STATISTICA SILETTI

Variabile
- Quantitativa (numerical): scala di misura definita da valori numerici (es età, n figli)
- Qualitativa (categorical): scala di misura definita da insieme di categorie (es genere)
- Nominale (nominal): non esiste ordinamento naturale delle modalità (es genere)
- Ordinale (ordinal): modalità hanno ordinamento naturale (es classe sociale: upper, middle, lower

Rappresentazioni grafiche
- Grafico a torta: per variabile qualitativa nominale e quando si vuole sottolineare la quota delle osservazioni nelle diverse categorie
- Grafico a barre: per variabile qualitativa ordinale e quando si ha un numero grande
- Grafico a bastoncini: per variabile quantitativa sintetizzata con valori nella distribuzione di freq.
- Istogramma: per variabile quantitativa, se i valori sono raggruppati in classi di valori nella distribuzione di frequenza

Moda: modalità o valore a cui è associata la frequenza massima individuabile per tutti i tipi di fenomeni
- Se la distribuzione è a singoli valori la moda è il valore a cui è associata la frequenza massima
- Se i valori sono raggruppati in classi si parla di classe modale, ovvero la classe a cui è associata la densità di frequenza massima, e la moda, convenzionalmente, è il valore centrale di tale classe

Mediana: è il valore associato all'unità statistica che sta nel mezzo della distribuzione ordinata del fenomeno
per fenomeno qualitativo ordinale, N in ordine non decrescente
modalità associata all'unità statistica che occupa la posizione centrale
Con n dispari —> (n+1)/2
Con n pari —> 2 mediane: $P_1 = N/2$; $P_2 = N/2 + 1$ (mediana non definita)

Se i valori sono raggruppati in classi si parla di classe mediana, ovvero la classe a cui è associata l'unità statistica nella posizione P = (N+1)/2 della distribuzione ordinata del fenomeno, la mediana viene quindi individuata facendo un ipotesi di equidistribuzione delle osservazioni all'interno della classe mediana

Media
Data una variabile quantitativa X, che assume su N unità statistiche i valori non decrescenti x_1, x_2 … x_n, sono denominate medie quelle particolari funzioni che soddisfano le seguenti proprietà:

- d'**internalità**, la media non deve essere inferiore al più piccolo dei valori osservati né superiore al più grande, ovvero deve essere interna al campo di osservazione;
- **moltiplicativa**, la media segue le variazioni dell'unità di misura, ad esempio, se la «statura» rilevata su N individui viene espressa in metri, anche la media è espressa in metri, ma se si modificano gli N valori in centimetri, anche la media subisce la stessa modifica
- di **monotonicità**, ovvero se almeno uno dei valori aumenta, o diminuisce, restando invariati tutti gli altri, la media aumenta, o diminuisce

Proprietà della media aritmetica:
- La somma dei quadrati degli scarti dei valori dalla loro media è minima
- La somma dei quadrati degli scarti dei valori dalla loro media è minima
- Se i valori x_i vengono trasformati linearmente allora la media subisce la stessa trasformazione (la media della trasformata è uguale alla trasformata della media)

Teoremi media
- La media della somma o della differenza di due o più variabili è uguale alla somma od alla differenza delle medie delle singole variabili
- La media di un miscuglio di k gruppi, o sottopopolazioni, per cui sono note la media e le numerosità di gruppo, è uguale alla media delle medie dei singoli gruppi ponderata con le numerosità di gruppo:

$$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{k} \bar{x}_i n_i = \frac{20 \cdot 3 + 40 \cdot 1}{4} = \frac{100}{4} = 25$$

Simmetria: presenza di alcune ripetizioni nella forma geometrica di un oggetto
—> distribuzione ordinata delle parti di un oggetto tale che si possa individuare un elemento geometrico (un punto, una linea, una superficie) rispetto al quale a ogni punto dell'oggetto posto da una parte di esso corrisponda, a uguale distanza, un punto dall'altra parte

Per le distribuzioni simmetriche la mediana e la media sono uguali
(l'uguaglianza tra media e mediana è una condizione «necessaria», ma non «sufficiente» alla simmetria)
Per le distribuzioni asimmetriche, la media si trova verso la coda più lunga rispetto alla mediana

MEDIANA vs MEDIA
- Per la media —> dati quantitativi
- Mediana vale anche per le scale ordinarie
- Media utilizza valori numerici delle osservazioni, non solo il loro ordinamento
- Media più appropriata della media quando distribuzione molto asimmetrica

Una misura di posizione da sola non è adeguata per descrivere i dati descrive un valore tipico, ma non la dispersione dei dati intorno a quel valore tipico

Range - Campo di variazione: differenza tra valore più grande e il più piccolo osservato; non è sensibile ad altre caratteristiche di variabilità dei dati

Variabilità: attitudine di un fenomeno quantitativo a manifestarsi con modalità/valori differenti

La misura di variabilità più utilizzata si basa sulle distanze (devianze-scarti) dei dati dalla loro media

Varianza: media degli scarti dalla media al quadrato, misura di variabilità più utilizzata $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$

Se i dati provengono da un campione $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Varianza con fr. assolut: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} x_i^2 \cdot n_i - \mu^2$; **Varianza con fr. relativa** $\sigma^2 = \sum_{i=1}^{k}(x_i - \mu)^2 \cdot f_i$

Proprietà Varianza:
- La varianza è invariante per traslazione
- Se si effettua una trasformazione lineare sulla variabile "perdiamo" l'intercetta, ma il coefficiente angolare "subisce" la stessa trasformazione della varianza

Teoremi varianza:
- La varianza della somma di due variabili risulta uguale alla somma delle varianze delle singole variabili solo se queste sono tra loro indipendenti
- La varianza di un miscuglio di k gruppi, o sottopopolazioni, per cui sono note le medie, le varianze e le numerosità di gruppo, è uguale a $\sigma^2 = \sigma_W^2 + \sigma_B^2$

Il primo addendo è uguale alla media delle varianze, e viene detto varianza within ovvero varianza di gruppi
Il secondo addendo è uguale alla varianza delle medie, e viene detto varianza between ovvero varianza tra i gruppi

Varianza within $\sigma_w^2 = \frac{1}{N}\sum_{i=1}^{k}\sigma_i^2 n_i$ varianza between $\sigma_B^2 = \frac{1}{N}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i$

Coefficiente di variazione (proposto da Pearson) $CV = \frac{\sigma}{|\bar{x}|}$
- CV < 1 : le distribuzioni sono considerate a bassa variabilità
- CV > 1 : le distribuzioni sono considerate ad alta variabilità

- A volte viene espresso anche in termini percentuali
- Confrontando serie, quella per la quale il CV è maggiore è la più variabile (meno omogenea)
- Quando il valore medio è vicino allo zero, il CV si avvicinerà all'infinito, è quindi sensibile a piccole variazioni della media (meaningful - scala rapporto)
- Nella scienza attuariale, il CV è noto come «rischio unitizzato» – «unitized risk»
- I CV sono spesso utilizzati come controlli di qualità per analisi di laboratorio quantitative
- Il CV soddisfa i requisiti come misura della disuguaglianza economica
- Gli archeologi utilizzano spesso i CV per confrontare il grado di standardizzazione dei manufatti antichi

Mutabilità: variabilità dei fenomeni qualitativi, attitudine di un fenomeno qualitativo a manifestarsi con modalità/qualità differenti.
Gli indici che misurano la mutabilità sono chiamati indici di eterogeneità ai quali vengono contrapposti gli indici di omogeneità

$$G = 1 - \sum_{i=1}^{k} f_i^2$$

Indice di Eterogeneità di Gini
- Assume valore minimo, ovvero di minima eterogeneità o massima omogeneità per G=0
- Assume valore massimo, ovvero di massima eterogeneità o minima omogeneità per G=(k-1)/k
- Il massimo dipende dal numero k delle modalità con cui si manifesta il fenomeno

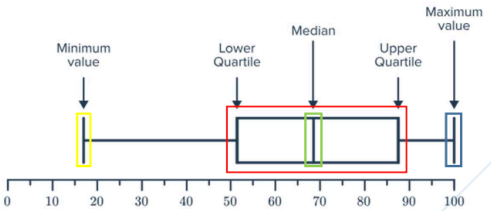
Per ottenere un indice normalizzato, ovvero un indice che varia tra 0, minima eterogeneità, ed 1, massima eterogeneità, è necessario dividere l'indice rispetto il suo massimo:

$$G_{norm} = \frac{G}{G_{max}} = \frac{1 - \sum_{i=1}^{k} f_i^2}{(k-1)/k} = \left(1 - \sum_{i=1}^{k} f_i^2\right)\frac{k}{k-1} = G\frac{k}{k-1}$$

Percentile: misura di posizione
p^{esimo} percentile: valore che divide un assegnato insieme di dati, cioè di valori, supposti ordinati in senso non decrescente, in modo che il numero dei valori inferiori a p costituisca p% delle osservazioni e il numero dei valori superiori a p costituisca (1-p)% delle osservazioni

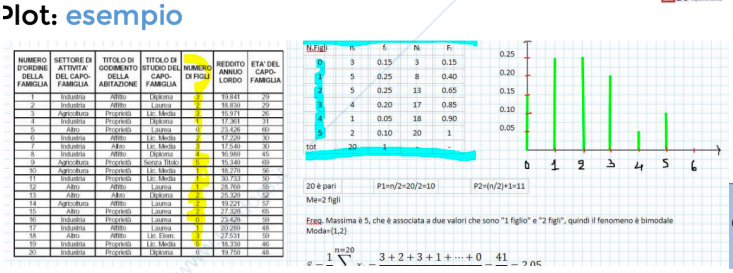
Intervallo/differenza interquartile: differenza tra i quartili superiore e inferiore (DI)
rispetto al range o alla deviazione standard non è sensibile ai valori anomali
Aumenta all'aumentare della variabilità ed è utile per confrontare la variabilità di diversi gruppi

Box plot: sintetizza graficamente il centro (mediana) e la dispersione (DI e range) dei dati quantitativi
composta da cinque indici: minimo, quartile inferiore, mediana, quartile superiore, massimo
permette un'analisi immediata della forma della distribuzione
simmetria: raccomandato se la distribuzione non è simmetrica



identificano anche i valori anomali /outlier: i baffi si estendono al minimo e al massimo solo se tali valori non sono valori anomali
Outlier: osservazione che ha valore maggiore di Q3 + 1.5 · (DI) o minore di Q1 - 1.5 · (DI)
Valore estremo: osservazione che ha valore maggiore di Q3 + 3 · (DI) o minore di Q1 - 3 · (DI)

Plot: esempio

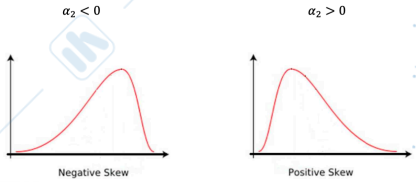


Non ci sono valori anomali:
$Q1 - 1.5(DI) = 1 - 3 = -2$ non è una valore osservabile, comunque $-2 < \min$
$Q3 + 1.5(DI) = 3 + 3 = 6$ non è una valore osservato, $6 > \max$

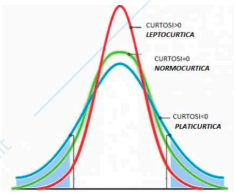
Standardizzazione
- Valori negativi: sotto la media
- Valori positivi: sopra la media

Distribuzione simmetrica se:
- Media aritmetica = mediana
- Primo e terzo quartile stessa distanza dalla mediana: Q2 - Q1 = Q3 - Q2
- Somma degli scarti dalla media elevati alla terza è nulla: sommatoria(xi-media)^3 = 0

L'indice di asimmetria $\alpha_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\sigma^3} = \begin{cases} > 0 & \text{se } X \text{ presenta asimmetria positiva} \\ = 0 & \text{se } X \text{ è simmetrica} \\ < 0 & \text{se } X \text{ presenta asimmetria negativa} \end{cases}$



L'indice di curtosi $= \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\sigma^4} = \begin{cases} > 3 & X \text{ leptocurtica o ipernormale} \\ = 3 & X \text{ è normale} \\ < 3 & X \text{ platicurtica o iponormale} \end{cases}$



Disuguaglianza di Chebyshev:
Tra le N osservazioni (x1, .., xn) si cui si conosce solo la media μ e la deviazione standard σ, la frequenza relativa di quelle che cadono nell'intervallo μ - kσ e μ + kσ è non inferiore a 1 - 1/k^2
K > 1; rappresenta il numero di σ dalla media

Intervallo (μ - kσ;μ + kσ) Formula = 1 - (1/k^2)

La **distribuzione di frequenza bivariata**: sintesi tabellare di due variabili rilevate congiuntamente
Tabella in cui:
- In riga sono riportate le differenti modalità/valori con cui si manifesta un fenomeno i
- In colonna quelle dell'altro fenomeno

Se non esiste alcuna relazione si dirà che X ed Y sono statisticamente indipendenti
Il metodo per stabilire se sono statisticamente indipendenti consiste nel confrontare le frequenze condizionate

Se al variare delle modalità del fenomeno condizionante le distribuzioni condizionate non variano, allora i due fenomeni sono statisticamente indipendenti
Se due fenomeni non sono statisticamente indipendenti allora esiste una relazione e si dirà che i fenomeni sono connessi

Misurare grado di connessione: considerare la differenza fra le frequenze congiunte osservate e le frequenze che si avrebbero in condizione di indipendenza statistica, ovvero le frequenze teoriche di indipendenza

Indice utilizzato per la connessione: **Chi quadro di Pearson** $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-n_{ij}^*)^2}{n_{ij}^*} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{{c_{ij}}^2}{n_{ij}^*}$

Formula alternativa (non richiede calcolo fr. teoriche) $\chi^2 = N\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{{n_{ij}}^2}{n_{i.}n_{.j}} - 1\right)$

Normalizzazione $\chi^2_{norm} = \frac{\chi^2}{N[min(r-1, c-1)]}$ —> indice di variazione da 0 ad 1

Indici di connessione: valutano l'intensità del legame in modo generico andando cioè a valutare l'entità dell'allontanamento dalla situazione di indipendenza

Se almeno uno dei due fenomeni è quantitativo ad esempio Y, ci si può chiedere se Y dipende in media da X, ovvero se al cambiare delle modalità di X cambiano le medie di Y

Indipendenza in media: dipende in media da X se la relazione di connessione tra le due variabili si riflette sulle medie condizionate di Y che risultano diverse tra loro al variare di X
C'è indipendenza se tutte le medie condizionate sono tra loro uguali e quindi uguali alla media marginale

Se entrambi i fenomeni sono quantitativi è necessario considerare anche la dipendenza in media di X da Y: si dice, allora, che X dipende in media da Y se la relazione di connessione tra le due variabili si riflette sulle medie condizionate di X che risultano diverse tra loro al variare di Y

L'indipendenza in media di X da Y non implica l'indipendenza in media di Y da X

Età quadro: indice di dipendenza per misurare la dipendenza in media $\sigma_Y^2 = \sigma_B^2 + \sigma_W^2$
Valori compresi tra 0 ed 1
0 = varianza FRA (between) i gruppi è nulla cioè quando Y è indipendente in media da X (e la varianza NEI gruppi coincide con la varianza marginale di Y)
1 = varianza FRA i gruppi coincide con la varianza marginale di Y cioè quando Y è perfettamente dipendente da X (e la varianza NEI (within) gruppi è nulla)

$$\sigma_Y^2 = \frac{1}{N}\sum_{j=1}^{c}(y_j - \bar{y})^2 n_{.j} = \frac{1}{N}\sum_{i=1}^{r}(\bar{y}|x_i - \bar{y})^2 n_{i.} + \frac{1}{N}\sum_{i=1}^{r}\sigma_Y^2|x_i \cdot n_{i.}$$

$$\eta_{Y|X}^2 = \frac{\sigma_B^2}{\sigma_Y^2} = \frac{\frac{1}{N}\sum_{i=1}^{r}(\bar{y}|x_i - \bar{y})^2 n_{i.}}{\frac{1}{N}\sum_{j=1}^{c}(y_j - \bar{y})^2 n_{.j}}$$

Covarianza: misura variabilità congiunta di due variabili X e Y $Cov(X,Y) = \sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$

Correlazione: indice di correlazione lineare di Bravais e Pearson è un indice normalizzato di correlazione $\rho = \frac{Cov(X,Y)}{DS(X)DS(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$