Descrizione numerica dei dati

Rappresentare graficamente i dati non è l'unica soluzione possibile per poterli studiare, infatti possono essere compresi attraverso:

- Misure di tendenza centrale
- Misure di variabilità
- Misure delle relazioni tra due variabili numeriche

Misure di tendenza centrale

Come nel caso dell'istogramma servono a vedere il punto in cui i dati tendono ad accentrarsi in modo da avere informazioni quantitative sull'osservazione tipica dei dati. Misure di tendenza centrale sono:

- Media → è la somma dei valori di tutte le osservazioni diviso il numero di esse. La media si suddivide in sua volta in:
  - Media della popolazione→ se tiene conto di tutta la popolazione
  - Media campionaria→ se tiene conto solo di un campione di popolazione
- Mediana → per poter utilizzare la mediana dobbiamo ordinare i dati in modo crescente o decrescente. È l'osservazione centrale di un insieme di osservazioni ordinate e si trova nella posizione 0.50 (n+1) della sequenza ordinata. Un lato negativo della mediana è che dati diversi possono avere la stessa mediana, ma è un indice robusto in quanto non è influenzato da valori estremi
- Moda→ la moda, se esiste, è la modalità che si presenta il maggior numero di volte. Può essere utilizzata in caso di dati di tipo qualitativo o quantitativo. Non è influenzata da valori estremi e ci possono essere più mode.
- Simmetria nei dati

Misure di variabilità

Le misure di tendenza centrale da sole non bastano per studiare i dati, per questo motivo si fa uso anche di:

- Campo di variazione→ differenza tra il massimo valore osservato e il minimo. Questo campo aumenta con l'aumentare con la variabilità dei dati dal centro della distribuzione. Questa osservazione a volte può non essere molto efficace in quanto se i dati presi in considerazione sono outlier si rischia che sia poco affidabile. Un altro difetto di questo metodo è che ignora il modo in cui i dati sono distribuiti
- Differenza interquartile→ per poter determinare questa differenza dobbiamo seguire diversi passi:
  - Passo 1) Ordinare l'insieme di dati in ordine crescente
  - Passo 2) Trovare il punto medio che divide la serie di dati a metà
  - Passo 3) Determinare la mediana del primo e terzo quartile delle due serie (eliminando il valore corrispondente al punto medio che si trova solo nella serie con un numero dispari di valori).
  - Passo 4) Determinare lo scarto interquartile campionario sottraendo i valori Q3 – Q1.

queste misure fino ad ora trattate tengono conto solamente di alcuni valori.

- Varianza→ ci sono due tipologie di varianza
  - Varianza della popolazione→è la somma delle differenze, al quadrato, tra ciascuna osservazione e la media della popolazione, divisa per la dimensione della popolazione n
  - Varianza campionaria→è la somma delle differenze, al quadrato, tra ciascuna osservazione e la media del campione, divisa per la dimensione del campione, n, meno 1. Il meno uno finale serve a rendere la varianza campionaria uno stimatore non distorto
- Devianza standard→ misura la variabilità comunemente usata e la mostra rispetto alla media. Ha la stessa unità di misura dei dati originali. È molto sensibile ai valori anomali, una sua alternativa robusta è lo scarto interquartile. Esistono due tipi di questo indice:
  - Devianza standard della popolazione→è la radice quadrata della varianza della popolazione
  - Devianza standard campionaria→è la radice quadrata della varianza campionaria
- Disuguaglianza di Chebychev→per un insieme qualunque si scelga arbitrariamente un valore k>=1, allora la proporzione di dati nell'intervallo [media+-k (deviazione standard)] è almeno $1-\frac{1}{k^2}$. Indipendemente da come i dati sono distribuiti almeno $1-\frac{1}{k^2}$ dei valori cadranno entro k deviazioni standard dalla media.

- Coefficiente di variazione → esprime la deviazione standard come una percentuale della media ed è una misura di variabilità relativa. Si divide in:
  - Coefficiente di variazione della popolazione→ (deviazione standard/media) *100
  - Coefficiente di variazione campionario→ (deviazione standard/media campionaria) *100

Media ponderata

Viene utilizzata nel caso in cui i dati siano raggruppati. Si ottiene facendo peso*dato per ogni dato e sommandoli tutti insieme, il risultato di questa operazione va diviso per la somma di tutti i pesi.

Misure delle relazioni tra variabili

Sono strumenti per determinare una relazione lineare e misurarne la direzione, questi strumenti sono:

- Covarianza→ è una misura della relazione lineare tra due variabili. Un valore positivo indica una relazione diretta, al contrario un valore negativo indica una relazione inversa. Ne esistono due tipi:
  - Covarianza della popolazione→ $\mathrm{Cov}(x,y)=\sigma_{xy}=\frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}$
  - Covarianza campionaria→ $\mathrm{Cov}(x,y)=s_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

X e y sono i valori osservati mentre n è la popolazione o la grandezza del campione. Il valore della covarianza tuttavia non è utile per misurare l'intensità della relazione lineare tra variabili perché dipende dall'unità di misura.

- Coefficiente di correlazione lineare→ viene utilizzato per misurare l'intensità e la direzione della relazione lineare tra variabili in quanto è una misura standardizzata. È calcolato dividendo la covarianza per la deviazione standard delle due variabili. Ne esistono due tipi:

- Coefficiente di correlazione lineare della popolazione→ $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- Coefficiente di correlazione lineare campionario→ $r_{XY} = \frac{s_{XY}}{s_X s_Y}$

La covarianza e il coefficiente di correlazione lineare hanno lo stesso segno.
Esiste una relazione lineare se $|r| > \frac{2}{\sqrt{n}}$.

Il valore del coefficiente di correlazione varia da 1 a -1.

- Se i valori osservati si avvicinano a 1→ allora essi sono vicini ad una retta crescente che indica una relazione diretta
- Se i valori osservati si avvicinano a -1→ allora essi sono vicini ad una retta decrescente che rappresenta una relazione inversa
- Quando r=0 non c'è alcuna relazione lineare
- R=1 → Xe Y hanno massima (perfetta) correlazione positiva= tutte le osservazioni stanno su una retta crescente
- R=-1 →Xe Y hanno massima (perfetta) correlazione negativa= tutte le osservazioni stanno su una retta decrescente

<u>Regola dei minimi quadrati</u>

La retta che viene ottenuta con questo metodo è detta retta di regressione ed è data da→ Y=b0+b1X

B1 è la pendenza della retta→ $b_1 = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$

B0 è l'ordinata della retta→ $b_0 = \bar{y} - b_1\bar{x}$

<u>Forme della distribuzione</u>

Descrive come i dati sono distribuiti:

- Obliqua a sinistra→ media<mediana
- Simmetrica→ media =mediana
- Obliqua a destra → media> mediana

Trasformazione dei dati

- Moltiplicazione per una costante c
  - C*media
  - $c^2 * varianza^2$
  - |c|*deviazione standard
- Addizione di una costante c
  - Media + c
  - Varianza
  - Devianza standard
- Standardizzazione→ speciale trasformazione dei dati e si trova facendo Z= $\frac{X - media}{deviazione\ standard}$

In seguito alla standardizzazione la media è 0 e la deviazione standard è 1. Questo metodo è utile per confrontare distribuzioni con diversa media e deviazione standard.