

Introduzione

- La STATISTICA è una scienza che consente di trarre informazioni efficaci in una popolazione limitandosi ad esaminare solo alcuni elementi

POPOLAZIONE: tutte le unità statistiche sulle quali si può ottenere una misura dell'unità di interesse

CAMPIONE: sottosamme della popolazione opportunamente rappresentativa

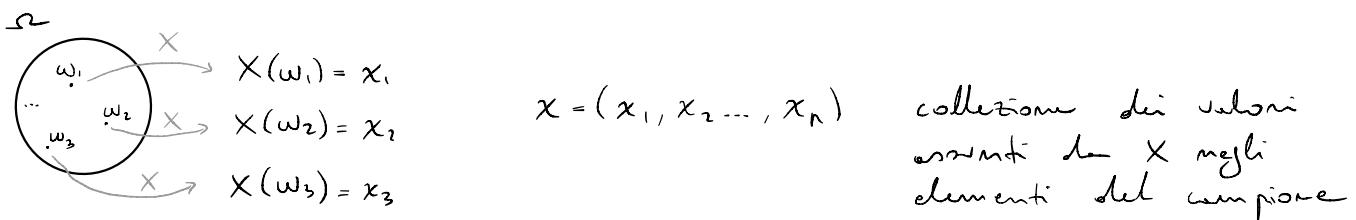
BIAS: errore commesso sistematicamente

DATI UNIVARIATI

- Considera una variabile alla volta

CAMPIONAMENTO: passo dalla popolazione a cui (unità statistiche)
→ si misurano una caratteristica (variabile aleatoria)

CAMPIONE → $X = (X_1, X_2, \dots, X_n)$ collezione di V.A. identicamente distribuite e indipendenti



MISURAZIONE: applicare la V.A. ai campioni estratti

LIVELLI DI MISURA E TIPOLOGIE DI DATI

DATI DI TIPO QUANTITATIVO

Numerici

- **Discrete Data**

Assume un numero finito o numerabile di dati

- **Continuous Data**

Non è possibile enumerarli

DATI DI TIPO QUALITATIVO

Non vengono espressi tramite numeri ma nomi

- **Nominal Data**

Valori puramente nominali

- **Ordinal Data**

I dati sono categorie che possono essere ordinate

- **Character Data**

Dati di tipo identificativo

- **Date and Time Data**

FATTORI

Statistica Descrittiva

- DATI NUMERICI UNIVARIATI -

DATA SET: riarrangiamento dei dati in una tabella con i dati (data frame) nelle righe e le variabili nelle colonne

QUANTILI CAMPIONARI - $\text{quantile}(x, \alpha)$

q_α : valore per cui $\alpha \cdot 100\%$ dei dati sono inferiori \rightarrow quantile di ordine α
La mediana è il quantile di ordine 0.50 $\rightarrow q_{0.50}$

QUARTILI CAMPIONARI - $\text{quantile}(x)$

$$Q_1 = q_{0.25}$$

$$Q_2 = q_{0.50}$$

$$Q_3 = q_{0.75}$$

Z-SCORE

Li dice quanto ogni singolo dato sia grande o piccolo rispetto agli altri

$$z_i = \frac{x_i - \bar{x}}{s}$$

COEFFICIENTE DI VARIAZIONE

$$CV = \frac{s}{\bar{x}} \%$$

$CV < 1$ distribuzione poco disperse
 $CV > 1$ distribuzione disperse

① INDICI DI POSIZIONE

- Medie
- Mediana
- (moda)

② INDICI DI VARIABILITÀ

- Range
- I.Q.R.
- Standard Deviation (Varianza)

③ INDICI DI FORMA

- Skewness
- Kurtosis

① INDICI DI POSIZIONE

MEDIA CAMPIONARIA - `mean(var$wt, na.rm=T)`

// `na.rm=T` pu toglie i dati not available

Dati $x = (x_1, x_2, \dots, x_n)$

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

differenza tra x_i e la media

- La media è influenzata dai valori estremi
- La media degli scarti dalla media è uguale a 0

MEDIANA CAMPIONARIA - `median(var$wt, na.rm=T)`

Valore centrale, maggiore di metà dei dati e minore dell'altra metà

$(x_1, x_2, \dots, x_n) \rightarrow (x_{(1)}, x_{(2)}, \dots, x_{(n)})$

non sono gli stessi x_n ma sono i valori ordinati

CASO n DISPARI: $\frac{n}{2} + 1$

CASO n PARI: $(\frac{n}{2} + (\frac{n}{2} + 1))/2$

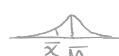
• MEDIANA = MEDIA $\bar{M} = \bar{x} \Rightarrow$ distribuzione simmetrica



• MEDIANA < MEDIA $\bar{M} < \bar{x} \Rightarrow$ media influenzata da valori molto grandi



• MEDIANA > MEDIA $\bar{M} > \bar{x} \Rightarrow$ media influenzata da valori molto piccoli



MODA

• Valore con frequenza di occorrenza più alta.

Viene usata per dati quantitativi continuvi divisi in classi o per dati qualitativi

② INDICI DI VARIABILITÀ / DEVIAZIONE

RANGE

$$\max_{\text{DATI}} - \min_{\text{DATI}}$$

INTER-QUARTILE RANGE

$$Q_3 - Q_1 \quad (\text{50% centrale dei dati})$$

VARIANZA CAMPIONARIA - $\text{var}(x, \text{na.rm} = T)$

$$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}$$

sensibile a grandi deviazioni
dalla media, ha dimensione x^2

DEVIAZIONE STANDARD - $\text{sd}(x, \text{na.rm} = T)$

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

③ INDICI DI FORMA

KURTOSI

misura delle code della distribuzione

$$K = n \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (-3) \quad \text{valore della distribuzione normale}$$

- $K > 0$: code più pesanti
- $K < 0$: code più leggere

SKEWNESS - indice di esimmetria

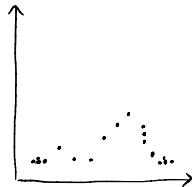
$$\Sigma = \frac{1}{n} \sum_{i=1}^n z_i^3 \quad \leftarrow z\text{-score}$$

- $\Sigma > 0$: esimmetria a destra
- $\Sigma < 0$: esimmetria a sinistra
- $\Sigma = 0$: simmetria

Rappresentazione Grafica dei Dati

DOT PLOT

Utilizzato principalmente per campioni di piccola taglia



ISTOGRAMMI - hist(var, freq=F)

// freq=T ci rappresenterebbe le frequenze assolute

- sull'asse delle ascisse segna i valori del Range
- l'asse x è diviso in bin, solitamente equipotenti
- Se ogni bin viene costruita una linea la cui altezza è proporzionale alle frequenze relative nelle classi che corrispondono ai bin
- sull'asse delle ordinate la stima della densità



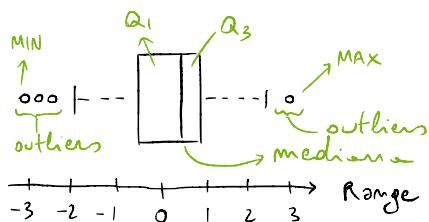
Σ area delle barre = 1 (corrispondenza empirica della densità della variabile continua)

REGOLA EMPIRICA PER IL N° DI BIN: $\approx \sqrt{n}$

$$\approx \frac{\text{range}}{1 + \log(n)}$$

BOXPLOT - boxplot(var, horizontal=T)

// box and whiskers che formiscono 1,5 volte il valore di std dev



- Se ci sono outliers gli estremi dei whiskers sono $Q_1 - 1,5 \text{ IQR}$, $Q_3 + 1,5 \text{ IQR}$ con $\text{IQR} = Q_3 - Q_1$.

Statistica Descrittiva

- VARIABILI QUALITATIVE -

dati categorici (categoriali, qualitativi)

ANALISI IMMEDIATA

- [FREQUENZE
GRAFICI]

① FREQUENZE

FREQUENZE ASSOLUTE - `table(x)`

Numero di occorrenze di ogni categoria a livello del campione
(non si considerano i dati mancanti)

FREQUENZE RELATIVE - `table(x) / sum(table(x))`

Quoziente tra le frequenze assolute e la loro somma

FREQUENZE PERCENTUALI

Frequenze relative moltiplicate per 100

② GRAFICI

BARPLOT

Altura delle barre proporzionale alle frequenze

DOTCHART

Poco usate, simili ai barplot

PIECHART - `pie(x)`

Gráfico a torta con segmenti proporzionali alle frequenze

Statistica per Dati Bivariati

In generale per più variabili misurate alle stesse unità statistiche si parla di dati multivariati

VARIABILE QUANTITATIVA + VARIABILE QUALITATIVA

Possiamo valutare per ogni gruppo:

- Indici
- Boxplot
- Histogrammi

→ divide il campione globale in gruppi determinati dai suoi livelli o categorie

VARIABILE QUANTITATIVA + VARIABILE QUANTITATIVA

Possiamo analizzare:

- Grafici
 - . Scatterplot
- Indici di Correlazione
 - . Pearson
 - . Spearman

SCATTERPLOT - $\text{plot}(x, y)$

Si vede se si può misurare ai punti del grafico come disposti attorno ad una retta

INDICE DI CORRELAZIONE DI PEARSON - $\text{cor}(x)$

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \rightarrow \begin{array}{l} \text{prodotto delle variabili} \\ \text{standardizzate (z-score)} \end{array}$$

$$-1 \leq r \leq 1 \iff |r| \leq 1$$

- $r \approx 0$ *nuova conoscenza*, non ci sono pattern, non c'è relazione
- $r > 0$ dipendenza diretta
- $r < 0$ dipendenza inversa

Il valore di r dà un'indicazione della forza della relazione tra le due variabili

INDICE DI CORRELAZIONE DI SPEARMAN - $\text{cor}(x, \text{method} = "spearman")$

indice di monotonia

Dà un'indicazione di quante osservazioni hanno la stessa concordanza

VARIABILE QUALITATIVA + VARIABILE QUALITATIVA

TABELLE DI CONTINGENZA / A DOPPIO INGRESSO

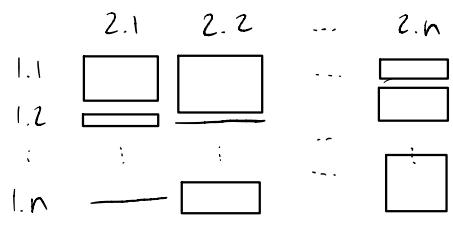
		Variabile 2				
Variabile 1		Categoria 2.1	Categoria 2.2	...	Categoria 2.n	
Categoria 1.1	n_{11}	n_{12}	...	n_{1n}		
Categoria 1.2	n_{21}	n_{22}			n_{2n}	
:	
Categoria 1.n	n_{n1}	n_{n2}			n_{nn}	
						n

n_1 : n° di categorie 1° variabile
 n_2 : n° di categorie 2° variabile

↙ taglio del campione

- Può essere visualizzato con un MOSAIC PLOT

MOSAIC PLOT - mosaicplot(table(x,y))



UNIVAR. Rettangoli con area proporzionale alle frequenze di ogni categoria

BIVAR. Rettangoli che compongono un rettangolo globale con area proporzionale alle freq. condizionate

τ DI KENDALL - cor(x,y,method="Kendall")

Misura l'associazione tra dati che possono essere ordinati (anche qualitativi)

$$\tau_A = \frac{\# \text{coppie concorrenti} - \# \text{coppie discordanti}}{tagli del campione} \rightarrow n(n-1)/2$$

campione

concordi: entrambe le coordinate sono maggiori o minori

- Si somma per tutti i punti