su cui si basa la realizzazione di una ricerca quantitativa.

Le fasi operative per la stesura di un questionario:

- *Definizione di bisogni conoscitivi*: consiste nel momento in cui si dovranno definire gli obiettivi della ricerca e le informazioni di cui si vuole venire in possesso.
- *Identificazione del contenuto delle domande da porre*: Fase molto delicata, in quanto richiede di decidere se inserire solo le domande le cui risposte siano utili alla realizzazione degli obiettivi formulare solo domande alle quali si è certi che il target intervistato sia in grado di rispondere; impostare le domande in modo da non creare imbarazzo al target intervistato
- *Definizione del tipo di domande*: domande aperte; domande chiuse; domande parzialmente chiuse; domande con scale verbali o semantiche; domande con scale numeriche; domande di differenziale semantico

I CARATTERI STATISTICI E LE TECNICHE DI RILEVAZIONE

- *Modalità di formulazione delle domande*: occorre ricordare che il testo di ogni singolo quesito deve avere per il target intervistato lo stesso significato attribuito da chi redige la domanda. Si deve pertanto porre attenzione a che i termini impiegati siano comprensibili, chiaramente definiti e di significato univoco.
- *Definizione della sequenza delle domande*: le domande iniziali hanno il compito di creare interesse; le domande difficili e quelle che riguardano i dati personali vanno poste alla fine del l'intervista
- *Verifica degli aspetti formali del questionario*: consiste nel controllo della funzionalità e dell'estetica (grafica, ordine, leggibilità, etc..).
- *Esecuzione di una fase pilota:* lo scopo è quello di verificare che le domande siano poste correttamente, che il questionario sia agevole per l'intervistato e la durata non vada oltre il tempo medio di attenzione.

FREQUENZE ASSOLUTE, RELATIVE E CUMULATE E LE TABELLE SEMPLICI, DOPPIE MULTIPLE

Distribuzione di un carattere

Le distribuzioni statistiche descrivono il modo in cui uno o più caratteri si manifestano (distribuiscono) in un dato collettivo.

Nel caso di un singolo carattere si è in presenza di distribuzioni semplici; per due caratteri di distribuzioni doppie e per tre o più caratteri di una distribuzione multipla.

L'elenco delle modalità osservate, unità per unità viene definita distribuzione unitaria.

Esempio: Distribuzione unitaria per sesso

Codice intervistato	Sesso
1	M
2	F
3	F

FREQUENZE ASSOLUTE, RELATIVE E CUMULATE E LE TABELLE SEMPLICI, DOPPIE MULTIPLE

Distribuzioni di frequenze

- Frequenza assoluta $\mathbf{n_j}$: numero di volte che la modalità di un carattere viene osservata nel collettivo **(N).**
- Distribuzione di frequenze assolute: associa alle modalità che può assumere un carattere X le corrispondenti frequenze assolute.
- Frequenza relativa f_j: è la frazione di collettivo che presenta la modalità *j-esima* ossia $f_j = \frac{n_i}{N}$
- Frequenza percentuale p_j: è uguale alla frequenza relativa moltiplicata per 100.
- Frequenza cumulata N_j: è data dalla somma della frequenza assoluta della modalità con quella della modalità precedente (può essere calcolata sia nella forma relativa che percentuale). $Nj = \sum_{j=1}^{i} fj$

-

Esempio: Distribuzione assoluta, relativa, percentuale, e cumulata

Sesso	n_j	f_j	p_j	N_j
M	8	0,4	40%	8
F	12	0,6	60%	20
TOTALE	20	1	100%	x

FREQUENZE ASSOLUTE, RELATIVE E CUMULATE E LE TABELLE SEMPLICI, DOPPIE MULTIPLE

Nel caso di due soli caratteri X e Y, la variabile statistica doppia si configura come l'insieme delle coppie di valori (x_i,y_i) corrispondenti alla stessa unità *i-esima* del collettivo ($i=1,2,3...,N$). Questa è rappresentata numericamente da una tabella doppia, detta di correlazione.

Esempio di variabile statistica doppia

Voti in Statistica (X)	Voti in Informatica (Y)					
	24	**25**	**26**	**27**	**28**	**TOTALE**
24	5	11	16	6	1	**39**
25	8	2	6	8	7	**31**
26	5	10	10	4	9	**38**
27	6	5	1	3	5	**20**
28	7	9	8	4	3	**31**
TOTALE	**31**	**37**	**41**	**25**	**25**	**159**

FREQUENZE ASSOLUTE, RELATIVE E CUMULATE E LE TABELLE SEMPLICI, DOPPIE MULTIPLE

Nel caso di combinazioni di modalità di due o più caratteri qualitativi, la mutabile (variabile) statistica doppia è rappresentata da una tabella detta di contingenza.

Esempio di mutabile statistica doppia

Professione/posizione	Settore attività economica			TOTALE
	Agricoltura	Industria	Altro	
Imprenditore	2	3	2	**7**
In proprio	1	1	2	**4**
Dirigenti	2	1	1	**4**
Lavoratori dipendenti	2	4	3	**9**
Coadiuvanti	1	2	3	**6**
TOTALE	**8**	**11**	**11**	**30**

LE RAPPRESENTAZIONI GRAFICHE

Attraverso la rappresentazione grafica si garantisce una migliore visualizzazione del fenomeno collettivo analizzato. La trasformazione della distribuzione semplice da forma tabellare a immagine grafica ha senso se tale operazione riesce a rendere più evidenti e di facile lettura le caratteristiche della distribuzione del carattere sul collettivo preso in esame.

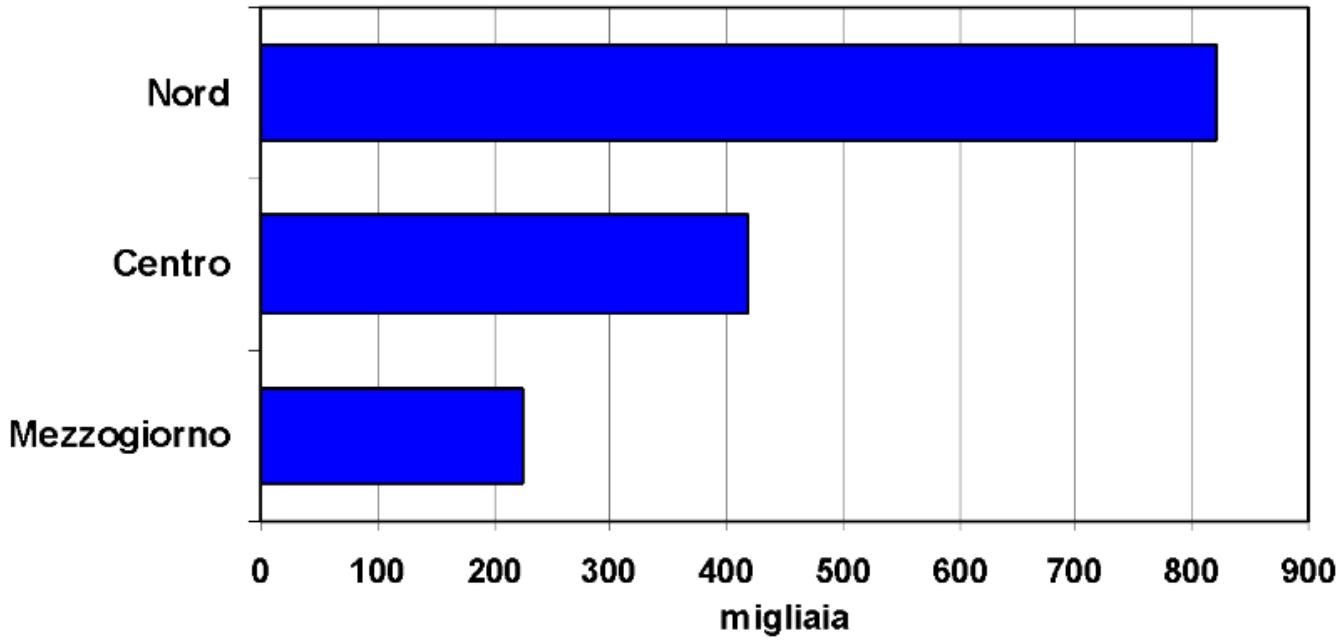
In particolare, si utilizzano:

- **Grafici a nastri** per caratteri qualitativi non ordinati
- **Grafici a barre** per caratteri qualitativi ordinati, caratteri quantitativi discreti
- **Grafici ad aree** per caratteri quantitativi continui nel tempo
- **Istogrammi** per caratteri quantitativi continui suddivisi in classi
- **Grafici a torta** per caratteri qualitativi non ordinati o ordinati ciclici
- **Grafici radar** per caratteri ciclici
- **Cartogrammi** per serie territoriali
- **Diagrammi cartesiani** per serie storiche

Grafici a nastri o a barre

Ogni frequenza viene rappresentata da un nastro così da ottenere una successione di rettangoli aventi la stessa altezza e basi proporzionali alle frequenze.

Esempio

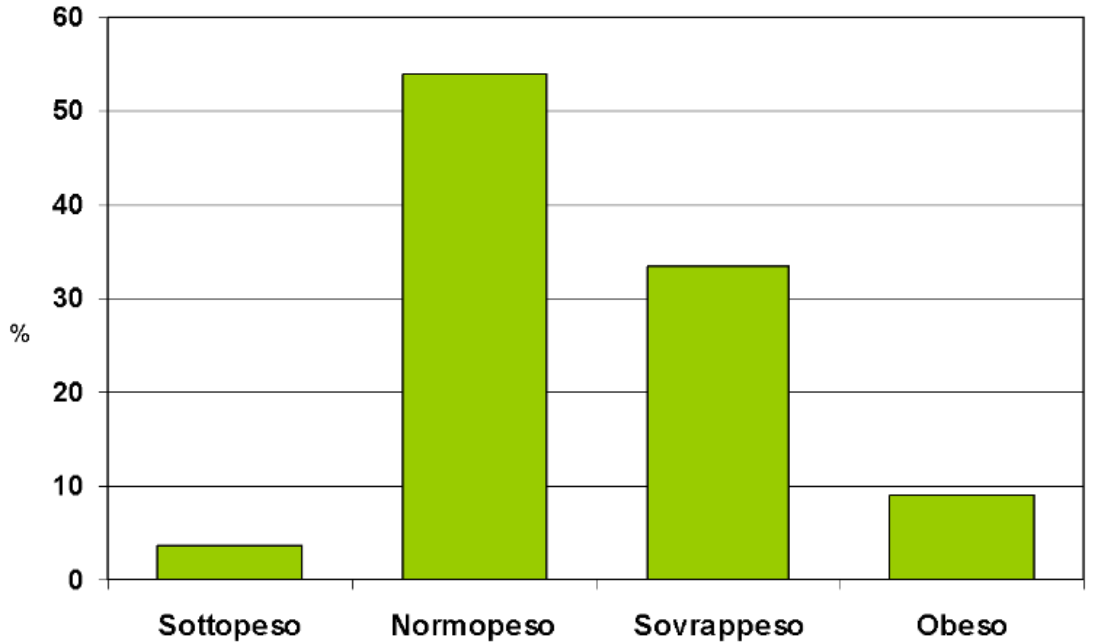


LE RAPPRESENTAZIONI GRAFICHE

Se il carattere è qualitativo ordinato o quantitativo discreto, è preferibile utilizzare il grafico a barre giacché le barre poste sull'asse orizzontale consentono di cogliere meglio l'ordinamento delle modalità. Ogni frequenza viene rappresentata da una barra così da ottenere una successione di rettangoli aventi la stessa base e altezze proporzionali alle frequenze.

N.B: Se per uno stesso carattere si sono osservate due o più distribuzioni semplici, relativamente per esempio a diversi collettivi, possiamo metterli a confronto riportando queste in un grafico a barre (o nastri) multipli.

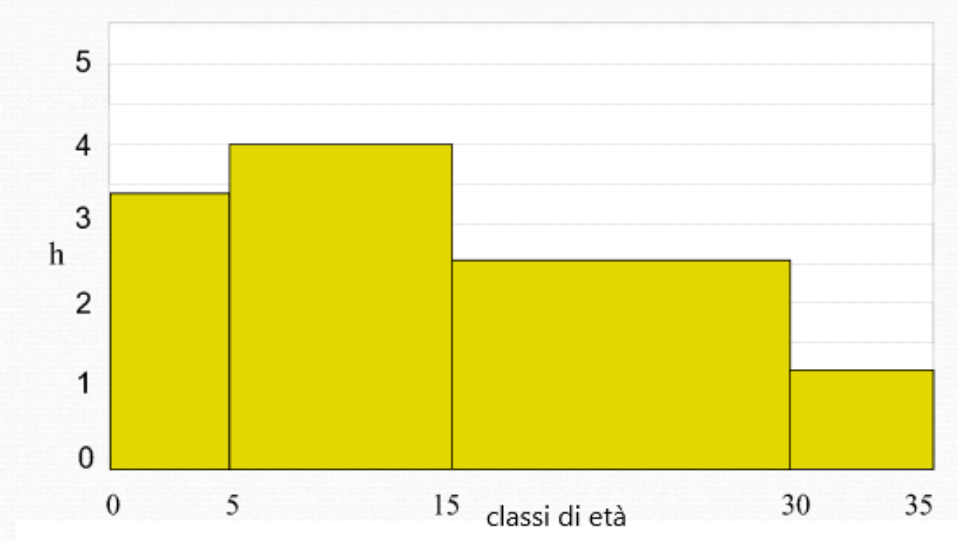
Esempio



Istogramma

L'istogramma è un grafico costituito da barre non distanziate, dove ogni barra possiede un'area proporzionale alla frequenza della classe. Occorre dunque calcolare la densità di frequenza **(h)** che si ottiene come rapporto tra la frequenza e l'ampiezza di classe.

Esempio



classi di età	amp. classe a_j	freq. % p_j	densità h_j
0-5	5	17,0	3,4
5-15	10	40,0	4,0
15-30	15	37,0	2,5
30-35	5	6,0	1,2

Grafico a torta

Si utilizza principalmente per rappresentare la composizione di un aggregato (sia in termini di frequenza assoluta, che relativa, che percentuale).

Esempio

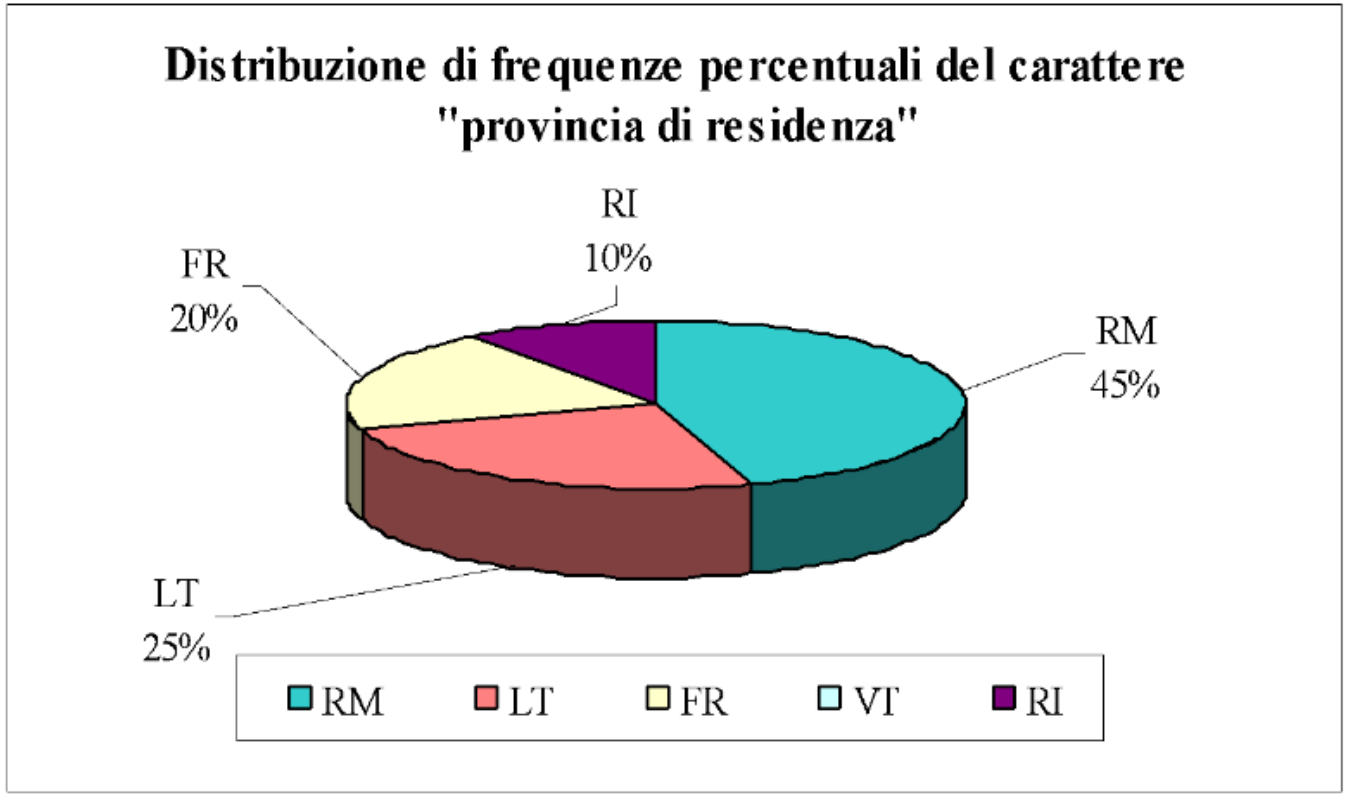
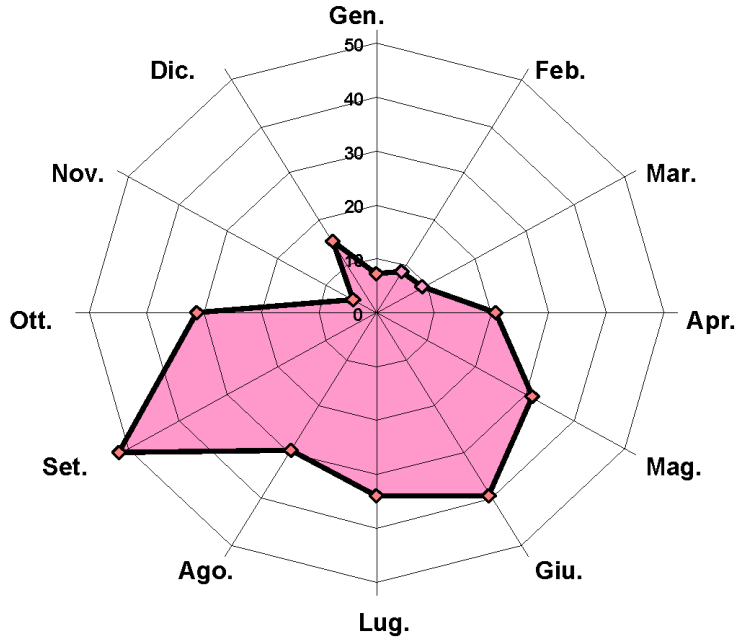


Grafico Radar

Si utilizza per caratteri ciclici. Si suddivide l'angolo di 360° con tanti raggi quante sono le modalità del carattere:

- agli angoli compresi tra coppie di raggi si attribuisce la stessa ampiezza.
- su ogni raggio si calcola un segmento di lunghezza proporzionale o uguale alla corrispondente frequenza.
- può essere efficacie unire con una spezzata gli estremi dei segmenti e colorare l'area interna al poligono che si viene a formare.

Esempio: numero di matrimoni in Italia per mese di celebrazione (anno 2019)



Cartogramma

Si utilizza per rappresentare le serie territoriali. Ha come base una mappa sulla quale sono visibili i contorni delle aree geografiche rispetto alle quali vengono analizzate le frequenze. Ogni area è colorata in base alla distribuzione di frequenza.

Esempio: tasso di disoccupazione per regione (Italia, anno 2003)

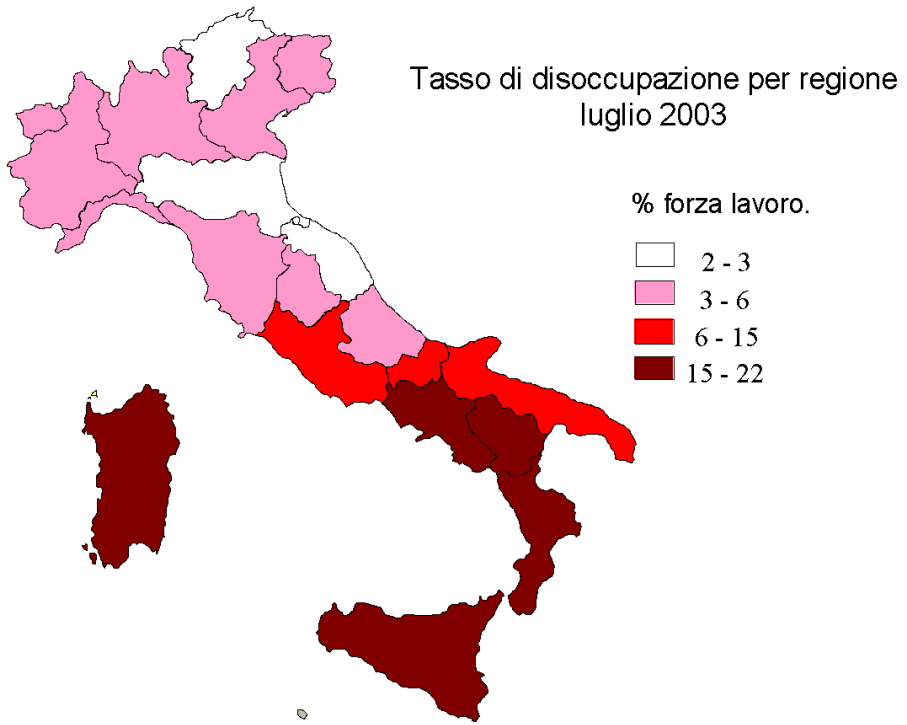
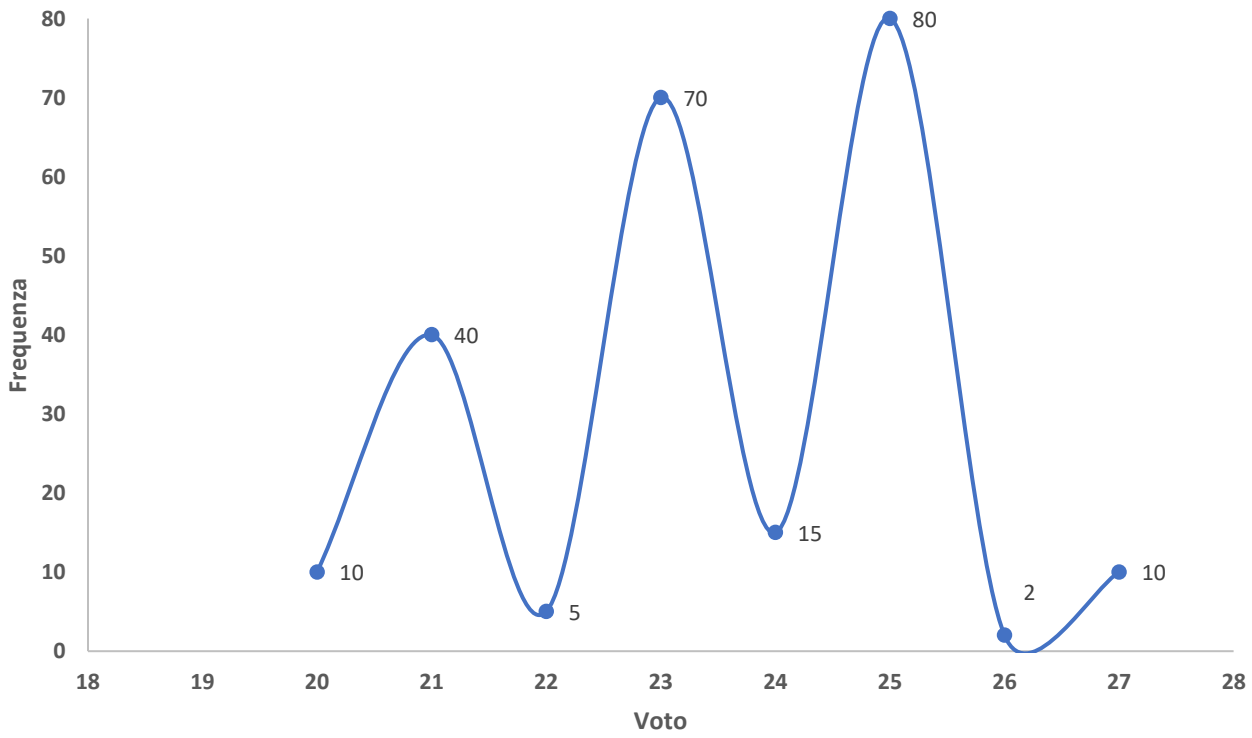


Diagramma cartesiano

Sull'asse delle ascisse ***X*** sono rappresentate le modalità x_i, e sull'asse delle ordinate ***Y*** sono riportate le corrispondenti frequenze n_j.

Esempio



LE MEDIE ALGEBRICHE (MEDIA ARITMETICA E GEOMETRICA) E DI POSIZIONE (MODA, MEDIANA; QUARTILI, PERCENTILI)

Analisi delle misure sintetiche

In statistica, si chiamano medie quei valori che offrono una sintesi di un insieme di dati:

Queste vengono dette anche indici di posizione, o indicatori di posizione, o indici di tendenza centrale o misure di tendenza centrale.

Si distinguono due tipi di medie:

1. medie di posizione, che considerano solo certi valori e, in generale, non si ottengono manipolando matematicamente i dati;
2. medie algebriche, che considerano tutti i valori e si ricavano applicando formule matematiche.

Qualunque sia la media, essa è sempre espressa nella medesima unità di misura delle modalità da cui si è ricavata.

MEDIE DI POSIZIONE

MODA

Si definisce moda (o modalità) degli $\mathbf{n}$ elementi $\mathbf{x_1, x_2, x_3, \ldots . x_n}$, l'elemento (o gli elementi) a cui corrisponde la massima frequenza assoluta.

- Nel caso sia una la modalità che presenta la frequenza maggiore di tutte le altre, allora la distribuzione si dirà unimodale, se due bimodale, se tre tri modale, etc.. (può anche non esistere);
- La moda campionaria non è influenzata da valori estremi
- La moda è l'unica tra le medie che ammetta caratteri sia quantitativi, sia qualitativi
- La moda dipende solo dalle frequenze
- La moda acquista validità solo se vi è una netta prevalenza di una modalità/intensità
- La moda si calcola su tutti i tipi di caratteri

MEDIE DI POSIZIONE

MEDIANA

Si definisce mediana degli n dati ordinati $\mathbf{x_1, x_2, x_3, \ldots . x_n}$, il valore centrale della serie, cioè il valore che occupa il posto $\frac{\mathbf{n+1}}{\mathbf{2}}$ nella serie se n è dispari o la media dei valori che occupano i posti $\frac{\mathbf{n}}{\mathbf{2}}$ ed $\frac{\mathbf{n}}{\mathbf{2}} + \mathbf{1}$ se n è pari.
Il carattere deve essere quantitativo o qualitativo ordinabile;
È quella modalità la cui frequenza percentuale cumulata vale 50%.

Se il carattere è suddiviso in classi, si può ottenere un valore ben approssimato tramite la formula:

$$\boldsymbol{M_e \approx I_M + \left(\frac{0{,}5 - F_{m-1}}{F_m - F_{m-1}}\right)\Delta_m}$$

in cui si assume implicitamente l'ipotesi che nella classe mediana le unità siano distribuite uniformemente

- La proprietà più importante è:

$$\sum_{i=1}^{n} |x_i - c| \text{ è minimo per } c = M_e$$

PERCENTILI

Definiamo percentili quei valori che dividono la distribuzione in cento parti di uguale numerosità.

- I percentili di uso più frequente sono il *25-esimo* e il *75-esimo* percentile, detti anche primo (Q_1) e terzo quartile (Q_3) che insieme alla mediana dividono la distribuzione in quattro parti uguali (la mediana corrisponde al secondo quartile, Q_2).

MEDIE ALGEBRICHE

Per ciascuna media algebrica, si distinguono i casi *semplice* e *ponderata*, la prima da applicarsi a distribuzioni unitarie e l'altra a distribuzioni di frequenze (la frequenza è chiamata anche peso).

MEDIA ARITMETICA

Si definisce media aritmetica degli degli n elementi $x_1, x_2, x_3, \ldots . x_n$, è quel valore che, sostituito ai dati, lascia invariata la loro somma.

Semplice	Ponderata
$\overline{X} = \frac{\sum_{i=1}^{n} xi}{n}$	$\overline{X} = \frac{\sum_{i=1}^{n} xi * fi}{\sum_{i=1}^{n} fi}$

MEDIE ALGEBRICHE

- Il valore della media aritmetica perde di significato quando i dati presentano valori eccezionali (*valori anomali – outliers*)
- La media aritmetica è tanto più indicativa quanto i dati sono tra loro omogenei (per quantificare la variabilità di un insieme di dati si vedano gli indici di variabilità)
- La media aritmetica è quel valore il quale, considerando il quadrato degli scarti tra ciascun dato ed essa rende minima la loro somma (il valore centrale che rende minima la somma degli scarti è la mediana)
- La somma delle differenze tra i valori e la loro media aritmetica è pari a zero
- Nel caso di una distribuzione di frequenze per un carattere **X** suddiviso in classi, possiamo approssimare la media utilizzando il valore centrale della classe $c_{j:}$

$$x_a \cong \frac{1}{n}\sum_{j=1}^{K} c_{j*n_j}$$

MEDIA GEOMETRICA

Si definisce media geometrica degli degli n elementi $x_1, x_2, x_3, \ldots . x_n$, è quel valore che, sostituito ai dati, lascia invariato il loro prodotto.

Semplice	Ponderata
$\overline{X} = \sqrt[n]{\prod_{i=1}^{n} xi}$	$\overline{X} = \sqrt[n]{\prod_{i=1}^{n} xi^{fi}}$

MEDIE ALGEBRICHE

- La media geometrica, in generale, serve per distribuzioni che seguono una progressione geometrica, ovvero si usa quando i dati oggetto dell'indagine sono valori che per loro natura vanno tra loro moltiplicati;
- La media geometrica prevede che nessuno dei valori sia nullo (perché annullerebbe il prodotto) e neanche negativo;
- La media geometrica è uguale all'esponenziale della media dei logaritmi delle *xi*:

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

LA TRIMMED MEAN (MEDIA TRONCATA)

La trimmed mean è la media aritmetica calcolata su una fissata percentuale di valori centrali di un insieme di dati.

- Elimina l'influenza dei valori anomali (ad esempio nella trimmed mean al 50% si escludono il 25% dei valori più piccoli e il 25% dei valori più grandi).

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

L'ANALISI DELLA VARIABILITÀ

In statistica, si definisce variabilità l'attitudine di un fenomeno ad assumere misure diverse tra loro.

Così, mentre le medie hanno un significato descrittivo dell'intensità dei fenomeni, le misure della variabilità hanno un significato descrittivo dell'uguaglianza o disuguaglianza dei fenomeni, sia che li si veda al proprio interno, sia che vengano posti a confronto essendo diversi.

Nella metodologia statistica si distinguono due aspetti della variabilità:

1) la dispersione, che caratterizza il maggiore o minore addensamento delle osservazioni intorno ad una media;
2) la disuguaglianza, che evidenzia la diversità delle varie osservazioni tra loro.

I conseguenti indici che si ottengono, si distinguono in:

- indici assoluti di variabilità, che sono espressi nella stessa unità di misura del fenomeno in esame e sono: la varianza, lo scarto quadratico medio, etc.
- indici relativi di variabilità, che prescindono dall'unità di misura del fenomeno esaminato e sono particolarmente adatti per effettuare confronti tra fenomeni diversi. Si ottengono rapportando un indice assoluto ad una media o al suo massimo.

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

CAMPO DI VARIAZIONE

Un indice assoluto della variabilità di una successione di dati, di immediata percezione e assai semplice da calcolarsi, è rappresentato dal campo di variazione (o range), che è dato dalla differenza tra il valore massimo e il valore minimo della successione. Di fatto costituisce l'ampiezza dell'intervallo dei dati.

$$\omega = X_{max} - X_{min}$$

- L'indice in questione è poco utilizzato in quanto prende in considerazione solo la dispersione esistente tra i valori estremi della distribuzione (risente dei valori anomali);
- Il campo di variazione è espresso nella stessa unità di misura dei dati, tanto è più piccolo più i dati sono concentrati, viceversa tanto più è grande tanto più i dati sono dispersi.

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

VARIANZA

La varianza di un insieme di dati o di una distribuzione di frequenza è una misura di dispersione che si ottiene come media dei quadrati degli scarti dalla media aritmetica, in simboli:

La varianza presenta, tuttavia, un notevole inconveniente nel senso che è espressa attraverso il quadrato dell'unità di misura delle osservazioni, per cui se le osservazioni ad esempio sono in metri, la varianza è espressa in metri al quadrato.

Semplice	Ponderata
$\sigma^2 = \frac{\sum_{i=1}^{n}(xi - \overline{x})^2}{n}$	$\sigma^2 = \frac{\sum_{i=1}^{n}(xi - \overline{x})^2 * f_i}{n}$

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

SCOSTAMENTO QUADRATICO MEDIO O DEVIAZIONE STANDARD

Per ovviare all'inconveniente dell'unità di misura si preferisce usare la radice quadrata della varianza e ottenere un importante indice di variabilità, tra tutti il più utilizzato, denominato scostamento quadratico medio o deviazione standard, in simboli:

Lo scostamento quadratico medio o deviazione standard è un indice altamente rappresentativo del maggiore o minore addensamento dei dati intorno al loro valore medio.

Semplice	Ponderata
$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(xi - \overline{x})^2}{n}}$	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(xi - \overline{x})^2 * f_i}{n}}$

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

DEVIANZA

Da ultimo infine consideriamo il numeratore della varianza in quanto si presenta come un'altra misura della dispersione denominata devianza**.** Per un carattere X la sua espressione analitica, a seconda che si abbia una successione di dati o una distribuzione di frequenza è:

Semplice	Ponderata
$D(X) = \sum_{i=1}^{n} (xi - \overline{x})^2$	$D(X) = \sum_{i=1}^{n} (xi - \overline{x})^2 * f_i$

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

COEFFICIENTE DI VARIAZIONE

In realtà confrontare le deviazioni standard non è di grande aiuto, perché esse dipendono fortemente dalle media dei dati su cui sono state calcolate. Per poter operare un confronto sulla variabilità di gruppi diversi è opportuno calcolare il coefficiente di variazione, un indice relativo di variabilità assai utilizzato e definito come rapporto tra scarto quadratico medio e media aritmetica. È in sostanza un numero puro che esprime $\boldsymbol{\sigma}$ in termini di $\bar{\boldsymbol{x}}$:

$$\boldsymbol{Cv} = \frac{\boldsymbol{\sigma}}{\bar{\boldsymbol{x}}}$$

- Il **Coefficiente di Variazione** non essendo espresso in alcuna unità di misura consente di effettuare confronti fra distribuzioni diverse per fenomeni omogenei;
- Il **Coefficiente di Variazione** è utilizzato per confrontare la variabilità relativa di un fenomeno in circostanze differente ed anche tutte le volte che si intende confrontare la variabilità di due fenomeni espressi in unità di misure diverse.
- **($\boldsymbol{Cv}$ = 0),** in questo caso la deviazione standard è pari a 0. Tutti i dati sono uguali tra loro e la media può essere considerata come un indice perfetto per rappresentarli.
- **($\boldsymbol{Cv} \geq$ 0.5),** in questo caso la deviazione standard è più della metà della media. La media, in questo caso, non può essere considerata un buon indice per rappresentare i dati.
- **($\boldsymbol{Cv} \leq$ 0.5),** in questo caso la deviazione standard è meno della metà della media. La media, in questo caso, può essere considerata un buon indice per rappresentare i dati.

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

COEFFICIENTE DI VARIAZIONE

In realtà confrontare le deviazioni standard non è di grande aiuto, perché esse dipendono fortemente dalle media dei dati su cui sono state calcolate. Per poter operare un confronto sulla variabilità di gruppi diversi è opportuno calcolare il coefficiente di variazione, un indice relativo di variabilità assai utilizzato e definito come rapporto tra scarto quadratico medio e media aritmetica. È in sostanza un numero puro che esprime $\boldsymbol{\sigma}$ in termini di $\bar{\boldsymbol{x}}$:

$$\boldsymbol{Cv} = \frac{\boldsymbol{\sigma}}{\bar{\boldsymbol{x}}}$$

- Il **Coefficiente di Variazione** non essendo espresso in alcuna unità di misura consente di effettuare confronti fra distribuzioni diverse per fenomeni omogenei;
- Il **Coefficiente di Variazione** è utilizzato per confrontare la variabilità relativa di un fenomeno in circostanze differente ed anche tutte le volte che si intende confrontare la variabilità di due fenomeni espressi in unità di misure diverse.
- **($\boldsymbol{Cv}$ = 0),** in questo caso la deviazione standard è pari a 0. Tutti i dati sono uguali tra loro e la media può essere considerata come un indice perfetto per rappresentarli.
- **($\boldsymbol{Cv} \geq$ 0.5),** in questo caso la deviazione standard è più della metà della media. La media, in questo caso, non può essere considerata un buon indice per rappresentare i dati.
- **($\boldsymbol{Cv} \leq$ 0.5),** in questo caso la deviazione standard è meno della metà della media. La media, in questo caso, può essere considerata un buon indice per rappresentare i dati.

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

ALTRI INDICI DI VARIABILITÀ

Lo scostamento semplice medio dalla media aritmetica:

$$S_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

Lo scostamento semplice medio dalla mediana:

$$S_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} |x_i - M_e|$$

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

LA STANDARDIZZAZIONE

La standardizzazione è una particolare trasformazione lineare che applicata ai dati originali riconduce qualsiasi variabile **X** con media $\overline{x}$ e deviazione standard σ a una nuova variabile con media nulla e varianza unitaria.

- Ogni osservazione x_i viene trasformata in un nuovo valore:

$$y_i = \frac{x_i - \overline{x}}{\sigma}$$

LE MISURE DELLA VARIABILITÀ: CAMPO DI VARIAZIONE VARIANZA, SCARTO QUADRATICO MEDIO, CV, DIFFERENZA INTERQUARTILE (BOX PLOT)

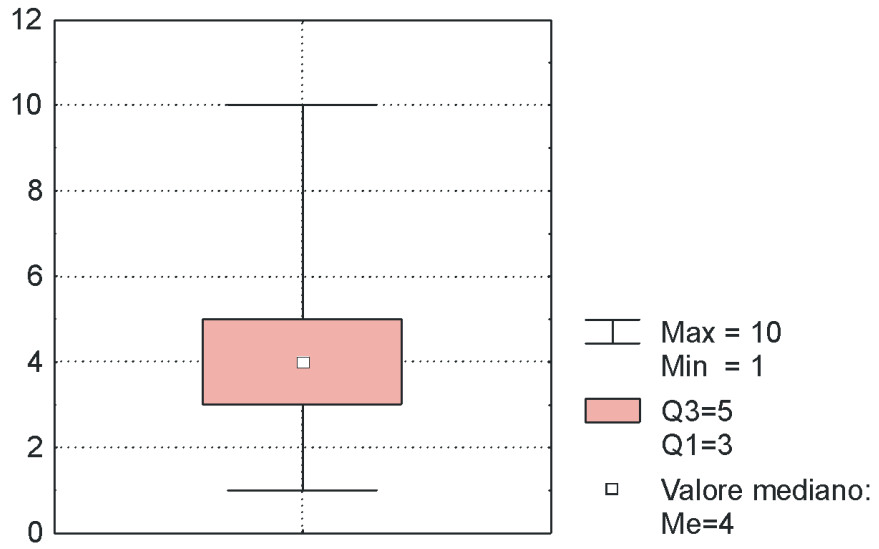
BOX-PLOT

Un modo per rappresentare graficamente la variabilità di una distribuzione è dato dal box-plot.

Il box-plot è un grafico caratterizzato da tre elementi:

- una linea o punto, che indicano la posizione della media della distribuzione
- un rettangolo (box) la cui altezza indica la variabilità dei valori "prossimi" alla media
- due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione.

Esempio



LA CONCENTRAZIONE

Osservando l'ammontare di un carattere quantitativo trasferibile su un collettivo statistico, si può essere interessati a sapere come questo ammontare sia ripartito fra le unità statistiche del collettivo in esame.

Un carattere quantitativo $\boldsymbol{X}$, con n valori osservati $\boldsymbol{x1}, \boldsymbol{x2}, \ldots\ldots \boldsymbol{xn}$, si dice equidistribuito se ognuna delle $\boldsymbol{n}$ unità possiede $\frac{\mathbf{1}}{\boldsymbol{n}}$ dell'ammontare complessivo del carattere.

$$\boldsymbol{A} = \sum_{i=1}^{n} x_i$$

Ossia per ogni *i* si ha:

$$\boldsymbol{x_i} = \frac{\boldsymbol{A}}{\boldsymbol{n}} = \overline{\boldsymbol{x}}$$

- Se non si verifica l'equidistribuzione, sussiste un certo grado di concentrazione del carattere che può essere misurato tramite opportuni indici.
- Tanto più un carattere è concentrato, tanto più è elevata la variabilità del carattere, mentre se non sussiste variabilità allora anche la concentrazione è nulla.

LA CONCENTRAZIONE

La situazione di massima concentrazione si ha quando l'intero ammontare del carattere, $\boldsymbol{A}$, è posseduto da una sola unità del collettivo e cioè:

$$\boldsymbol{x_1 = x_2 = \cdots = x_{n-1} \;\; e \;\; x_n = A}$$

Per un carattere quantitativo trasferibile **X,** si ha:

$\boldsymbol{A_i = x_1 + x_2 + \cdots + x_i}$; l'ammontare di carattere posseduto dalle *i* unità

$\boldsymbol{Q_i = \frac{A_i}{A_n}}$; la corrispondente frazione di ammontare

$\boldsymbol{F_i = \frac{i}{n}}$; la frequenza relativa cumulata delle *i* unità

LA CONCENTRAZIONE

Le distribuzioni di $\boldsymbol{Q_i}$ e $\boldsymbol{F_i}$ possono essere messe a confronto. Nel caso in cui non sussista equidistribuzione del carattere vale la relazione $\boldsymbol{Q_i \leq F_i}$. Possiamo sintetizzare tali differenze attraverso il seguente indice:

$$\boldsymbol{C = \sum_{i=1}^{n-1}(F_i - Q_i)}$$

La sommatoria arriva fino al termine (n-1)-esimo, perché *l'n-esima* differenza è sempre uguale a 0 essendo $\boldsymbol{Q_n = F_n = 1}$. Questo indice assume valore minimo quando tutte le differenze sono pari a zero, cioè nel caso di equidistribuzione, e il suo valore massimo nel caso di massima concentrazione (dove le $\boldsymbol{Q_i = 0}$ e $\boldsymbol{Q_n = 1}$).

$$\boldsymbol{C = \sum_{i=1}^{n-1} F_i}$$

Per trasformare l'indice C in un indice di concentrazione relativo, variabile tra 0 e 1, basterà dividerlo per il suo valore massimo. Con tale operazione di ottiene l'indice di concentrazione chiamato rapporto di concentrazione di GINI.

RAPPORTO DI CONCENTRAZIONE DI GINI

Date le distribuzioni delle F_i e delle Q_i relative alla distribuzione di un carattere quantitativo **X**, osservato su n unità, con valori ordinati $x_1, x_2, \ldots, xn$, si definisce rapporto di concentrazione di Gini l'indice:

$$R = \frac{1}{\sum_{i=1}^{n-1} F_i} * \sum_{i=1}^{n-1} (F_i - Q_i)$$

Dato che il numeratore può essere scritto anche come:

$$\sum_{i=1}^{n-1} F_i - \sum_{i=1}^{n-1} Q_i$$

Il rapporto di concentrazione di Gini può essere formulato come segue:

$$R = 1 - \frac{\sum_{i=1}^{n-1} Q_i}{\sum_{i=1}^{n-1} F_i}$$

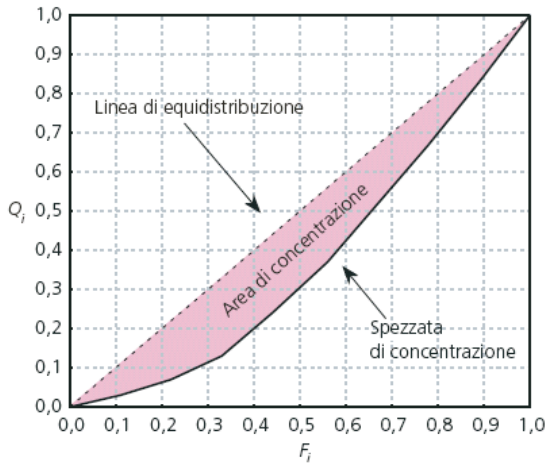
Mediante le coppie di valori Qi *e* Fi identificando vari punti, è possibile realizzare un interessante grafico su un piano cartesiano. I punti limitrofi possono essere congiunti da segmenti tali da formare una curva detta **spezzata di concentrazione** o **curva di Lorenz**, dal nome del primo autore che ne propose l'impiego.

CURVA DI LORENZ

Nel grafico, oltre alla spezzata di concentrazione, viene rappresentata la **linea di equidistribuzione** che è il segmento che congiunge il punto (0;0) al punto (1;1).

- Se l'ammontare del carattere fosse equidistribuito fra tutte le unità del collettivo, i punti corrispondenti giacerebbero sulla linea di equidistribuzione.
- La curva di Lorenz giace sotto la linea di equidistribuzione, poiché $\boldsymbol{Q_i \leq F_i}$.
- L'area della superficie compresa tra la curva di Lorenz e la linea di equidistribuzione viene detta **area di concentrazione**.
- La curva di Lorenz cambia la sua forma a seconda che il carattere osservato sul collettivo sia più o meno concentrato: più è vicina alla linea di equidistribuzione e più l'ammontare del carattere è equidistribuito fra le unità; e viceversa.

Esempio



ANALISI D'ASSOCIAZIONE TRA CARATTERI

Un caso di grande rilevanza nelle applicazioni riguarda l'analisi dell'associazione tra due caratteri quantitativi. Lo studio di queste relazioni viene complessivamente chiamato analisi dell'associazione.

- Concordanza: se modalità di ordine elevato di ***X*** si associano più frequentemente a modalità di ordine elevato di ***Y***, mentre modalità di ordine basso di ***X*** si associano più frequentemente a modalità di ordine basso di ***Y***.
- Discordanza: se modalità di ordine elevato di ***X*** si associano più frequentemente a modalità di ordine basso di ***Y***, mentre modalità di ordine basso di ***X*** si associano più frequentemente a modalità di ordine elevato di ***Y***.

Quando i caratteri della distribuzione doppia sono quantitativi, possiamo rappresentare la distribuzione doppia attraverso il grafico di dispersione.

Nel grafico di dispersione le coppie di modalità di due caratteri quantitativi, osservate per ogni unità del collettivo, vengono rappresentate come punti di un piano cartesiano i cui assi ortogonali corrispondono ai due caratteri.

Questo grafico è molto utile quando si ha a disposizione la distribuzione unitaria e quindi le singole coppie di valori osservati dei caratteri. In questa maniera ogni coppia di modalità individua la posizione di un'unità sul piano cartesiano.

COVARIANZA

La Covarianza misura l'intensità e il verso della relazione lineare tra due variabili quantitative (X e Y). Il termine covarianza rimanda all'idea di una misura di quanto due variabili variano insieme, covariano per l'appunto.

$$\boldsymbol{cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}}$$

Come nel caso della varianza, anche la covarianza ha una espressione alternativa che consente di pervenire a uno schema di calcolo più semplice, ovvero la Codevianza.

$$\boldsymbol{codev(X,Y) = \sum_{i=1}^{n} x_i\, y_i * n\bar{x}\bar{y}}$$

- La covarianza, pur essendo una misura della relazione lineare fra due variabili quantitative, ha un grave difetto, in quanto può assumere qualsiasi valore che dipenda sia dalla grandezza dei fenomeni considerati che dalle unità di misura delle variabili.
- Tramite la covarianza non si riesce a determinare la forza del legame, per far ciò è necessario calcolare un altro indice chiamato coefficiente di correlazione.

COEFFICIENTE DI CORRELAZIONE

Il coefficiente di correlazione si ottiene dividendo la covarianza di ***X*** e ***Y*** per i relativi scarti quadratici medi. In questo modo si eliminano sia le due unità di misura delle due variabili, sia le grandezze relative. Il numero così ottenuto è un numero puro, privo cioè di unità di misura, e normalizzato, ovvero con un campo di variazione ben determinato (-1, 1)

$$r = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

- I valori del coefficiente di correlazione variano tra -1, che indica una perfetta correlazione negativa, e +1, che indica una perfetta correlazione positiva.
- "Perfetta correlazione" significa che, se si disegnano i punti su un diagramma a dispersione, tutti i punti sono allineati, ovvero sono disposti su una retta (relazione lineare).
- La correlazione da sola non dimostra che esiste un effetto di causalità, cioè che la variazione del valore di una variabile ha causato il cambiamento di altre variabili. Una forte correlazione indica che le due variabili variano congiuntamente in un verso o nell'altro (la sola correlazione non implica *causazione*).
- Esso indica soltanto la tendenza ad associarsi delle variabili.

INDICE RHO DI SPEAMAN

L'indice di correlazione R per ranghi di Spearman è una misura statistica non parametrica di correlazione che si applica per caratteri qualitativi ordinati che rappresentano delle graduatorie.

- A livello pratico il coefficiente *ρ* è semplicemente un caso particolare del coefficiente di correlazione di Pearson dove i valori vengono convertiti in ranghi prima di calcolare il coefficiente
- La differenza tra i ranghi dell'i-esima unità è $d_i = r_i - s_i$ (essendo r_i e s_i rispettivamente il rango della prima variabile e della seconda variabile della *i-esima* osservazione)

$$p_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n * (n^2 - 1)}$$

- *P=1;* i ranghi sono in perfetta concordanza
- *P=-1;* i ranghi sono in perfetta discordanza
- *P=0;* le due graduatorie non mostrano associazione.

LA DIPENDENZA STATISTICA

X e Y si diranno indipendenti se le distribuzioni relative condizionate di un carattere rispetto all'altra sono uguali. Due variabili statistiche sono indipendenti se le modalità di una non influenzano le modalità dell'altra.

- La conoscenza della modalità di uno dei due caratteri non migliora la previsione della modalità dell'altro.
- Se ***X*** è indipendente da ***Y*** allora anche ***Y*** è indipendente da ***X.*** Dette ***X*** e ***Y*** le due variabili statistiche, la distribuzione delle frequenze delle loro modalità x_1, x_2, x_q e y_1, y_2, y_p può essere rappresentata attraverso una tabella a doppia entrata in cui si associa ad ogni coppia $(x_i;y_i)$ la sua frequenza assoluta detta frequenza congiunta.

SEGUE...

ANALISI D'ASSOCIAZIONE TRA CARATTERI

Y	X				TOTALE
	x_1	x_2	x_3	x_q	
y_1	$f_{(1,1)}$	$f_{(1,2)}$	$f_{(1,3)}$	.....	$f_{(1,o)}$
y_2	$f_{(2,1)}$	$f_{(2,2)}$	$f_{(2,3)}$	.....	$f_{(2,0)}$
y_3	$f_{(3,1)}$	$f_{(2,3)}$	$f_{(3,3)}$	.....	$f_{(3,0)}$
y_p	.....	.....	.....	.....	$f_{(p,0)}$
TOTALE	$f_{(0,1)}$	$f_{(0,2)}$	$f_{(0,3)}$	$f_{(0,q)}$	f

LA DIPENDENZA STATISTICA

- La prima riga è quella delle modalità del carattere ***X***, la prima colonna è quella delle modalità del carattere ***Y***
- La colonna dei totali e la riga dei totali sono le frequenze marginali della variabile ***X*** e della variabile **Y**, sono dette distribuzioni marginali e rappresentano le distribuzioni di ognuno dei due caratteri considerati singolarmente (distribuzioni univariate)
- $f_{(q,j)}$ sono le frequenze congiunte
- Le colonne e le righe interne della tabella sono le distribuzioni condizionate
- Dalla tabella si possono ricavare le due distribuzioni marginali (UNIVARIATE) della ***X*** e della ***Y***. Di ciascuna di tali distribuzioni marginali si possono calcolare, ad esempio, la media, la varianza, e lo scarto quadratico medio.

Per determinare se due variabili statistiche sono dipendenti o indipendenti bisogna utilizzare le distribuzioni marginali delle frequenze della tabella a doppia entrata.

La variabile ***X*** e la variabile **Y** sono indipendenti se la frequenza congiunta $\boldsymbol{f_{(q,p)}}$ (quella interna alla tabella), è il prodotto delle corrispondenti frequenze marginali, divise per il numero di dati n:

$$\boldsymbol{f}_{(q,p)} = \frac{\boldsymbol{f}_{(q,0)} * \boldsymbol{f}_{(o,p)}}{\boldsymbol{n}} \quad (1)$$

Se tale condizione non è rispettata, le due variabili NON sono indipendenti ma si dicono DIPENDENTI:

- Dipendenza perfetta di Y da ***X*** quando ad ogni modalità di ***X*** è associata una sola modalità di ***Y***
- Interdipendenza perfetta tra ***X*** e **Y** se a ogni modalità di uno dei due caratteri corrisponde una e una sola modalità dell'altro e viceversa.

LA DIPENDENZA STATISTICA

Per verificare se le variabili sono indipendenti si può costruire la tabella delle frequenze teoriche di indipendenza. Partendo dalle frequenze marginali e utilizzando la **formula (1)** si costruisce la tabella teorica delle frequenze che le due variabili dovrebbe avere se fossero indipendenti, se tale tabella coincide con quella data, le due variabili sono perfettamente indipendenti. La tabella così costruita si chiama tabella teorica di indipendenza.

La situazione di "perfetta indipendenza statistica" si verifica raramente, la tabella teorica di indipendenza rappresenta una situazione "ideale" , quello che serve è valutare quanto la tabella dei dati reali $\boldsymbol{freal}_{(q,p)}$ si discosta da quella di perfetta indipendenza per capire in che misura le due variabili sono dipendenti. A tal fine viene costruita la tabella di contingenza costruita dalla seguente differenza:

$$\boldsymbol{c}_{(q,p)} = \boldsymbol{freal}_{(q,p)} - \boldsymbol{f}_{(q,p)}$$

A questo punto è utile introdurre una misura in grado di quantificare il grado di associazione tra due caratteri ricorrendo all'indice Chi-quadrato espresso da:

$$\boldsymbol{X^2} = \sum_{q=1}^{H} \sum_{p=1}^{K} \frac{\boldsymbol{c}_{(q,p)}^2}{\boldsymbol{f}_{(q,p)}}$$

LA DIPENDENZA STATISTICA

- $\boldsymbol{X^2 > 0}$, esiste un'associazione tra caratteri
- $\boldsymbol{X^2 = 0}$, sussiste indipendenza tra ***X*** e ***Y***.
- il Chi-quadrato dipende dalla numerosità del collettivo e dal numero di modalità dei due caratteri, a tal fine è possibile calcolare l'indice di contingenza quadratica:

$$\boldsymbol{\theta^2 = \frac{X^2}{n}}$$

- L'indice V di Cramer:

$$\boldsymbol{V = \sqrt{\frac{\theta^2}{(\min[(H-1),(k-1)]}}}$$

- $0 \leq V \leq 1$, indipendenza tra X e Y
- V= 1 se vi è dipendenza o interdipendenza perfetta.