ELEMENTI DI STATISTICA DESCRITTIVA

Si dice **induzione** o **metodo induttivo** il metodo di indagine scientifica caratteristico delle scienze sperimentali:
- si osservano fenomeni che si presentano spontaneamente o che vengono provocati con esperimenti,
- ci si domanda se tali fenomeni presentano qualche uniformità;
- in caso affermativo si cerca di formulare le leggi a cui questi fenomeni ubbidiscono.

Con tale metodo si procede dal particolare all' universale.

E' necessario anche il cammino inverso, perché una volta formulata una legge, ci si deve accertare, con nuove osservazioni, che sia effettivamente sempre valida.

Il metodo induttivo, quando opera in termini quantitativi, costituisce il metodo statistico: esso trova applicazione sia nel campo naturale (Fisica, Chimica, Biologia, ecc.), sia nel campo sociale (Demografia, Economia, Psicologia, ecc.).

Osserviamo che, mentre nel campo naturale i fenomeni possono essere oggetto, non solo di osservazione, ma anche di esperimento (ciò consente di fare osservazioni di un fenomeno molto numerose e nelle stesse condizioni), i fenomeni sociali non si possono provocare e la diversità delle condizioni ambientali in cui si verificano le diverse osservazioni porta a delle perturbazioni la cui influenza deve essere eliminata con appropriati mezzi statistici..
Analizziamo ora le varie fasi dell'indagine statistica:

FASE n°1 : FORMULAZIONE DELLE IPOTESI

In base ad osservazioni casuali, vengono formulate delle ipotesi che si vuole sottoporre a verifica sperimentale. Tale fase è compito dello studioso delle singole discipline.

La statistica si occupa invece delle modalità tecniche della rilevazione, dello spoglio e della elaborazione dei dati.

FASE n°2 : RILEVAZIONE DEI DATI

Premettiamo alcune definizioni.

Chiamiamo popolazione un insieme E di elementi di natura qualunque.

Gli elementi dell'insieme E si dicono individui o unità statistiche.

La popolazione E viene divisa in sottoinsiemi E1, E2, ..., En detti classi, disgiunti tra loro ed aventi come unione tutto E (ossia i sottoinsiemi E1, E2, ..., En formano una partizione di E: P(E).).

Ogni classe Ei viene individuata mediante un certo carattere comune agli individui che la compongono.

Ad ogni classe Ei appartiene un certo numero (eventualmente nullo) di individui, chiamato frequenza della classe : f(Ei).

Esempio di rilevazione dati:
Popolazione: alunni candidati all'esame di maturità di un ITIS (sono 338)
Dividiamo la popolazione in classi contraddistinte dal voto preso come carattere:

E1 insieme degli alunni che hanno preso 60
E2 insieme degli alunni che hanno preso 61
e cosi via
E42 alunni non maturi.

Ei	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
F(Ei)	20	5	7	20	10	12	10	8	20	2	30	3	9	6	8	20	8	8	6	2	10
Ei	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	<60
F(Ei)	4	10	7	1	12	2	0	7	6	9	4	10	0	3	10	5	2	5	0	10	7

.
La tabella definisce una funzione f chiamata funzione di distribuzione (perché dice come si distribuisce la popolazione fra le varie classi) dall'insieme P(E) in un insieme B di numeri reali così definita:
f: P(E) -------> B ad ogni classe Ei associo la sua frequenza, ossia il numero dei suoi elementi.
Ei -------> f(Ei)
Si dice che la funzione f caratterizza la distribuzione delle frequenze.

Date queste definizioni passiamo ad analizzare la fase n°2

Per fare una indagine statistica relativa ad una popolazione E si deve:

a) individuare il carattere in base al quale suddividere la popolazione in classi Ei.
Un carattere si dice **QUANTITATIVO** quando ogni classe è determinata da un numero (n° di elementi o misure di grandezze).
Si parla di **intensità** di un carattere quantitativo.
Un carattere si dice **QUALITATIVO** quando ogni classe è determinata da un aggettivo o nome (professione, colore ecc.).
Si parla di **modalità** di un carattere qualitativo.

b) individuare le unità statistiche aventi un dato carattere (si parla di **rilevazione statistica).**
Una rilevazione si dice COMPLETA quando comprende tutti gli elementi della popolazione E; **PARZIALE** quando riguarda solo una parte rappresentativa della popolazione detta **campione.**
Il campione deve essere scelto opportunamente.

La formazione del campione pone alcuni problemi:

-**PROBLEMA QUANTITATIVO**: di quanti elementi deve essere formato il campione? Bisogna contemperare le due esigenze di poco costo e sufficiente significatività. Per fare questo bisogna anche sapere come la dimensione del campione può influire sulla precisione dei risultati. Tale problema presenta notevoli difficoltà.
-**PROBLEMA QUALITATIVO**: una volta stabilite quante si deve stabilire come scegliere le unità statistiche. Si può procedere per estrazione a sorte: campione casuale (ogni individuo della popolazione ha ugual probabilità di essere estratto e ogni possibile campione ha la stessa probabilità

di essere formato); spesso si preferisce fare un campione stratificato (poiché il campione è usato per stimare le caratteristiche dell'intera popolazione da cui è estratto, questo dovrà essere scelto in modo da riflettere le caratteristiche della popolazione stessa).

-PROBLEMA DI INTERPRETAZIONE DEI RISULTATI: una rilevazione parziale pone notevoli problemi di interpretazione dei risultati: come estendere tali risultati all'intera popolazione? La moderna tecnica statistica ha fatto notevoli progressi in tale campo e si è preoccupata soprattutto di valutare l'attendibilità dei risultati parziali.

Di questo problema, che presenta notevoli difficoltà, ne daremo un cenno più avanti, parlando della variabilità.

c) fissare l'estensione (nel tempo e nello spazio) della rilevazione.

Fissato l'oggetto della rilevazione, bisogna predisporre il PIANO per realizzarla concretamente.

Si devono raccogliere le varie informazioni relative a ciascuna unità in appositi MODELLI DI RILEVAZIONE, che possono essere registri, schede, questionari, liste o tabelle (es tabelle relative alle analisi delle acque).

FASE n°3 : SPOGLIO DEI DATI

Le unità statistiche vengono raggruppate in classi omogenee Ei corrispondenti alle singole modalità dei caratteri qualitativi o alle singole intensità (spesso sono invece classi di intensità) dei caratteri quantitativi che interessano.

Una tabella statistica si forma associando ad ogni classe Ei la sua frequenza f(Ei).

Osservazione: il carattere quantitativo può essere continuo o discreto. Quando è continuo le classi Ei di intensità sono espresse in intervalli, se invece è discreto le classi Ei possono essere espresse dalle singole intensità, anche se spesso si ricorre ugualmente ad intervalli.

Una tabella che raccoglie dati statistici può essere rappresentata graficamente, ad es. con istogrammi, settori circolari, diagrammi cartesiani.

Rappresentazione con istogrammi :

Si raggruppano i valori della variabile indipendente in un certo numero di intervalli, detti intervalli di classe, (di ampiezza costante o meno), si riportano sull'asse orizzontale gli intervalli di classe e in corrispondenza si costruiscono dei rettangoli che hanno area proporzionale alle frequenze.

Se gli intervalli di classe sono tutti uguali anche le altezze sono proporzionali alle frequenze.

Esempio

Per rendere più leggibili i dati, dalla tabella statistica relativa agli esami di maturità possiamo ricavare una nuova tabella statistica estraendo le seguenti 5 fasce:

- dal 60 al 74 : 170 alunni
- dal 75 al 85 : 88 alunni
- dal 86 al 95 : 51 alunni
-dal 96 al 100: 22 alunni
- non maturi : 7 allievi

Otteniamo così una distribuzione ponderata per classi

voto	$60 \le voto \le 74$	$75 \le voto \le 85$	$86 \le voto \le 95$	$96 \le voto \le 100$	voto<60
N° alunni	170	88	51	22	7

Rappresentiamo i dati con un istogramma

(manca grafico)

Notiamo che le classi non hanno tutte la medesima ampiezza. L'ultima è aperta, si può pensare di chiuderla con un valore ragionevole, ad esempio 50.
Vedere rappresentazione grafica allegata n 1
Notiamo che in questo caso sono le aree dei rettangoli che sono proporzionali alle frequenze e non le altezze perché le classi non hanno tutte la medesima ampiezza.

Osservazione: essendo il carattere discreto si poteva costruire anche un **ortogramma**.

Rappresentazione a settori circolari:

Si divide un cerchio in un numero di settori circolari uguale al numero delle classi Ei considerate, ogni settore ha un'area proporzionale alla frequenza della classe Ei.
Per far questo viene disegnato un cerchio e lo si divide in tanti settori quanti sono i dati ; gli angoli di ogni settore sono proporzionali alle quantità che essi rappresentano.
In generale, per determinare l'angolo di ogni settore si utilizza la proporzione seguente:

frequenza del dato: frequenza totale = α °: 360°

Il grafico relativo al nostro esempio è:

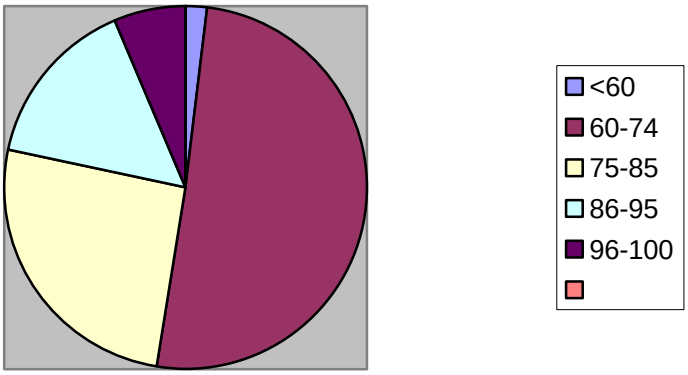

Diagramma cartesiano:

Si riportano nel piano i punti aventi per ascissa il numero che è il centro di ogni intervallo della classe e per ordinata le corrispondenti frequenze. Unendo con segmenti tali punti si ottiene la spezzata o poligono di frequenze. Nel caso, più frequente, che gli intervalli di base siano tutti uguali, l'area coperta dalla spezzata è la stessa di quella dell'istogramma. Sugli stessi dati della spezzata si può costruire una curva detta curva di frequenze o di distribuzione (quando l'ampiezza di ciascuna classe è molto piccola ed il numero totale dei casi osservati è sufficientemente grande la spezzata di frequenze tende ad una curva continua che si può studiare con metodi matematici).
Vedere rappresentazione allegata numero 3 (manca grafico)

FASE n°4 : **ELABORAZIONE DEI DATI.**

Ha più propriamente carattere matematico. Permette di esprimere i risultati dell'indagine statistica in modo sintetico e tale da facilitarne l'interpretazione.

Principali forme di elaborazione dati:

1 **RAPPORTI STATISTICI**: si intende un qualunque rapporto fra due dati statistici
. E' la più elementare forma di elaborazione dati.
Esempio: frequenze relative, dette anche rapporti di parte al tutto.
Altri esempi di rapporti statistici sono: rapporto tra i dimessi ed i ricoverati in un ospedale, tra i nati e i morti in una nazione, tra le entrate e le uscite in una ditta.

2 **LE MEDIE:** la media è un numero, che in qualche modo, ne riassume molti, e permette di avere una visione sintetica di questi, naturalmente nascondendo la varietà dei dati da cui proviene. Prenderemo in considerazione tre tipi di medie: la media aritmetica o valor medio, e due medie di posizione: la moda e la mediana.

Dicesi **media aritmetica** semplice degli n numeri x1,x2,...,xn quel numero $\bar{x}$ che sostituito a ciascuno di essi lascia inalterata la loro somma, cioè x1+x2+...+xn =nx da cui $\bar{x}$ =(Σ xi)/n
La media aritmetica esprime un valore di **equiripartizione** quando il carattere è additivo.
Dati m numeri x1,x2,...,xm, con rispettivi pesi o frequenze f1,f2,...,fm, dicesi loro **media aritmetica ponderata** il numero $\bar{x}$ che sostituito a tutti gli xi lascia inalterato il valore dell'espressione x1f1+x2f2+...+xmfm, ossia $\bar{x}$ = (Σxifi)/Σfi.

Osservazione: nel caso di distribuzione ponderata per classi è più conveniente , per calcolare la media aritmetica, utilizzare il centro della classe.
Quindi possiamo dare la seguente definizione:
La media di una distribuzione ponderata per classi è il quoziente tra lo somma dei prodotti del centro di ogni classe per la sua frequenza e la somma delle frequenze.

Esempio: calcoliamo la media aritmetica della distribuzione di frequeze relativa alla statistica dei voti dell'esame di maturità nei due modi:

Dati n numeri x1,x2,...,xn e detta M una qualunque loro media, si dice scarto da M ciascuno degli n valori s1=x1-M,..., sn=xn-M.
Gli scarti calcolati rispetto alla media aritmetica **x** godono di due importanti proprietà:
1) La loro somma è sempre nulla: cioè scarti + e scarti - si compensano.
Dimostrazione:

$$\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{x}=\sum_{i=1}^{n}x_i-n\bar{x}=\sum_{i=1}^{n}x_i-n\frac{\sum_{1=1}^{n}x_i}{n}=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i=0$$

2)La somma dei loro quadrati è minore della somma dei quadrati degli scarti calcolati rispetto a qualunque altro valore M.

Dimostrazione :

Considero la funzione f(M)= $\sum_{i=1}^{n}(x_i - M)^2$ e dimostro che ha un minimo in M = $\bar{x}$

Calcolo la derivata prima rispetto alla variabile M

f '(M)= $\sum_{i=1}^{n}[2(x_i - M)(-1)] = -2\sum_{i=1}^{n}(x_i - M) = -2\sum_{i=1}^{n}x_i + 2\sum_{i=1}^{n}M = -2\sum_{i=1}^{n}x_i + 2nM = -2n\bar{x} + 2nM$

Studio il segno della derivata prima

2nM-2n$\bar{x} \geq 0$ per M$\geq \bar{x}$ e quindi la funzione f(M) e decrescente per M< $\bar{x}$ e crescente par M> $\bar{x}$, per M = $\bar{x}$ ha quindi un minimo

Si dice **moda** o **valore normale** di una distribuzione di frequenze il valore xM avente maggior frequenza. E' molto importante perché rappresenta i casi più frequenti del fenomeno studiato.

Si dice **mediana** quel numero m tale che la somma delle frequenze fino ad m è uguale alla somma delle frequenze dopo m (è il valore che occupa il posto di mezzo).
La mediana si può calcolare solo se il carattere è quantitativo o qualitativo ordinabile.
Si mettono i dati in ordine crescente e, se sono in numero dispari, si prende quello che occupa il posto centrale, se sono in numero pari si fa la media aritmetica dei due dati centrali.
Se si vuole studiare il reddito annuo di ogni persona di una certa popolazione, la mediana m è quel numero tale che metà della popolazione ha un reddito <m e metà un reddito >m. In tal caso la mediana è più indicativa del valor medio, in quanto, se molte persone hanno un reddito basso e poche un reddito molto alto, la media aritmetica non ci dà uno specchio attendibile della distribuzione del reddito essendo molto influenzata da un numero piccolo di individui.
Le medie danno un idea sintetica del fenomeno studiato, ma non dicono assolutamente nulla della variabilità.

3 **LA VARIABILITA'**. Una prima semplice misura della variabilità è il **campo di variazione**; esso è la differenza fra il valore massimo ed il valore minimo osservati: R= Xmax - Xmin
E' una misura grossolana, infatti su di esso non influiscono i valori intermedi.
La variabilità si misura più adeguatamente considerando le differenze fra ogni termine ed un valor medio (cioè gli scostamenti o scarti). Di tali scarti si prende poi una media (per avere una sola misura sintetica della variabilità); bisogna però sempre trascurare i segni (altrimenti si avrebbe una compensazione tra differenze positive e negative), perciò si ricorre ai valori assoluti o ai quadrati.
La misura classica della variabilità è lo **scarto quadratico medio**
Siano xi i valori assunti dalla variabile statistica in esame (o i centri degli n intervalli delle classi se il carattere è continuo) ed fi le rispettive frequenze.

s.q.m. relativo all'intera popolazione $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n}}$ oppure $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2 fi}{\sum_{i=1}^{n} fi}}$

s.q.m. in un campione di n elementi $s = \sqrt{\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{n-1}}$ oppure $s = \sqrt{\frac{\sum_{i=1}^{n}(xi-\bar{x})^2\, fi}{\sum_{i=1}^{n} fi - 1}}$

(**s.q.m. corretto** che viene chiamato anche **deviazione standard**)

Osserviamo che n-1 è proprio il numero dei gradi di libertà del sistema (gli n scarti dalla media sono linearmente dipendenti). Fissato n, $\sigma < s$ e per n molto grande $\sigma = s$.
Si può dimostrare rigorosamente (teoria dei buon estimatori) che la media del campione è una buona stima della media della popolazione, mentre se si vuole una buona stima dello s.q.m. della popolazione dallo s.q.m. del campione bisogna usare la 2a formula (stimatore corretto)

La deviazione standard può anche essere espressa percentualmente ed allora prende il nome di **coefficiente di variazione** $CV = s/\bar{x}$ riportato in forma percentuale, ossia $CV = (100\, s\,\%)/\bar{x}$
In questo modo ottengo un numero puro indipendente dall'unità di misura e così posso confrontare dispersioni diverse.

La quantità $D = \Sigma (xi - \bar{x})^2$ viene detta **devianza**.
Allo s.q.m. spesso si preferisce la **varianza** che non è altro che il quadrato dello s.q.m. (ossia la media dei quadrati degli scarti), perché si evita di estrarre la radice.

Conclusione: la varianza, oppure lo s.q.m., individua la dispersione dei valori delle frequenze attorno alla media: se i valori sono concentrati intorno al valor medio la varianza è abbastanza piccola; viceversa, essa diventa grande se sono dispersi lontano dal valor medio.