Statistica breve introduzione

la statistica è un insieme di metodi di trattazione dei dati: come e quanti dati raccogliere, stabilire se un'esperienza consenta di ottenere conclusioni affidabili ecc.

Ricavare delle informazioni dai dati è cosa molto complicata e piena di trabocchetti, come del resto accade in tutte le scienze empiriche che pongono problemi e difficoltà di ordine superiore.

I metodi per ottenere risultati soddisfacenti nel delicato procedimento di passaggio dal campione alla popolazione sono studiati da quella parte della statistica detta **statistica induttiva** (o inferenza statistica).

Noi ci limiteremo a introdurre alcuni degli strumenti matematici utilizzati per descrivere i dati relativi a un certo gruppo scelto come popolazione. Ovvero ci occuperemo di quella che viene chiamata **statistica descrittiva**.

Curiosità : La Statistica ha questo nome perché all'inizio essa studiava principalmente i dati utili al governo degli Stati.

Statistica breve introduzione : che cos'è la Statistica

Come dicevamo, con il termine "statistica" indichiamo l'insieme di metodi di analisi quantitativa di un fenomeno variabile, per ricavarne un' informazione sintetica sul suo andamento.

La statistica si occupa dei modi di raccogliere e analizzare dati relativi a un certo gruppo di persone (gli studenti di una scuola, gli abitanti di un quartiere, gli elettori di una regione ecc.) o di oggetti (le automobili, i dischi, i libri ecc.), per trarne conclusioni e fare previsioni.

Le fasi fondamentali di un'indagine statistica sono due:

- rilevamento dei dati;
- elaborazione dei dati.

Il gruppo preso in considerazione viene detto **popolazione** o **universo**. Se la rilevazione dei dati viene effettuata su tutta la popolazione, parliamo di **censimento**.

Gli elementi di una popolazione si chiamano **unità statistiche**.

Spesso non consideriamo tutta la popolazione, ma soltanto una parte.

Tale parte viene detta **campione** ed è scelta in modo che rappresenti l'intero gruppo.

La maggior parte delle raccolte dati è di tipo campionario, cioè si riferisce solo ad un gruppo. Dai dati ottenuti da quel campione, poi, si cerca di ricavare risultati validi per tutta la popolazione.

Le tecniche utilizzate per la raccolta dei dati possono essere l'intervista diretta o indiretta. Nel caso di intervista indiretta, si possono ottenere le informazioni volute facendo compilare un questionario che viene poi spedito o consegnato a un incaricato dall'intervistato (pensa, per esempio, al censimento).

Si propongono di solito questionari anonimi con la sola richiesta dell'indicazione del sesso e dell'età. Una volta raccolti i questionari compilati,

- vengono contati, per stabilire il numero effettivo delle unità che costituiscono il campione;
- si contano le diverse risposte date a ciascuna domanda predisponendo tabelle di spoglio;
- i dati ottenuti vengono rappresentati graficamente;
- si elaborano i dati con i metodi matematici più opportuni;
- infine gli Statistici provvedono ad interpretare i dati e a trarre conclusioni che possano essere valide per tutta la popolazione.

Vedremo poi meglio come si svolge un'indagine statistica. Intanto torniamo ai concetti che ci torneranno utili.

Caratteri qualitativi e quantitativi

Come detto, gli elementi di una popolazione si chiamano anche **unità statistiche**. Possiamo studiare diverse caratteristiche di tali unità.

Ogni caratteristica rappresenta un carattere della popolazione.

Ogni carattere viene descritto mediante le **modalità** con cui esso si può manifestare.

Per esempio, se uno studente è di nazionalità italiana e un altro è di nazionalità francese, diciamo che nel primo il carattere «nazionalità» presenta la modalità «italiana», mentre nell'altro la modalità «francese».

Uno degli obiettivi della statistica è lo studio di come si distribuisce una data popolazione, in relazione a uno o più caratteri.

I caratteri possono essere di due tipi:

- **qualitativi**, se le loro modalità sono descritte da attributi (francese, italiano, biondo, moro,...);
- **quantitativi(VARIABILI)**, se le loro modalità sono descritte da numeri.

In altre parole: i ***caratteri qualitativi*** sono di solito *espressi in forma verbale*, spesso come *aggettivi*. *Invece i **caratteri quantitativi** sono espressi da numeri* (per esempio, la statura, il peso ecc.). Le modalità di un carattere quantitativo saranno, allora, espresse da numeri che si chiamano anche *valori* di quel carattere.

Un *carattere quantitativo* può essere, a sua volta:

- ***discreto**, se assume valori appartenenti a un insieme finito (o infinito numerabile)*. Esempi possono essere: il numero dei componenti di una famiglia, il numero delle nascite in un anno, il numero di iscritti in una scuola ecc.;
- ***continuo**, quando può assumere tutti i valori reali appartenenti a un certo intervallo* (a, b). Esempi di caratteri continui sono: il peso, le temperature, i volumi ecc.

ESEMPI:

Se vogliamo sapere con quali mezzi di trasporto gli studenti arrivano a scuola, useremo il carattere «mezzo di trasporto». Tale carattere ha più modalità: «treno», «autobus», «motorino», ...

E' un carattere qualitativo.

Se invece vogliamo conoscere la distribuzione delle altezze di una classe di studenti, useremo il carattere «altezza». Esso ha più modalità: 140 cm, 145 cm, 160 cm, ... L'altezza è un carattere quantitativo, essendo espresso con numeri.

FREQUENZA DI UN DATO

Come scopriremo nella prossima lezione, per l'analisi di una distribuzione di dati sono fondamentali dei numeri che ne rappresentano i valori medi, detti *indici centrali*, compresi fra il valore più piccolo e il valore più grande della distribuzione di dati; gli indici centrali più utilizzati sono la moda, la mediana, la media aritmetica e la media geometrica.

Alla base di tutti questi problemi c'è il fondamentale concetto di frequenza di un dato.

La **frequenza (assoluta)** di una modalità è il numero di volte in cui tale modalità si presenta.

In altre parole :

la *frequenza assoluta* (F) è il numero che indica quante volte un dato compare.

La funzione che associa ad ogni modalità di un carattere la rispettiva frequenza si chiama FUNZIONE DI DISTRIBUZIONE DELLE FREQUENZE.

Tale funzione si rappresenta di solito con una tabella a due colonne. Nella prima colonna vengono riportate le modalità, nella seconda le frequenze.

Nota : L'insieme delle coppie ordinate di cui il primo elemento è la modalità e il secondo la frequenza corrispondente viene detto **distribuzione di frequenza**.

Nell'ultima riga della tabella è poi utile riportare il totale delle frequenze, che rappresenta l'ampiezza del campione considerato.

ESEMPIO

In un questionario abbiamo chiesto a 28 operai di una fabbrica di indicare con le seguenti lettere i mezzi di trasporto con cui vanno di solito a lavoro:

- A: automobile;
- **M**: motorino o scooter;
- **P**: a piedi;
- **C**: bicicletta.
- **B**: autobus o pullman;

Abbiamo ottenuto i seguenti risultati:

A, B, M, M,P, A, A, B, P, B, C, A, B, B, B, C, P, B, A, C, C, A, M, B, M, B, A, C.

Contiamo quante volte si presenta ciascuna modalità, ovvero la sua frequenza. Costruiamo la seguente tabella di frequenza.

distribuzione di frequenza	
MODALITA'	FREQUENZA (assoluta)
automobile	**7**
a piedi	**3**
autobus/pullman	**9**
motorino/scooter	**4**
bicicletta	**5**
Totale delle unità statistiche	***28***

Nota : L'insieme delle coppie ordinate di cui il primo elemento è la modalità e il secondo la frequenza corrispondente viene detto **distribuzione di frequenza**.

IN GENERALE. Se X è un generico CARATTERE e $x_1, x_2, \ldots, x_k$ sono le MODALITÀ osservate di quel carattere su n individui, indichiamo con $f_1, f_2, \ldots, f_k$ le frequenze delle varie modalità. La distribuzione di frequenze del carattere X si può rappresentare con una tabella che può essere di due tipi:

X	FREQUENZA (ASSOLUTA)
x_1	f_1
x_2	f_2
…	**…**
x_k	f_k
Totale delle unità statistiche	***n***

X	x_1	x_1	x_1	x_1	*Totale unità statistiche*
Frequenza	f_1	f_1	f_1	f_1	n

Se il carattere è di tipo QUANTITATIVO le modalità vanno ordinate in senso crescente o decrescente

Le serie e le seriazioni statistiche

Le tabelle che riportano nella prima colonna le modalità di un carattere *qualitativo* vengono dette **serie statistiche**. Nella seconda colonna può comparire o il *numero di volte in cui una modalità si presenta* (*frequenza*) o la sua *misura* (*intensità*), che può essere considerata come un tipo particolare di frequenza.

Le tabelle che invece riportano nella prima colonna un carattere *quantitativo* vengono dette **seriazioni statistiche**. Nella seconda colonna compare la frequenza, cioè il numero di volte in cui si presenta la relativa modalità

Le tabelle a doppia entrata

Questo tipo di tabelle ci permettono l'osservazione delle unità statistiche sotto due modalità. Quando entrambe le modalità sono quantitative, parliamo di **tabelle di correlazione**. Se almeno una delle modalità è qualitativa, si hanno **tabelle di contingenza.**

Ne vedremo diversi esempi in seguito

LE CLASSI DI FREQUENZA

In alcuni casi, per semplificare la rappresentazione dei dati, è utile accorpare le varie modalità in INTERVALLI, detti CLASSI, e poi costruire la distribuzione di frequenze delle classi.

Studiamo per esempio l'altezza di un gruppo di studenti. Otteniamo la seguente tabella con i dati "grezzi":

NUMERO D'ORDINE	MISURA ALTEZZA (in metri)
1	1,56
2	1,64
3	1,62
4	1,68
5	1,69
6	1,76
7	1,75
8	1,72
9	1,61
10	1,69
11	1,65
12	1,73
13	1,68
14	1,67
15	1,66
16	1,60
17	1,64
18	1,67
19	1,74

In casi come questo, è utile raggruppare le modalità in **classi**, determinando la frequenza di ogni classe. Per esempio, è utile suddividere le possibili altezze in cinque intervalli:

(1,55-1,60], (1,60-1,65], (1,65-1,70], (1,70-1,75], (1,75-1,80]

Di solito l'estremo inferiore di ciascuna classe viene considerato escluso dalla classe, mentre quello superiore incluso. Per esempio, nella tabella "CLASSI DI FREQUENZA", il valore 1,60 è relativo alla classe 1,55-1,60 e non alla classe 1,60-1,65.

Il raggruppamento in classi fornisce meno informazioni (per esempio, non sappiamo quanto misurano esattamente le 7 altezze comprese fra 1,65 e 1,70 m), però fornisce una sintesi più leggibile del fenomeno.

Di ogni classe è spesso utile calcolare il **valore centrale**, che si ottiene dividendo per 2 la somma degli estremi della classe. Per esempio, il valore centrale della classe 1,60-1,65 è (1,60+1,65)/2, ossia 1,625.

Una volta raggruppati i dati nelle classi, possiamo costruire una tabella in cui associamo ad ogni classe la sua frequenza assoluta (e quella relativa, che introdurremo a breve).

CLASSI DI FREQUENZA		
CLASSE	**FREQUENZA**	**FREQUENZA RELATIVA PERCENTUALE**
1,55-1,60	2	11%
1,60-1,65	5	26%
1,65-1,70	7	37%
1,70-1,75	4	21%
1,75-1,80	1	5%

Spesso interessa il valore della frequenza confrontato con il numero totale delle unità statistiche. Infatti siamo in situazioni diverse se, per esempio, la frequenza di una modalità è 7 rispetto a un totale di 28 o se, invece, è 7 rispetto a un totale di 280. Per questo motivo viene calcolata la **frequenza relativa**, di cui diamo la definizione.

DEFINIZIONE : Frequenza relativa
La frequenza relativa di una particolare modalità è il rapporto fra la frequenza della modalità stessa e il numero totale delle unità statistiche.

$$f = \frac{F}{tot} = \frac{Frequenza}{totale\ unità\ statistiche}$$

Per esempio, la frequenza relativa delle persone che vanno a lavoro in auto è

$$f = \frac{F}{tot} = \frac{7}{28} = 0{,}25$$

La frequenza relativa può essere espressa anche in **percentuale**, moltiplicandola per 100: la frequenza percentuale della modalità automobile è 25%. Questo significa che, in una distribuzione con le stesse caratteristiche di quella data, su un campione di 100 operai, 25 vanno al lavoro in automobile

Anche per le frequenze relative di un carattere, possiamo costruire la distribuzione delle frequenze relative o percentuali.
Riprendendo la distribuzione delle frequenze precedente, possiamo completarla come segue, inserendo la distribuzione delle frequenze relative e percentuali.

DISTRIBUZIONE DELLE FREQUENZE RELATIVE			
MODALITA'	FREQUENZA	**FREQUENZA RELATIVA**	**FREQUENZA RELATIVA PERCENTUALE**
automobile	**7**	**0,25**	**25%**
a piedi	**3**	**0,11**	**11%**
autobus/pullman	**9**	**0,32**	**32%**
motorino/scooter	**4**	**0,14**	**14%**
bicicletta	**5**	**0,18**	**18%**
Totale delle unità statistiche	***28***	***1***	***100%***

Nota bene : La somma delle frequenze relative alle diverse modalità è 1, in percentuale è 100%.

FREQUENZA CUMULATA:

Nel caso di un carattere quantitativo, si definisce "FREQUENZA CUMULATA" relativa ad una data modalità, la somma delle frequenze di tutte le modalità minori o uguali ad essa.

Per esempio, nel caso del carattere altezza, se vogliamo sapere quanti ragazzi hanno altezza minore o uguale a 1,75 m, ci basta SOMMARE le frequenze assolute di tutte le modalità minori o uguali a 1,75 m.

Anche per le frequenze cumulate possiamo costruire la distribuzione delle frequenze cumulate di un carattere.

Dalle frequenze relative alle frequenze
Se vengono forniti le frequenze relative *f* e il numero totale *T* delle unità statistiche, è possibile calcolare le frequenze *F* di ogni modalità:

$$F = f \cdot T.$$

La frequenza di una modalità è il prodotto tra la frequenza relativa e il numero totale delle unità statistiche.
ESEMPIO
Se sappiamo che, in un campione di 3500 persone, il 27% ha guardato una certa trasmissione televisiva, il numero delle persone del campione che ha guardato la trasmissione è 0,27 • 3500=945.