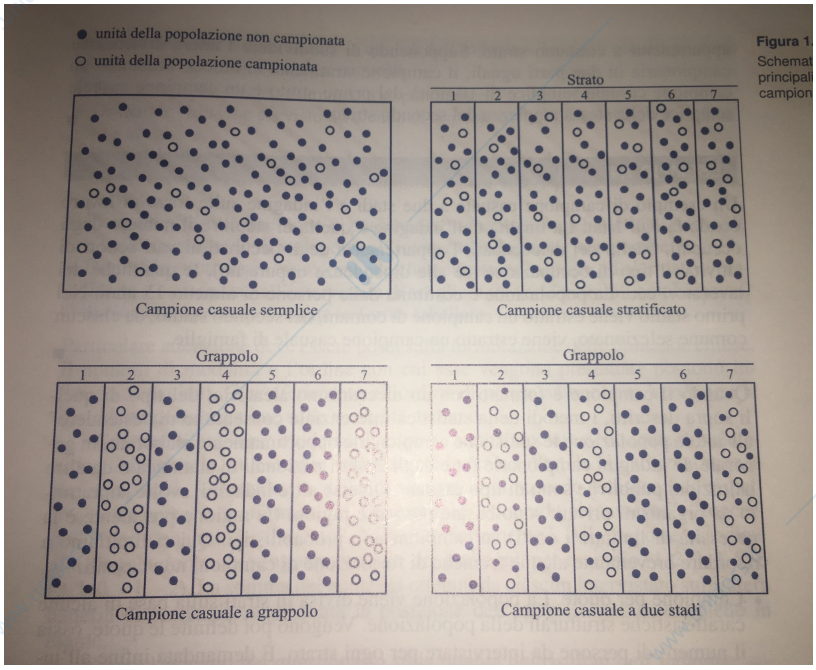
Statistica

I dati statistici traggono origine da un'attività internazionale rivola all'acquisizione di informazioni sul fenomeno o sui fenomeni di interesse.
Da qui dobbiamo parlare dei diversi processi che danno origine ai dati statistici:
- indagine statistica
- esperimento
- studio di osservazione o sul campo.

Indagini statistiche

Si ha quando lo studio statistico riguarda un collettivo statistico, le cui unità sono entità (persone, imprese ecc) individuabili ed osservabili che chiameremo popolazione reale o finita.
Si parla di indagine censuaria quando lo studio statistico è condotto con l'osservazione della totalità delle unità statistiche della popolazione e di indagine campionaria quando l'osservazione è limitata a una parte delle unità della popolazione, ossia ad un campione. Un esempio di indagine censuaria è il censimento della popolazione condotto dall'Istat ogni dieci anni.

Le modalità per formare un campione casuale sono molteplici:
- campione casuale -> le unità del campione sono selezionate con un meccanismo aleatorio (a caso) tale da assicurare a tutte le unità della popolazione la stessa probabilità di essere inserite nel campione. Se indichiamo con N il numero delle unità della popolazione e con n la numerosità del campione, l'estrazione di un campione casuale semplice equivale ad estrarre a sorte n palline da un'urna contenente N palline;
- campione sistematico -> dobbiamo supporre che il numero delle unità della popolazione N sia multiplo della dimensione del campione e che le unità siano messe in un certo ordine. Da qui dobbiamo introdurre un nuovo concetto, ovvero passo di campionamento che sarebbe l'inverso della frazione sondata p = N/n. Se r è il numero casuale estratto, si definisce campione sistematico l'insieme delle unità contraddistinte dai numeri (r, r+p+2p,…,r+(n-1)p). In certe condizioni il campionamento sistematico porta alla formazione di un campione molto simile al campionamento casuale semplice;
- campione casuale stratificato -> se avessimo informazioni aggiuntive sulla popolazione, suddivideremo la popolazione in più classi (sottopopolazioni) denominati strati, ognuno dei quali contiene unità tra loro omogenee secondo un certo criterio; poi da ciascun strato estraiamo un campione casuale semplice;
- campione casuale a grappolo -> suddividiamo la popolazione in un determinato numero di sottoinsiemi chiamati grappoli; di tali grappoli prendiamo un campione casuale semplice e consideriamo tutte le unità appartenenti ai grappoli prescelti;
- campione casuale a due stadi -> questo campionamento si limita a rilevare il campionamento casuale semplice estratto dal grappolo. In questo modo, si hanno due stadi di campionamento: nel primo stadio si estraggono i grappoli, nel secondo le unità statistiche oggetto di studio.



Vi sono comunque dei campioni non casuali:
- campione per quote -> la popolazione viene divisa in strati sulla base di alcune caratteristiche strutturali della popolazione. Vengono poi definite le quote ovvero il numero di persone da intervistare per ogni strato;
- campione a valanga -> si individua un primo gruppo di persone da intervistare che possieda le caratteristiche richieste. Queste persone a loro volta ne individuano altre che presentano le medesime caratteristiche;
- campione per testimoni privilegiati -> si tratta di individuare le persone che sono esperte sul fenomeno oggetto di studio. Questo tipo di campione è utilizzato in indagini che riguardano argomenti complessi e delicati.

Il questionario
Costituisce un insieme coordinato di domande da sottoporre a un campione di unità statistiche o all'intera popolazione, a seconda che l'indagine sia campionaria o censuaria. I singoli quesiti possono essere suddivisi in:
- domande aperte -> sono domande a risposta libera;
- domande chiuse -> elenco di risposte e se ne deve scegliere una;
- domande miste -> si può dare una risposta diversa da quelle previste;
- domande filtro -> consentono di passare direttamente da una batteria di domande ad un'altra;
- domande strutturate -> si ha possibilità di scegliere tra varie possibili combinazioni di risposte, spesso presentate in forma tabellare.

Somministrazione del questionario
la somministrazione del questionario si realizza con l'intervista che può essere diretta, telefonica, postale, via Internet.

Esperimenti
Si parla di esperimento quando persone, animali o oggetti vengono sottoposti a un trattamento per osservare su di essi la risposta, ossia la reazione al trattamento.

Per trattamento si intende una specifica condizione sperimentale nella quale le unità statistiche vengono osservate. La condizione sperimentale è determinata dal livello assunto da uno o più caratteri, detti variabili esplicative o fattori.

Studi di osservazione o sul campo
Esiste una situazione intermedia rispetto ai due processi generatori dei dati fin qui esaminati, situazione che si manifesta negli studi di osservazione o sul campo, in cui non esiste una popolazione finita da indagare, né vi sono unità statistiche che il ricercatore decide di assegnare ai diversi trattamenti: all'opposto, sono le unità stesse che si assegnano all'uno o all'altro trattamento.

Statistica descrittiva e inferenza statistica
Si usa suddividere il campo della statistica metodologica in due settori:
- statistica descrittiva -> i principi e i metodi della statistica descrittiva riguardano la programmazione delle indagini censuarie, la rilevazione dei dati, la costruzione delle distribuzioni di frequenze o di quantità, la presentazione di queste in forma grafica o tabellare, le elaborazioni statistiche mirate alla sintesi dei dati, come il calcolo di rapporti statistici, di indici di posizione, di variabilità e di forma;
- inferenza statistica -> si intende l'insieme dei metodi che ci permettono di generalizzare i risultati basati su un'osservazione parziale del fenomeno d'interesse, come nel caso delle indagini campionarie, dove viene analizzato un campione casuale estratto da una popolazione reale o come nel caso degli esperimenti o degli studi di osservazione, dove il campione casuale è generato dalla ripetizione dell'esperimento o dell'osservazione sul campo nelle stesse condizioni.
La statistica inferenziale si avvale di due metodologie fondamentali: la verifica delle ipotesi e la stima dei parametri, intesi come grandezze riassuntive della popolazione finita oggetto d'indagine. Entrambe le metodologie sono basate sul calcolo delle probabilità.

Bisogna introdurre alcune operazione utili: se a e b sono i livelli di uno stesso fenomeno, espressi nella stessa unità di misura, ma riferiti a situazioni diverse, il confronto tra le quantità a e b può essere effettuato tramite la differenza assoluta:

$$\text{differenza assoluta} = b - a$$

Dalla differenza assoluta poi si passa alla differenza relativa dividendo per a o per b:

$$b - a \,/\, a$$

Moltiplicando la differenza relativa per 100, si ottiene la differenza percentuale:

$$\text{differenza percentuale} = b - a \,/ a \times 100$$

Terminologia essenziale
I dati statistici sono numeri in un contesto. Per fare un esempio, la temperatura osservata in una certa stazione meteorologica alle ore 13 del 21 giugno non è un dato statistico, lo diventa se viene considerata nel quadro delle temperature rilevate nello stesso istante in altre stazioni meteorologiche della regione o del Paese, perché in questo caso si possono effettuare delle valutazioni e confronti di possibile interesse.

Si chiama collettivo statistico la molteplicità, l'insieme di casi individuali, in cui si manifesta il fenomeno oggetto di studio.

Si chiama unità statistica il caso individuale componente del collettivo statistico oggetto di studio.

Si chiama carattere ogni aspetto elementare, ogni caratteristica oggetto di rilevazione nelle unità statistiche del collettivo.

Si chiamano modalità del carattere i diversi modi con cui questo si manifesta nelle unità statistiche del collettivo. Devono essere esaustive, ossia devono rappresentare tutti i possibili modi di essere del carattere e non sovrapposte, ciò significa che ogni unità statistica si può associare a una sola modalità.

I caratteri sono di due tipi:
- qualitativi -> hanno modalità costituite da singole parole o da espressioni verbali;
- quantitativi -> hanno come modalità dei numeri.
Le modalità dei caratteri qualitativi possono essere sconnesse oppure ordinabili.
Sulla base di questa distinzione, si parla di:
- caratteri sconnessi quando le modalità non presentano un ordine naturale, ossia se date due sue modalità distinte è possibile solo affermare se queste sono uguali o diverse; il sesso, il luogo di nascita ecc sono esempi di caratteri qualitativi sconnessi. I caratteri sconnessi possono distinguersi in dicotomici quando possono assumere solo due modalità (es carattere del sesso); e politomici quando assumono un numero finito di modalità distinte (es colore di una stanza).
- caratteri ordinabili quando le modalità presentano un ordine naturale, ossia se date due sue modalità è possibile dare un rodine, specificando che una precede l'altra. Sono caratteri ordinabili ad esempio il grado di soddisfazione per un servizio erogato dalla comunità, posizione in una graduatoria o il titolo di studio ecc. Essi si distinguono in rettilinei se possiedono una modalità iniziale e una finale (es grado di soddisfazione), ciclici se non hanno vere e proprie modalità iniziali e finali (es il giorno della settimana o direzione del vento).
I caratteri quantitativi si dicono
- discreti se le loro modalità sono quantità distinte, individuabili ed elencabili, quasi sempre espresse da numeri interi. Es -> il numero di vani delle abitazioni;
- continui quando possono assumere tutti i valori di un certo intervallo di numeri reali. Es-> l'altezza delle persone.
I caratteri quantitativi si distinguono inoltre in trasferibili e non trasferibili a seconda che abbia o non abbia senso ipotizzare il trasferimento di parte del carattere da un'unità all'altra. Esempi di caratteri trasferibili il reddito e il patrimonio delle persone.

Distribuzioni statistiche
Consideriamo un collettivo statistico d N unità, dove sia osservato il carattere X. si chiama distribuzione unitaria semplice disaggregata (o unitaria) l'insieme delle osservazioni relative alle N unità del collettivo. Può essere presentata in questo modo: x_1, x_2.., x_n.
Questi simboli si riferiscono ad ogni diversa modalità del carattere.

Distribuzioni di frequenza

Le unità o osservazioni della distribuzione disaggregata vengono classificate e aggregate in gruppi omogenei sulla base di uno o più caratteri.

La scelta della modalità è condizionata dal livello di disaggregazione con cui i dati sono stati rilevati.

Si chiama distribuzione di frequenze lo schema con cui si associa a ciascuna modalità del carattere X la rispettiva frequenza. La distribuzione di frequenze viene rappresentata con una tabella:

X	Frequenza assoluta	X	Frequenza assoluta
x_1	n_1	x_1	n_1
x_2	n_2	x_2	n_2
...	...	...	...
x_j	n_j	x_j	n_j
...	...	...	...
x_K	n_K	x_K	n_K
Totale	n	**Totale**	n

dove n1, n2,....,nk sono le frequenze delle modalità x1, x2,...,xk.

Con la distribuzione di frequenze i dati passano dallo stato grezzo a una forma di presentazione organizzata e sintetica, fondamentale per la comunicazione dell'informazione, ma anche per la comprensione del fenomeno a cui i dati si riferiscono.

Per frequenza si intende il numero di volte che una data modalità si presenta nel collettivo di riferimento.

Vi sono vari tipi di frequenza:

- frequenze relative
- frequenze assolute (che abbiamo visto con la tabella sopra)
- frequenze cumulate
- frequenze percentuali
- frequenze per classi di valori

Frequenze relative

Si ottengono rapportando le frequenze assolute al totale delle unità N:

$f_i = ni / N$ i = 1, 2,…,k.

Le frequenze relative sono dei particolari rapporti di composizione e possono dare una valutazione rapida del "peso", dell'importanza, della singola modalità nell'ambito della distribuzione di frequenze.

Il totale deve essere sempre 1.

Le frquenze percentuali

Si ottengono moltiplicando per 100 le frequenze relative:

$p_i = f_i \times 100$ i = 1, 2,…, k.

Il totale deve dare 100.

Voto riportato	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
4	1	1/12 = 0,083	8,30%
5	3	3/12 = 0,25	25,00%
6	3	3/12 = 0,25	25,00%
7	3	3/12 = 0,25	25,00%
8	2	2/12 = 0,167	16,70%
Totale	12	12/12 = 1	100%

Frequenze cumulate

La frequenza cumulata associata ad una modalità del carattere misura il numero di casi che presentano un valore non superiore a quella modalità. (ci riferiamo a caratteri qualitativi)

x	frequenza cumulata assoluta	frequenza cumulata relativa	frequenza cumulata percentuale
A_1	$N_1 = n_1$	$F_1 = f_1$	$P_1 = p_1$
A_2	$N_2 = n_1+n_2$	$F_2 = f_1+f_2$	$P_2 = p_1+p_2$
...	...	...	...
A_j	$N_j = n_1+...+n_j$	$F_j = n_1+...+n_j$	$P_j = p_1+...+p_j$
...	...	...	...
A_k	$N_k = n_1+...+n_k = N$	$F_k = f_1+...+f_k = 1$	$P_k = p_1+...+p_k = 100$

Quando il carattere è quantitativo e il numero di osservazioni è elevato, la presentazione dei dati richiede che le modalità siano aggregate tramite la formazione di classi, cioè di intervalli numerici che comprendono più modalità.

Peso (kg)	Frequenze
40-50	4
50-60	7
60-70	3
70-80	2
80-90	1

WWW.OKPEDIA.IT

Dobbiamo fare una distinzione tra:

- classe aperta a destra e chiusa e sinistra [a, b) -> tali classi comprendono il valore xi e non il valore xi +1, esclude unità che presentano modalità esattamente uguali all'estremo sinistro;
- classe aperta a sinistra e chiusa e destra (a, b] -> tali classi comprendono il valore xi +1 e non il valore xi, esclude unità che presentano modalità esattamente uguali all'estremo destro della classe.

A seguito, definiamo:
- ampiezza della generica classe $C_i + 1$ la quantità d_i = differenza tra l'estremo destro superiore ed estremo sinistro inferiore -> $C_{i-1} - C_i$

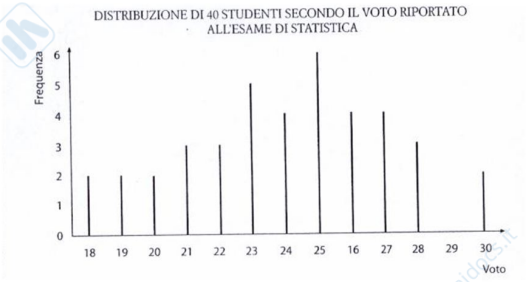
Valore centrale della generica classe $C_{i-1} - C_i$ la quantità:

x_i = somma degli estremi di classe / 2 = $C_{i-1} + C_i$ / 2

Rappresentazioni grafiche
Le rappresentazioni grafiche hanno lo scopo di illustrare le distribuzioni di frequenze. Rispetto alle tabelle, i grafici presentano dei vantaggi:
- consentono di visualizzare immediatamente le caratteristiche delle distribuzioni;
- rendono possibile il confronto tra più distribuzioni in spazi ristretti;
- agevolano l'investigazione dei fenomeni, mettendo in rilievo dati anomali, andamenti ecc;
- sono un efficace strumento per la divulgazione dei dati.

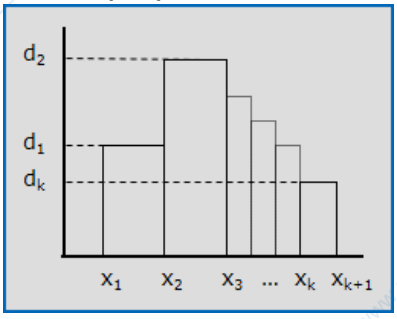
Distribuzioni di frequenze per caratteri quantitativi
Diagramma ad aste
La rappresentazione grafica si effettua ponendo sull'asse delle ascisse le modalità x1, x2,…, xk e sull'asse delle ordinate le frequenze corrispondenti n1, n2,…, nk.

Diagramma cartesiano ad aste (Leti, Cerbara, 2009)



Abbiamo rappresentato la distribuzione, ponendo sull'asse delle ascisse le modalità del carattere e sull'asse delle ordinate le frequenze

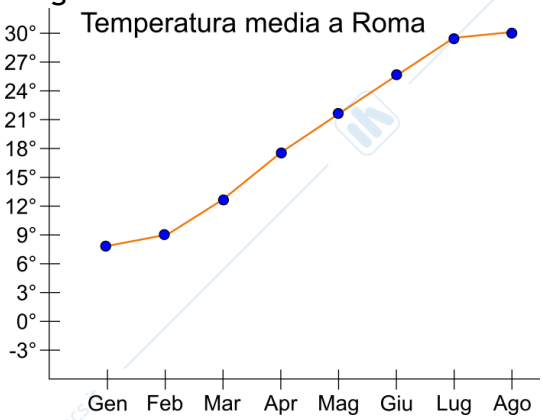
Istogramma di frequenza
è utilizzato quando la distribuzione si riferisce ad un carattere quantitativo continuo. In un sistema di assi cartesiani si collocano una serie di rettangoli che hanno come base l'ampiezza delle varie classi e come altezza le frequenze se le classi hanno la stessa ampiezza.
Se le classi sono di ampiezza diversa, le frequenze non sono direttamente confrontabili ed è necessario rendere proporzionale l'altezza.



La densità è il rapporto tra frequenza di una classe e la sua ampiezza:

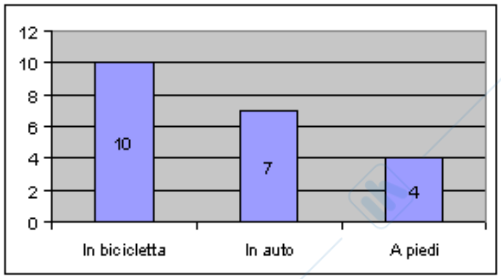
$$d = ni / xi_{+1} - xi = ni / xi$$

Diagramma cartesiano



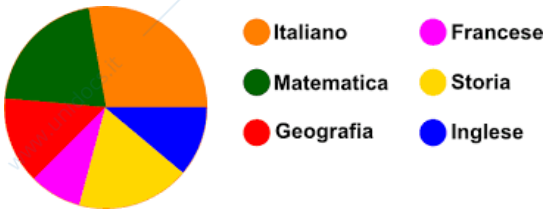
Con questa rappresentazione viene messo in luce il ritmo di accrescimento delle frequenze cumulate. Rappresenta il carattere quantitativo discreto.

Ortogramma



Rappresenta caratteri quantitativi continui.

Grafico a torta



è un tipo di rappresentazione usato per riprodurre una sola serie di dati.
La torta è suddivisa in tanti settori quanti sono gli elementi da visualizzare.
Per trovare l'angolo al centro biusogna fare:

$$\alpha i = 360^\circ \times ni/N$$

Medie

Le medie servono per sintetizzare le distribuzioni statistiche.
Vi sono vari tipi di medie:
- medie analitiche
- medie lasche.

Le medie analitiche sono quelle che si ottengono dall'applicazione di opportune

operazioni matematiche a tutti i valori del carattere che formano la distribuzione statistica presa in considerazione.
Rientrano in questa categoria.
- la media aritmetica
- media geometrica
- media armonica
- media quadratica.

Le medie lasche sono invece caratterizzate dal fatto che nel loro calcolo intervengono solo alcuni valori specifici della distribuzione, tipicamente quelli che occupano particolari posizioni nella graduatoria. Rientrano in questa categoria:
- la mediana
- i quartili
- i decili
- il valore centrale
- la moda.

Le medie analitiche sono applicabili solo a distribuzioni statistiche di caratteri quantitativi; le medie di posizione sono determinabili anche nel caso di distribuzioni statistiche di caratteri qualitativi a modalità ordinabili; la moda è definibile per qualsiasi tipo di carattere.

Media aritmetica

La media aritmetica è quel valore che possiamo attribuire singolarmente ad ogni unità lasciando invariato l'ammontare globale del carattere.

X1 + ….+ Xi + …+ Xn = M + …+ M...+ M

$$M_{aritmetica} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Le proprietà della media aritmetica:
- la somma dei termini della distribuzione è uguale alla media aritmetica moltiplicata per il numero di unità:

$$\sum_{i=1}^{n} x_i$$ = Nu.

- la media aritmetica è sempre compresa tra il minimo e il massimo delle modalità della variabile (soddisfa la condizione di internalità di Cauchy):

$$x_{(1)} \leq \mu \leq x_{(n)}$$

- la somma algebrica degli scarti della media aritmetica è nulla:

$$\sum_i (x_i - \mu) = 0$$

Gli scarti sarebbe la somma delle differenze tra i valori.
Vi possono essere scarti positivi e negativi e si compensano esattamente; quindi la media aritmetica è quel valore che bilancia esattamente scarti positivi e negativi.
- la somma del quadrato degli scarti rispetto alla media aritmetica è minima:

$$\sum (x_i - \mu)^2 = \min$$

La somma del quadrato degli scarti è più piccola rispetto a qualsiasi altro valore che non sia la media aritmetica.
- Se la variabile x presenta modalità x1,.., xi, x.., xn allora la variabile a + bx possiede la seguente media aritmetica:
1) se si aggiunge 0 si sottrae una costante a alla x, la rispettiva media sarà modificata dallo stesso ammontare;
2) se la variabile x è moltiplicata per un coefficiente b costante, la media risulterà moltiplicata per lo stesso ammontare.

$Y = a + bX$ $Y = a + bX$ - se un collettivo statistico di N unità venisse suddiviso in L sottoinsiemi disgiuntivi aventi numerosità N (1), N(2),.., N(L) e medie aritmetiche M(1), M(2),.., M(L), la media aritmetica del collettivo può essere calcolata così:

M = M(1) x N(1) + M(2) x N(2) + M(L) x N(L) / N(1) + N(2) + N(L)

Media quadratica
La media quadratica di una distribuzione statistica disaggregata x1, x2, …, xn è la radice quadrata della media aritmetica dei quadrati dei termini della distribuzione:

$$\mu_q = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 + \ldots + x_{N-1}^2 + x_N^2}{N}}$$

$$\mu_q = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$

Media Trimmed
In una serie di osservazioni di un fenomeno collettivo possono esserci osservazioni molto elevate e/o basse (dette outliers).
Poiché nel calcolo delle medie analitiche vengono considerati tutti i valori osservati, si ha che queste possono essere influenzate, in modo più o meno forte, dalla presenza di tali outliers che sono dei fenomeni anomali.
Nel caso siano presenti pochi outliers possono conservare la media Trimmed, che non fa altro che lavorare sulle modalità decrescenti.
Si prendono in considerazione solo le osservazioni NON outliers, quindi tutto ciò

che rientra tra il massimo e il minimo (senza prendere in considerazione queste due).

Medie analitiche per le distribuzioni di frequenze
La media aritmetica nel caso di distribuzione di frequenze (vedere o libro o quaderno)

$$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i n_i$$

La media aritmetica nel caso di distribuzione di frequenza con classi di valori
Viene utilizzato il valore centrale (vedere il quaderno o libro)

ale:

$$\mu = \frac{1}{N}\sum_{i=1}^{k} \tilde{x}_i n_i$$

Media quadratica per le distribuzioni di frequenze

$$m_q = \sqrt{\frac{\sum_{i=1}^{n} x_i^{\,2}}{n}}$$

SI ha la media aritmetica ponderata quando si attribuisce a ciascuna osservazione un peso che ne esalta o diminuisce l'importanza

$\sum_{i=1}^{k}$ = mettendolo sotto radice.

Mediana
La mediana Me (o m) è quella modalità ella variabile che bipartisce la distribuzione ordinata delle modalità. È la modalità che occupa il posto centrale nella distribuzione ordinata delle osservazioni.
Nel caso di distribuzione statistica disaggregata:

X1....Xi....XN

dopo aver ordinato:
X (1) ≤ ..... ≤ X (i)≤ ....≤ X(N)

Si possono avere due casi:
N dispari ----------> Me = X (N+1 / 2)
N pari ----------> Me = ˣ(N/2) + ˣ(N/2+1)
Bisogna ricordare che alla mediana non interessa del valore, ma solo della posizione.

Proprietà della mediana:
- è sempre compresa tra il massimo e il minimo delle modalità della variabile, soddisfa quindi la condizione di internalità di Cauchy:

$$X(1) \leq Me \leq X(N)$$

- è quel valore che minimizza la somma degli scarti assoluti:

$$\sum_{j=1}^{n} |x_j - Me| = minimo$$

Quartili e quantili
Si definiscono quartili quei valori che dividono la distribuzione in 4 parti di uguale numerosità.
La mediana è un quartile, più precisamente il secondo.

Mediana, quartili e quantili nel caso delle distribuzioni di frequenze
Mediana
- Si calcolano le frequenze cumulate:
N1 = n1
N2 = n1 + n2
Ni = n1 + n2 +…+ ni
Nk = n1 + n2 +…+ ni + …+ nk
- Si calcolano la mediana con le formule precedenti, a seconda che N sia pari o dispari.

Valore centrale
Il valore centrale è la media aritmetica dei valori estremi:

$$C = \frac{x_{(1)} + x_{(N)}}{2}$$

Moda
La moda Mo di una distribuzione di frequenze è la modalità che presenta la frequenza più alta.
- le distribuzioni statistiche che presentano un picco sono dette unimodali;
- le distribuzioni statistiche che presentano due picchi di uguale altezza sono dette bimodali;
- le distribuzioni che presentano più mode sono dette plurimodali.

Caso di distribuzioni di frequenze raggruppate in classi di valore
Mediana
Data una distribuzione di frequenze per un carattere suddiviso in classi la formula che dobbiamo usare è la seguente:

$$Me = Xi\text{-}1 + (N/2) - Ni\text{-}1 \;/\; ni \times (Xi - Xi\text{-}1)$$

Indici di variabilità
Per variabilità si intende l'attitudine dei fenomeni naturali e sociali a manifestarsi in modi differenti. La variabilità costituisce la ragione stessa dell'esistenza della statistica.
Esprime la pendenza delle unità statistiche di un collettivo ad assumere differenti

modalità del carattere.
La variabilità è tanto maggiore quanto più sono grandi le differenze tra le modalità o rispetto ad un valore caratteristico.
Abbiamo visto che la media è un indice che fornisce una sintesi della distribuzione di un fenomeno.
Ovviamente, la distribuzione è ben rappresentata dalla media quanto più le unità presentano modalità prossime ad essa.
Prendendo in considerazione l'esempio che abbiamo fatto (quello riguardo gli alberghi) noi vediamo che:
- nella distribuzione (A) tutti i termini sono uguali tra loro: la media fornisce una informazione sufficiente a descrivere completamente la distribuzione;
- nella distribuzione (B) i termini sono poco diversi: il numero di presenze negli alberghi, quando non uguale, è all'incirca equivalente;
- nella distribuzione (C) vi è una notevole diversità tra gli alberghi rispetto al numero di presenze;
- nella distribuzione (D) la diversità tra i termini è ancora più alta ed è massima, compatibilmente con N e μ, in (E), ove un solo albergo concentra su di sé tutte le presenze.
Le distribuzioni (A), (B), (C), (D), (E) sono presentate in ordine di variabilità crescente. Il carattere "numero di presenze" ha variabilità nulla in (A): i termini sono tutti uguali tra loro. Nella distribuzione (B) la variabilità è bassa, cresce in (C), è ancora più alta in (D) ed è massima in (E).
Conoscendo solo N e μ non sappiamo in quale situazione di variabilità si colloca il fenomeno nel collettivo in esame.
Inoltre si ha che la media aritmetica μ, pur essendo uguale in tutti i casi, ha una minore efficacia rappresentativa via via che si passa dal caso (A) al caso (E).
Dall'esempio visto si ha quindi che la misura della variabilità riduce la perdita di informazione che si ha considerando soltanto la media.
È necessario, quindi, affiancare alla media una misura di variabilità (indice di variabilità).

Una misura di variabilità deve avere le seguenti proprietà:
- Assumere valore minimo (essere nulla) quando tutte le unità della distribuzione presentano la medesima modalità del carattere;
- Crescere all'aumentare della "diversità" tra le modalità assunte dalle diverse unità.

Le misure di variabilità si possono distinguere in 2 categorie:
- Variabilità delle singole modalità rispetto ad un valore caratteristico (ad esempio, la media aritmetica, la mediana, etc.) mediante una sintesi degli scarti tra le singole modalità e il valore caratteristico preso come riferimento. In tal caso la variabilità è intesa come DISPERSIONE e parleremo quindi di MISURE DI DISPERSIONE.
- Variabilità reciproca (mutua) tra tutte le modalità considerate a 2 a 2. In tal caso la variabilità è intesa come DISUGUAGLIANZA e parleremo quindi di MISURE DI DISUGUAGLIANZA.

Vogliamo misurare di quanto i valori rilevati differiscono in media dalla grandezza che si è assunta a rappresentare l'intensità del carattere.
Supponendo di prendere la media aritmetica μ dei valori osservati come grandezza che li rappresenta, è intuitivo che la misura cercata possa essere data dalla quantità che sintetizza gli scarti tra le singole modalità xi e la grandezza μ; cioè è ragionevole effettuare un'opportuna media degli scarti (xi-μ).

L'indice di dispersione più importante per misurare la variabilità di una distribuzione è la cosiddetta VARIANZA, che esprime la media degli scarti (dalla media aritmetica) al quadrato. Nel caso di distribuzione statistica disaggregata (distribuzione unitaria o serie di osservazioni) la varianza è espressa nel modo seguente:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

varianza **varianza**

Ove il numeratore $\sum_{i=1}^{N}(x_i - \mu)^2$ rappresenta la DEVIANZA.
La varianza è sempre non negativa ed è un indice assoluto espresso nell'unità di misura del fenomeno al quadrato. La varianza varia tra 0 e infinito.

Una difficoltà nella interpretazione della varianza deriva dal fatto che essa è espressa nell'unità di misura del fenomeno al quadrato.
Possiamo quindi considerare lo SCARTO QUADRATICO MEDIO (STANDARD DEVIATION) che rappresenta la media quadratica degli scarti (dalla media aritmetica):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Si nota che lo scarto quadratico medio (s.q.m.) è la radice quadrata della varianza ed è espressa nella medesima unità di misura del carattere.

Calcolo semplificato della varianza:

$$\sigma^2 = \mu q^2 - \mu^2$$

Il calcolo semplificato dello scarto quadratico è lo stesso ma mettendo tutto sotto radice.

Proprietà della varianza:

- $\sigma 2 \geq 0$ (è uguale a zero se la distribuzione è degenere);
- Considerata la trasformazione lineare di xi con media μ e varianza σ2 (ad esempio trasformo in decimi il voto espresso in trentesimi):

$$yi = \alpha\, xi + \beta$$

si ottiene che la varianza della trasformazione lineare è:

$$\alpha^2\, \sigma^2.$$

Indici di variabilità relativi

Gli indici di variabilità finora considerati vengono chiamati indici di variabilità assoluti, nel senso che sono espressi nella stessa unità di misura dei termini della distribuzione.

Per questo motivo (e per altri) essi non sono sempre sufficienti per poter eseguire correttamente il confronto tra la variabilità di distribuzioni differenti.

Possiamo infatti distinguere, ad esempio, i seguenti casi:

- le modalità delle distribuzioni a confronto sono espresse in unità di misura diverse (tra le quali non intercorre alcuna relazione) (ad esempio, mt. e Kg; dollari e litri, etc.);
- le modalità delle distribuzioni a confronto sono espresse nella stessa unità di misura, ma le intensità medie di queste distribuzioni sono differenti (ad esempio, pesi delle madri e pesi dei neonati; prezzi di merci ordinarie e prezzi di merci pregiate; etc.).

In questi casi occorre considerare indici di variabilità relativi, che, contrariamente ai primi (assoluti) (espressi nell'unità di misura del carattere), sono numeri puri.

Se indichiamo con Va un indice di variabilità assoluto, un indice di variabilità relativo può ottenersi da una delle seguenti formule:

$Vr = Va/\mu$ $Vr' = Va/maxVa$

Vr si definisce indice di variabilità relativo rispetto alla media in quanto esprime la variabilità in termini della media;

Vr' si definisce indice di variabilità relativo rispetto al massimo in quanto esprime la variabilità in termini della variabilità massima.

Entrambi Vr e Vr' sono indipendenti dall'unità di misura in cui è espressa la variabilità statistica in esame.

Un importante caso particolare di indici di variabilità relativo alla media Vr è il coefficiente di variazione:

$$Cv = \frac{\sigma}{\mu} \quad Cv = \frac{\sigma}{\mu}$$

Altri indici di variabilità

Alcuni semplici indici di variabilità sono:

- Campo di variazione ------> $\Delta c = x(N) - x(1)$ oppure $\Delta c = yN - y1$ ove $y1 = x(1) = xmin$; $yN = x(N) = xmax$. Inconveniente: basandosi sui valori estremi, Δc è influenzato dall'eventuale presenza di outliers (valori anomali);
- Differenza interquartilica ---------> $\Delta q = q3 - q1$ ove q1=1° quartile; q3= 3° quartile. Δq è meno sensibile di Δc dall'influenza di eventuali outlier.

Altre misure di variabilità:

- misure di variabilità per caratteri trasferibili: misure di concentrazione;
- misure di variabilità per caratteri qualitativi: misure di eterogeneità.