CLEAM

A.A. 2024/2025

BLAB

DISPENSA

STATISTICA

-PRIMO PARZIALE-

A CURA DI
MARCO FORMISANO

TEACHING DIVISION

STATISTICA

processi decisionali in condizioni di incertezza ⇒ **metodo statistico**

insieme di procedure necessarie e funzionali all'analisi di dati, finalizzata all'estrazione di informazioni di immediato valore pratico, ad eventuale supporto di svariati processi decisionali

valutazione di scenari, definizione di strategie aziendali o manovre politiche
- individuare le abitudini di acquisto dei clienti di un'azienda
- stabilire la strategia di marketing più efficace per un target
- prevedere l'ammontare degli ordini per il prossimo anno

RACCOLTA E PREPARAZIONE DEI DATI

FONTI PRIMARIE
- osservazioni
- indagini
- esperimenti

FONTI SECONDARIE
- elaborazioni di fonti primarie (formato cartaceo o elettronico)

i dati possono riferirsi a una popolazione o a un campione

popolazione (universo) ⇒ insieme di tutte le unità di interesse in uno studio

campione (sottoinsieme) ⇒ frazione di elementi scelti all'interno della popolazione

i campioni devono essere estratti in modo casuale affinché siano rappresentativi di tutte le unità nella popolazione e non riflettano caratteristiche di gruppi specifici di soggetti

campionamento casuale semplice
- → ogni unità è selezionata in maniera casuale all'interno della popolazione
- → ogni unità ha la stessa probabilità di essere selezionata
- → ogni possibile campione di unità ha la stessa probabilità di essere scelto

---

- **caso / unità statistica**: entità (unità della popolazione) su cui viene rilevata l'informazione
- **variabile**: una caratteristica (dei casi) di interesse
- **dati / misurazioni**: osservazioni sulla variabile di interesse misurate sui casi presi in considerazione
- **modalità**: manifestazioni della variabile, valori assunti dalla variabile

i dati sono tipicamente organizzati in **dataset** con i casi sulle righe e le variabili sulle colonne

CLASSIFICAZIONE DELLE VARIABILI

qualitative (categoriali) ⇒ le modalità delle variabili qualitative sono etichette che indicano attributi, ovvero l'appartenenza a gruppi o categorie con specifiche caratteristiche

es. sesso, genere, regione di nascita, settore d'attività

- → nominali: le modalità delle variabili non possono essere ordinate in nessun modo
- → ordinali: le modalità delle variabili si possono ordinare, senza quantificarne le differenze

quantitative (numeriche) ⇒ le modalità delle variabili quantitative sono valori numerici

es. età, altezza, numero di figli, ammontare speso

- → discrete: i dati sono generati da un processo di conteggio (numeri interi)
- → continue: i dati sono generati da un processo di misurazione (numeri reali)

PARAMETRI E STATISTICHE

in generale si è interessati a misurare alcune caratteristiche di un insieme di dati; bisogna distinguere fra:

- **parametri** → misura che descrive o sintetizza una caratteristica della popolazione
- **statistica** → misura che descrive o sintetizza una caratteristica di un campione

STATISTICA DESCRITTIVA

- metodi grafici e numerici per la sintesi e l'elaborazione dei dati
- è applicabile a dati riguardanti l'intera popolazione o un campione
- include tecniche per la predisposizione, la sintesi e la presentazione dei dati

es. sintesi: media, varianza, correlazione
es. presentazione: tabelle e grafici

STATISTICA INFERENZIALE

- metodi tramite cui è possibile fare inferenza e previsioni su caratteristiche della popolazione (parametri) partendo da informazioni estratte da dei dati campionari (statistiche)
- l'affidabilità dell'analisi statistica è legata al rischio in cui si incorre utilizzando informazioni campionarie per fare inferenza sulla popolazione
  - ↳ bisogna considerare il meccanismo casuale e l'aleatorietà insita nell'estrazione del campione

STATISTICA UNIVARIATA

- distribuzione di frequenza
  - tabelle
  - grafici
- misure di sintesi
  - misure di posizione centrale
    - media
    - mediana
    - moda
  - misure di posizione non centrale
    - quantili
  - misure di variabilità
    - range
    - range interquartile
    - varianza
    - scarto quadratico medio
    - coefficiente di variazione
- forma
  - box plot

STATISTICA DESCRITTIVA

DISTRIBUZIONE DI FREQUENZA

TABELLE E GRAFICI

i dati relativi ad una variabile possono essere rilevati sugli N casi di una popolazione o su n casi di un campione → dati rilevati su un campione di n unità : $x_1, x_2, \ldots, x_n$

per analizzare e interpretare correttamente i dati grezzi è necessario organizzarli efficacemente

⇒ si ricorre a schemi come tabelle e grafici costruiti tenendo conto di { tipologia dei dati (qualitativi o quantitativi); numero K di valori/modalità distinti $x^*_1, x^*_2, \ldots, x^*_K$

i dati vengono organizzati in una distribuzione di frequenza, ossia una tabella che riporta:

modalità
valori distinti per ogni variabile

frequenza assoluta
numero di casi di ogni modalità

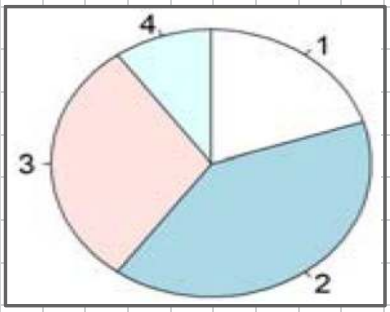
frequenza relativa
proporzione di casi di ogni modalità sul numero di casi totali

frequenza relativa = frequenza assoluta / numero casi

VARIABILI QUALITATIVE

per rappresentare graficamente la distribuzione di frequenza relativa a una variabile qualitativa si usano

DIAGRAMMA A TORTA



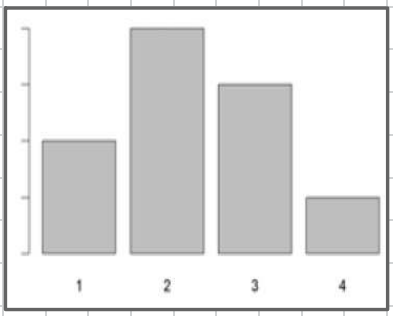
un cerchio diviso in spicchi (modalità) le cui aree sono proporzionali alle frequenze delle modalità osservate

il grafico fornisce informazioni sull'importanza relativa di ogni modalità

DIAGRAMMI A TORTA → SOLO VARIABILI QUALITATIVE NOMINALI

```
distr.plot.x(x,freq="proportions",plot.type="pie",data)
```

DIAGRAMMA A BARRE



un insieme di barre (modalità) pari ampiezza le cui altezze sono proporzionali alle frequenze delle modalità osservate

il grafico consente di apprezzare l'ordine delle modalità osservate

DIAGRAMMI A BARRE -> VARIABILI QUALITATIVE ORDINALI E NOMINALI

```
distr.plot.x(x,freq="perc",plot.type="bars",data)
```

R STUDIO

- la distribuzione di frequenza di una variabile è facilmente determinabile tramite la **funzione table(x)**

argomenti – nome del vettore contenente i dati (colonna del dataframe)

per ottenere le frequenze relative è sufficiente dividere tale valore per il numero di righe (casi) del dataframe

→ la funzione stampa la distribuzione delle frequenze assolute di ogni modalità

- per costruire una tabella che riporti i diversi tipi di frequenze si ricorre alla **funzione distr. table. x(...)**

```
distr.table.x(x, freq= c("counts","prop"), total=TRUE, data)
```

argomenti
- x è un vettore o fattore (nome di una delle colonne del dataframe)
- freq indica quali frequenze si riportano nella tabella [assolute(counts), relative (prop) e/o percentuali (perc)]
- total è un valore logico che specifica bisogna riportare i totali (TRUE) o meno (FALSE)
- data è il nome del dataframe di riferimento

- per rappresentare graficamente una tabella esistono in R diverse funzioni come **funzione distr. plot. x(...)**

```
distr.plot.x(x, freq="counts", plot.type, bw=FALSE, data)
```

argomenti
- x è un vettore o fattore (nome di una delle colonne del dataframe)
- freq indica quale frequenza si riporta nel grafico [assoluta (counts), relativa (prop) o percentuale (perc)]
- plot type è il tipo di grafico da produrre (varia a seconda del grafico: pie, bars, spike)
- bw specifica se il grafico va prodotto a colori (FALSE) o in scale di grigio (TRUE)
- data è il nome del dataframe di riferimento

in caso di dati di variabili ordinali è raccomandabile utilizzare un fattore per ordinarne i livelli in maniera opportuna

```
nome_variabile.F <- factor(nome_variabile, levels = c("..."))
```

VARIABILI QUANTITATIVE

per le variabili quantitative (numeriche), bisogna considerare che:

le variabili discrete possono presentare un numero di modalità sia ridotto che piuttosto elevato

le variabili continue assumono in genere un valore diverso per ogni osservazione (numero di modalità osservate = numero di casi)

DIAGRAMMA AD ASTE

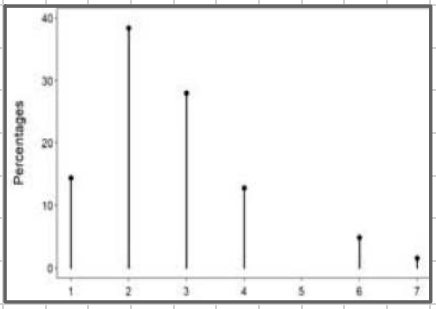


grafico che associa ad ogni modalità osservata un'asta la cui altezza è la frequenza assoluta o relativa

tiene conto sia delle modalità osservate che delle loro distanze sull'asse orizzontale (a differenza del diagramma a barre*)

DIAGRAMMI AD ASTA → SOLO VARIABILI QUANTITATIVE DISCRETE

```
distr.plot.x(x,freq="perc",plot.type="spike",data)
```

* nel diagramma a barre la distanza fra valori sull'asse orizzontale è sempre uguale (adatto quando le modalità sulle ascisse sono qualitative, non adatto per variabili quantitative con modalità numeriche)

CLASSI DI INTERVALLO

se il numero di modalità è elevato l'efficacia descrittiva della tabella di frequenza è minima (per via dell'alto numero di righe-modalità e delle basse frequenze che le caratterizzano) ed è quindi necessario semplificarla

⇒ conviene suddividere i dati in classi di intervallo con specifiche caratteristiche ⇐

- intervalli esaustivi e mutualmente esclusivi [intervalli adiacenti e non sovrapposti che includono tutte le modalità osservate]
- limiti degli intervalli definiti chiaramente [per convenzione l'estremo inferiore è incluso, il superiore no tranne che per l'ultimo]

la distribuzione delle frequenze in questo caso è una tabella che associa alle varie classi di intervallo:
- frequenze assolute: numero di casi che presentano modalità appartenenti ad ogni intervallo
- frequenze relative: proporzione di casi che presentano modalità appartenenti ad ogni intervallo

ampiezza [w] → dimensione di ogni classe di intervallo } le classi possono avere ampiezze uguali o diverse

la distribuzione delle frequenze per dati in classi di intervallo si rappresenta graficamente con l'ISTOGRAMMA

ISTOGRAMMA

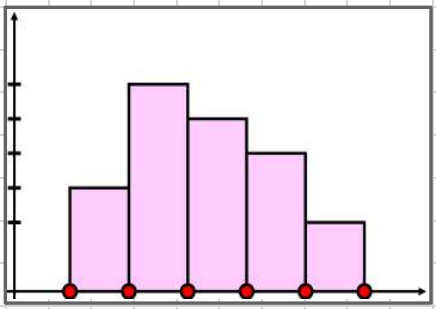


diagramma a rettangoli accostati che associa ad ogni classe di intervallo un rettangolo che ne misura la densità di frequenza

- base = ampiezza dell'intervallo (k-esimo intervallo: w_k)
- area = frequenza, tipicamente relativa (k-esimo intervallo: p_k)
- altezza = densità di frequenza = frequenza/ampiezza (k-esimo intervallo c_k)

ISTOGRAMMA → VARIABILI QUANTITATIVE DISCRETE E CONTINUE

```
distr.plot.x(x, freq= "counts", plot.type="histogram", breaks, data)
```

CLASSI DI AMPIEZZA UGUALE

per suddividere i dati in classi con stessa ampiezza (w) essa si ottiene con la formula: $w = \frac{(\text{Max-Min})}{\text{Nr di classi}}$

* max e min: massimo e minimo dei valori osservati (l'ampiezza si può arrotondare per semplificare la classificazione)

se le classi hanno stessa ampiezza ⇒ è possibile utilizzare oltre le densità di frequenza anche le frequenze assolute o relative nell'istogramma (asse delle ordinate) senza modificarne l'aspetto

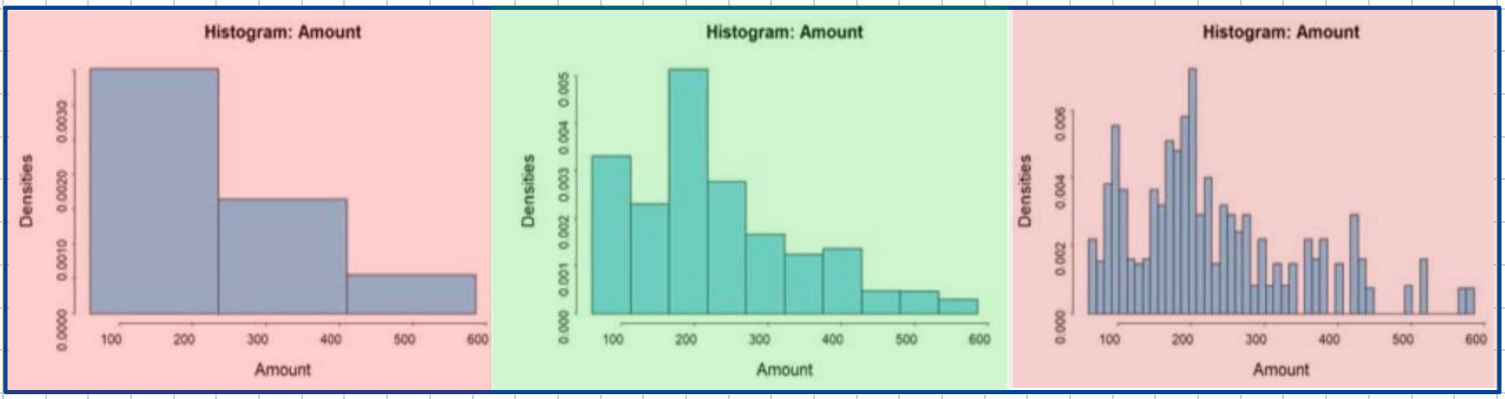
è preferibile usare sempre le densità

NUMERO DI CLASSI

la scelta del numero di intervalli (in genere 5 - 25) dipende dalle caratteristiche della variabile e dal numero di osservazioni pertanto non c'è un criterio ottimale per stabilirlo a priori , tuttavia è bene tenere presente che:

l'istogramma deve

- descrivere le caratteristiche di forma della distribuzione e dispersione dei dati
  - → distr. simmetrica o asimmetrica
  - → presenza di "code"
- rendere evidenti le differenze nei dati semplificandone l'osservazione
  - → né tante classi con bassa frequenza
  - → né poche classi con alta frequenza

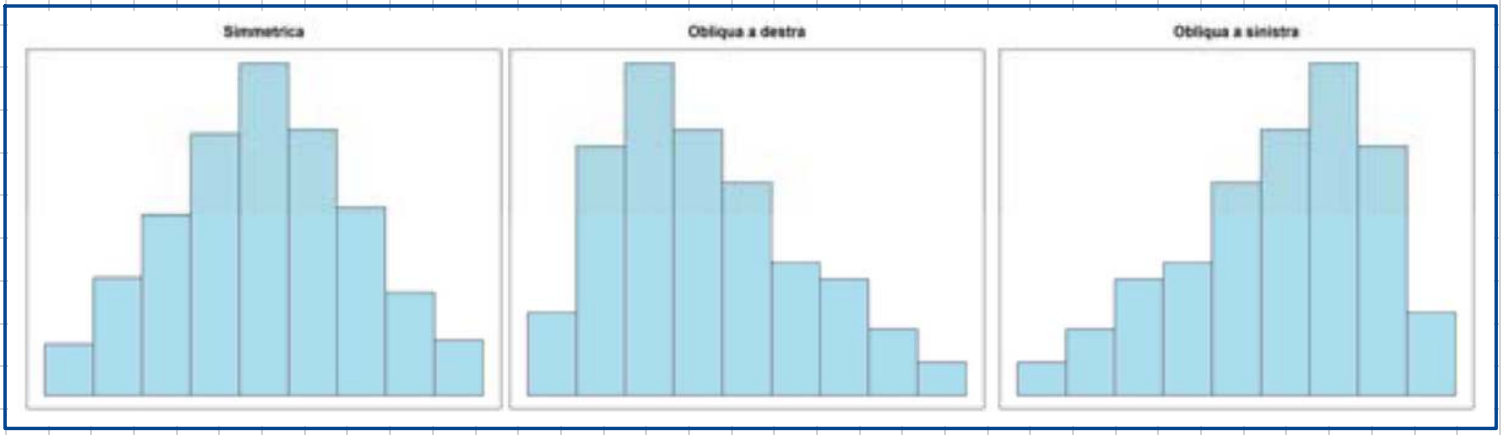


FORMA DELLA DISTRIBUZIONE

caratteristiche rilevanti di una distribuzione sono la sua simmetria e la presenza di eventuali code

distribuzione simmetrica → dati distribuiti in modo regolare intorno al centro dell'istogramma
distribuzione asimmetrica → presenta una coda che si estende lungo una sola direzione

tutte le distribuzioni, simmetriche e oblique, possono presentare code particolarmente lunghe



R STUDIO

- per analizzare una variabile con dati numerici attraverso classi di intervallo si usa l'argomento breaks

```
distr.table.x(x, freq= c("counts","prop","dens"), breaks, data)
```

```
distr.plot.x(x, freq="counts", plot.type="histogram", breaks, data)
```

argomenti

- freq permette di richiedere che la tabella e/o il grafico riporti le densità ("densities")
  *se le classi hanno ampiezze diverse l'istogramma viene rappresentato usando le densità
- breaks consente di specificare la classificazione desiderata secondo due modalità:
  - se breaks è un singolo valore numerico esso indica il n° di intervalli di pari ampiezza da creare → breaks = x
  - se breaks è un vettore di valori crescenti questi definiscono i limiti degli intervalli → breaks = c(x,y,z,...)

per creare classi di pari ampiezza su molti dati → breaks = seq (min, max, by = ampiezza)

CLASSI DI AMPIEZZA DIVERSA

spesso e volentieri si osservano molti dati concentrati in un intervallo di valori relativamente piccolo e i rimanenti dispersi su un intervallo molto ampio (con la conseguente formazione di lunghe code)
per semplificare la distribuzione di frequenza e la sua rappresentazione grafica è possibile costituire classi di intervallo di ampiezza diversa maggior dettaglio sugli intervalli in cui cade la maggior parte dei dati

se le classi hanno ampiezza diversa ⇒ poiché le basi cambiano, l'istogramma dovrà necessariamente essere costruito con le densità di frequenza (le altezze non possono essere frequenze assolute o relative)

DATI IN CLASSI DI INTERVALLO

può capitare che i dati vengono rilevati direttamente in classi di intervallo (dati sensibili o approssimativi)

→ R non può riconoscere la natura quantitativa dei dati, in quanto gli intervalli vengono considerati come categorie, pertanto è necessario ricorrere ad un fattore che permetta di ordinare opportunamente i dati

R STUDIO

- per analizzare variabili con dati numerici già rilevati in classi di intervallo si usa l'argomento interval

```
distr.table.x(x, freq= c("counts","prop","dens"),interval=TRUE, data)
```

```
distr.plot.x(x, freq= "counts", plot.type="histogram", interval=TRUE, data)
```

argomenti

- interval segnala che x è una variabile rilevata in classi

le funzioni analizzano le codifiche degli intervalli e tentano di individuarne gli estremi cosicché, se consistenti, le classi possano essere ordinate e la variabile tabulata / rappresentata graficamente

estremi inconsistenti: la tabella viene prodotta con un messaggio di warning, l'istogramma no

in distr. table. x si utilizzano i parametri f.digits, p.digits, d.digits per specificare il numero di decimali per frequenze relative, percentuali e densità

```
distr.table.x(x, freq=c("counts","prop","perc",
"dens"),interval=TRUE,f.digits=3,d.digits=3,data)
```

FREQUENZE CUMULATE

la distribuzione delle frequenze cumulate si ottiene associando ad ogni modalità (o classe) la somma della sua frequenza relativa e di quelle delle modalità (o classi) che la precedono

$F_k = p_1 + p_2 + \ldots + p_k$ → in genere si fa riferimento alle frequenze relative anche se è possibile cumulare anche frequenze assolute o percentuali

per una variabile numerica è possibile valutare la frequenza cumulata per qualsiasi valore sull'asse reale

funzione cumulativa delle frequenze o funzione di ripartizione

associa ad ogni numero (reale) x la frequenza relativa con cui si osservano valori ad esso inferiori o uguali

$F(x) = Freq(X \leq x)$

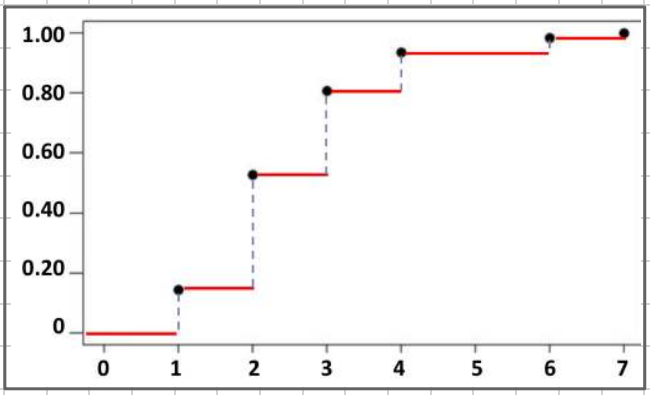
X → variabile, insieme di valori
x → modalità, valore specifico

DIAGRAMMA A SCALINI

i valori fra due modalità non sono osservati pertanto non contribuiscono alla cumulata

- valori minori del minimo → $F_k = 0$
- valori maggiori del massimo → $F_k = 1$

DIAGRAMMI A SCALINI → VARIABILI QUALITATIVE O QUANTITATIVE CON POCHE MODALITÀ (DISCRETE)



è possibile determinare le frequenze cumulate anche per una variabile qualitativa ordinale (tuttavia quando non riconosciuta da R è necessario modificarne i livelli tramite un fattore)

---

approssimando il grafico a scalini, a partire dai dati grezzi, e quindi la funzione cumulativa si ottiene l'ogiva

se non sono disponibili i dati grezzi, la funzione si approssima in ipotesi di ripartizione uniforme dei valori nelle varie classi

dato un valore x compreso nel k-esimo intervallo [a,b)

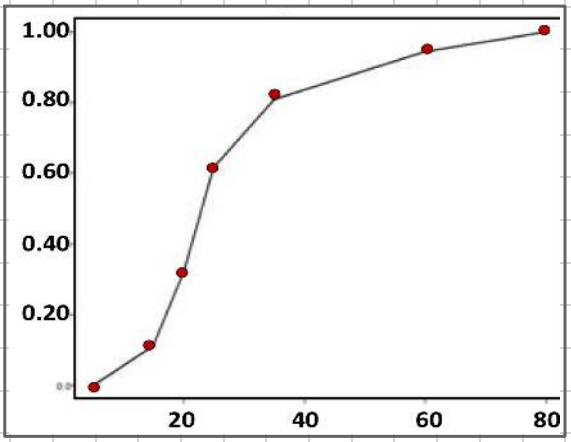
$F(x) = Freq(X \leq x) = F_k + (x - a) * c_k$

OGIVA

l'ogiva o curva delle frequenze cumulate è un grafico basato sulla classificazione in intervalli (spezzata che connette le frequenze cumulate in corrispondenza degli estremi superiori)

l'ogiva è costruita assumendo che l'incremento della frequenza cumulata in ogni intervallo sia costante

OGIVA → VARIABILI QUANTITATIVE (IN GENERE CONTINUE)



si utilizza l'ogiva:
- quando non si dispone dei dati grezzi, ma la variabile è già rilevata in classi
- per semplificare lo studio di una variabile con un numero elevato di modalità

la funzione distr.plot.x() produce un diagramma a scalini solo se i dati non sono classificati in intervalli → se si specifica che i dati sono rilevati in classi (interval = TRUE) la funzione crea il grafico dell'ogiva

R STUDIO

- per calcolare e rappresentare le frequenze cumulate va aggiunta la voce "cumulative" nell'argomento freq

```
distr.table.x(x, freq= c("counts","prop","cum"), total=TRUE, data)
```

```
distr.plot.x(x, freq= "prop", plot.type="cumulative", data)
```

- per rappresentare graficamente l'ogiva si usa la stessa scrittura delle frequenze cumulate specificando:

→ che la variabile si presenta già rilevata in classi [interval = TRUE]
→ che si vuole classificare la variabile in classi di intervallo [breaks = c(...)]

```
distr.plot.x(x, freq="prop", plot.type="cumulative",interval=TRUE, data)
```

```
distr.plot.x(x, freq="prop", breaks=c(a,b,c...), plot.type="cum", data
```

VARIABILI QUALITATIVE / QUANTITATIVE (poche modalità)

TABELLE

- frequenze assolute
- frequenze relative e percentuali
- frequenze cumulate
  (qualitative ordinali o quantitative discrete)

GRAFICI

- diagramma a torta [pie]
- diagramma a barre [bars]
- diagramma ad aste [spike]
- diagramma a scalini [cum]
  (qualitative ordinali o quantitative discrete)

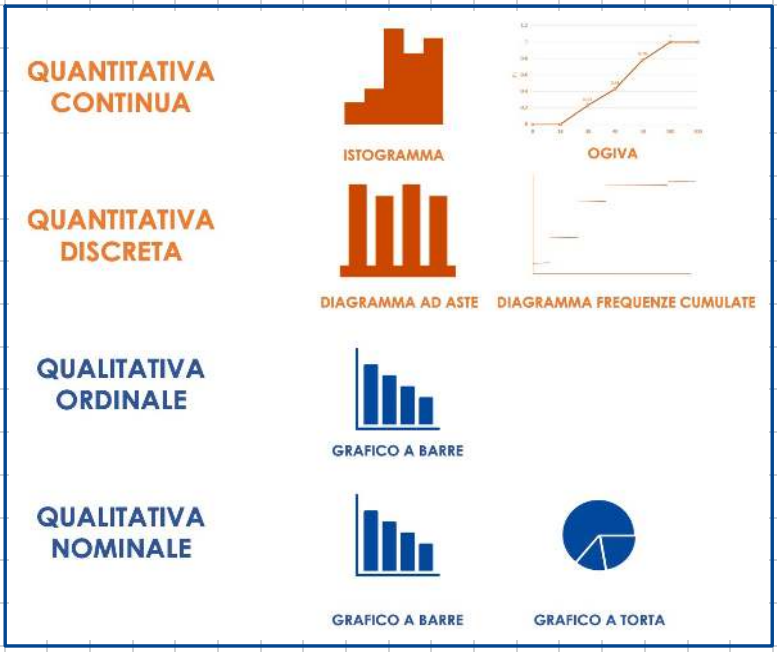
VARIABILI QUANTITATIVE

TABELLE

- frequenze assolute
- frequenze relative e percentuali
- densità di frequenza
- frequenze cumulate

GRAFICI

- istogramma [hist]
- ogiva [cum]



MISURE DI SINTESI

talvolta può essere conveniente sintetizzare le caratteristiche più rilevanti di una serie di modalità

le **misure di sintesi** consentono di comunicare informazioni in modo semplice e intuitivo

*la scelta delle misure di sintesi dipende sempre da tipo di dati e caratteristiche della distribuzione della variabile

MISURE DI POSIZIONE CENTRALE

una **misura di tendenza centrale** è un singolo valore che sintetizza in un dato modo tutti i dati osservati

- ha lo scopo di descrivere il centro dei dati
- può essere di tre differenti tipologie

} a seconda di come si decide di definire la centralità

MODA

la **moda** è la modalità con massima frequenza osservata in un insieme di dati

- si considera "centro" dei dati il valore che descrive il "comportamento più tipico" dei casi della variabile

la moda è ottima per dati qualitativi e quantitativi discreti (con poche modalità), mentre è meno utile per modalità quantitative continue (molte modalità, in quanto hanno tipicamente frequenza unitaria)

la moda può
- non essere unica → due o più modalità con stessa frequenza
- essere debole → due o più modalità con frequenza simile
- non esistere → i dati hanno tutti distribuzione equifrequente

nel caso di variabili continue, discrete con molte modalità o rilevate per intervalli si considera la **classe modale:** la classe con la più alta densità di frequenza (la quale dipende dalla classificazione)

MEDIANA

la **mediana** è il valore posizionato centralmente nella sequenza ordinata dei dati

- si considera "centro" il valore che divide i dati in due blocchi della stessa dimensione (minori e maggiori)

essendo basata esclusivamente sull'ordinamento delle variabili e non sui loro valori, è possibile calcolare la mediana per dati quantitativi e qualitativi ordinali, mentre non si applica ai dati qualitativi nominali

calcolo della mediana
- numero di valori dispari → mediana = valore centrale della sequenza ordinata
- numero di valori pari → mediana = media aritmetica dei due valori centrali

dati grezzi ⇒ la mediana corrisponde al primo valore in corrispondenza del quale la frequenza cumulata raggiunge o supera 0.5

se è 0.5 in corrispondenza di un dato valore, si fa la media fra questo e il successivo

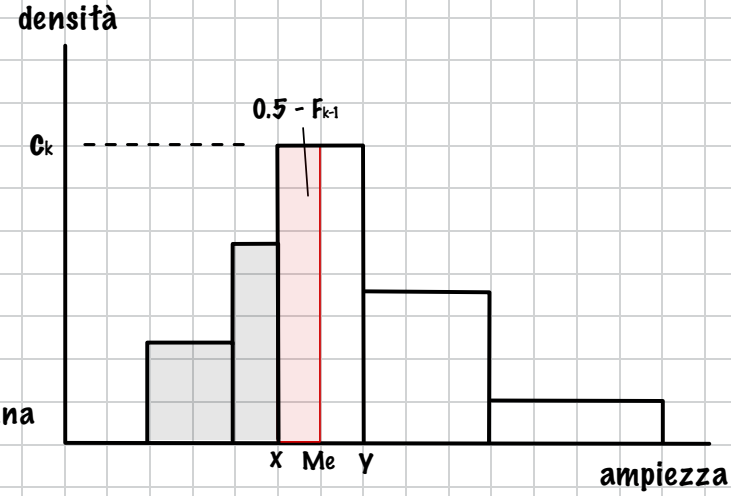
dati rilevati in classi ⇒ la mediana può essere solo approssimata, ipotizzando che la frequenza di ogni intervallo sia uniformemente distribuita

classe mediana: classe in corrispondenza della quale la frequenza cumulata raggiunge o supera 0.5

la mediana è una statistica robusta → relativamente insensibile ai dati estremi / outliers

MEDIANA PER VARIABILI RILEVATE IN CLASSI

$$F_{k-1} + (Me - x_k) * c_k = 0{,}5$$



dove:

- F_{k-1} → è il valore della cumulata della classe precedente
- Me → è il valore della mediana da trovare
- x_k → è il valore dell'estremo inferiore della classe mediana
- c_k → è la densità di frequenza della classe mediana
- $0{,}5$ → è il valore necessario per individuare la classe mediana

---

MEDIA

la media (aritmetica) è la misura data dalla somma dei dati divisa per il numero di casi

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

- $\bar{x}$ → calcolata su un campione (statistica)
- μ → calcolata sulla popolazione (parametro)

la media si calcola solo per le variabili quantitative

- si considera "centro" quel valore tale che la somma delle deviazioni positive e negative dei dati da esso è 0

10	25	30	45	55	60	70	85	90	100

media: 57

(10 - 57)	(25 - 57)	(30 - 57)	(45 - 57)	(55 - 57)	(60 - 57)	(70 - 57)	(85 - 57)	(90 - 57)	(100 - 57)
-47	-32	-27	-12	-2	3	13	28	33	43

- 120 + 120

centro di gravità dei dati osservati: la somma degli scarti fra ogni valore e la media è nullo ⇒

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

la media è una statistica non robusta → molto sensibile ai dati estremi / outliers

variabili discrete

partendo dalle distribuzioni di frequenza dei dati, la media è

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{k=1}^{K} \frac{x_k^* f_k}{n} = \sum_{k=1}^{K} x_k^* \frac{f_k}{n} = \sum_{k=1}^{K} x_k^* p_k$$

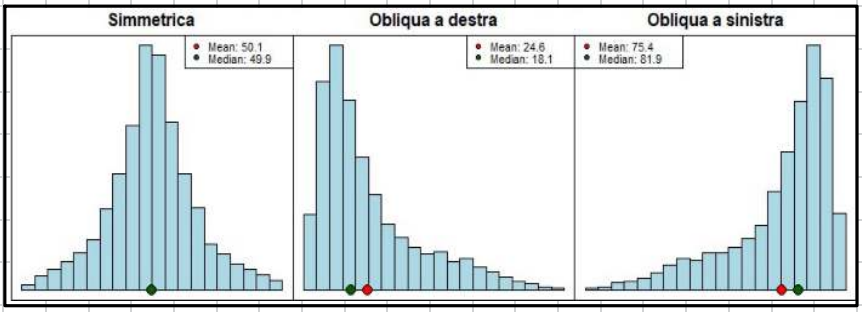
- somma dei prodotti fra modalità osservate e frequenze relative
- somma dei prodotti fra modalità osservate e frequenze assolute divisa per il numero di casi

dati in classi di intervallo → la media può essere solo approssimata a partire dalle classi considerate

⇒ si considerano i punti centrali di ogni classe e si associa ad essi la rispettiva frequenza relativa ⇒ la media approssimata è la somma dei prodotti fra i punti centrali delleclassi e le frequenze relative

MODA E MEDIANA: FORME DELLA DISTRIBUZIONE

la forma della distribuzione, la presenza di valori estremi o eventuali asimmetrie, implica delle differenze fra media e mediana per via della loro diversa robustezza



- se la distribuzione è simmetrica, le due misure sono vicine fra loro
- se la distribuzione è asimmetrica, la media si discosta dalla mediana
  - → media ↑ se la distribuzione è obliqua a destra
  - → media ↓ se la distribuzione è obliqua a sinistra

lo scostamento è tanto più marcato quanto più lunghe sono le code

a partire da media e mediana è possibile dedurre l'asimmetria della distribuzione ⇛ media >> mediana → distribuzione obliqua a destra; media << mediana → distribuzione obliqua a sinistra

in linea generale vale che le misure di posizione centrale sono tanto più rappresentative quanto più i dati sono concentrati intorno al centro

R STUDIO

- per calcolare media e mediana di dati numerici si usano le funzioni mean(x, na.rm) e median(x, na.rm)

argomenti
- x è il nome del vettore contenente i dati (colonna del dataframe: nome_dataframe$nome_variabile)
- na.rm è un valore logico che se TRUE specifica di ignorare eventuali missing values (NA remove)

- per determinare una collezione di misure di tendenza centrale si usa la funzione distr.summary.x(...)

```
distr.summary.x(x, stats, digits=2,f.digits=4, data)
```

argomenti
- stats permette di richiedere le statistiche in output
  - → stats = "central" corrisponde a stats = c("mode","median","mean")
- digits e f.digits consentono di indicare il numero di decimali a cui arrotondare rispettivamente le statistiche richieste e le eventuali frequenze stampate in output (di default 2 e 4)

*distr.summary.x() tratta le variabili rilevate in classi come variabili qualitative, calcolando solo la moda (facendo riferimento alla massima frequenza relativa piuttosto che alla massima densità di frequenza)

moda, mediana e media ⇛

```
distr.summary.x(x, stats="central", data)
```

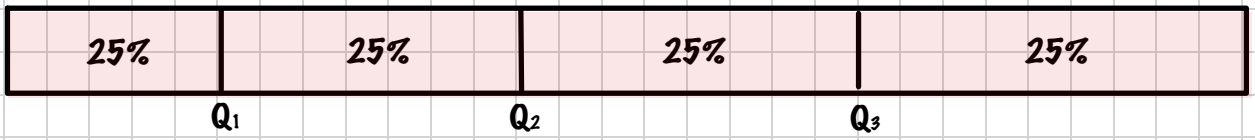
MISURE DI POSIZIONE NON CENTRALE

in caso di distribuzioni fortemente asimmetriche (lunghe code) le misure di tendenza centrale descrivono solo parzialmente le caratteristiche della distribuzione stessa pertanto bisogna ricorrere ad altri mezzi

→ si rivelano utili in questi casi le **misure di posizione non centrale (quantili)**: quartili, decili, percentili...

QUARTILI

i quartili sono valori che suddividono la sequenza ordinata dei dati in 4 blocchi che includono tendenzialmente lo stesso numero di casi



- primo quartile: separa il 25% dei dati più piccoli (minori di Q_1) dal restante 75% (maggiori di Q_1)
- secondo quartile: mediana, separa due blocchi di pari dimensioni (50%)
- terzo quartile: separa il 25% dei dati più grandi (maggiori di Q_3) dal restante 75% (minori di Q_3)

in certi casi i quartili sono compresi fra due modalità ed è necessario approssimarne il valore o la posizione

⇛ in presenza di variabili quantitative discrete o qualitative ordinali si associano i quartili ai valori in corrispondenza dei quali la frequenza cumulata raggiunge o supera per la prima volta 0.25, 0.5 e 0.75

*in caso di valori numerici, si trovano così valori diversi da quelli ottenuti con R, che approssima i quartili tramite opportune interpolazioni

```
distr.summary.x(x, stats="quartiles", data)
```

variabili rilevate in classi di intervallo

come per la mediana i quartili possono essere solo approssimati considerando le frequenze cumulate

$F_{k-1} + (Q_1 - x_k) * c_k = 0{,}25$

$F_{k-1} + (Q_2 - x_k) * c_k = 0{,}5$

$F_{k-1} + (Q_3 - x_k) * c_k = 0{,}75$

PERCENTILI

i percentili sono valori che suddividono la sequenza ordinata dei dati in 100 gruppi che includono tendenzialmente lo stesso numero di casi

→ si indica con P_q il q-esimo percentile che separa il q% dei dati più piccoli dal restante (100-q)%

- i quartili forniscono informazioni sui dati intorno al centro
- i percentili forniscono informazioni sulle code (casi estremi)

* i percentili sono calcolati sui dati grezzi e non dipendono dalla classificazione dell'analista

R STUDIO

- la funzione distr. summary. x() consente di determinare i vari tipi di quantili attraverso l'argomento stats

quartili ⇛
```
distr.summary.x(x, stats="quartiles", data)
```

quintili ⇛
```
distr.summary.x(x, stats="quintiles", data)
```

decili ⇛
```
distr.summary.x(x, stats="deciles", data)
```

percentili ⇛
```
distr.summary.x(x, stats="percentiles", data)
```

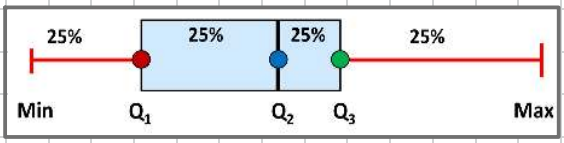
```
distr.summary.x(x, stats=c("p1","...","p50","...","p99"), data)
```

BOX PLOT

il box and whiskers plot è un grafico molto utile per rappresentare in modo univoco e schematico la distribuzione di una variabile numerica senza ricorrere all'istogramma

nella sua versione più semplice il box plot si basa su cinque essenziali numeri di sintesi

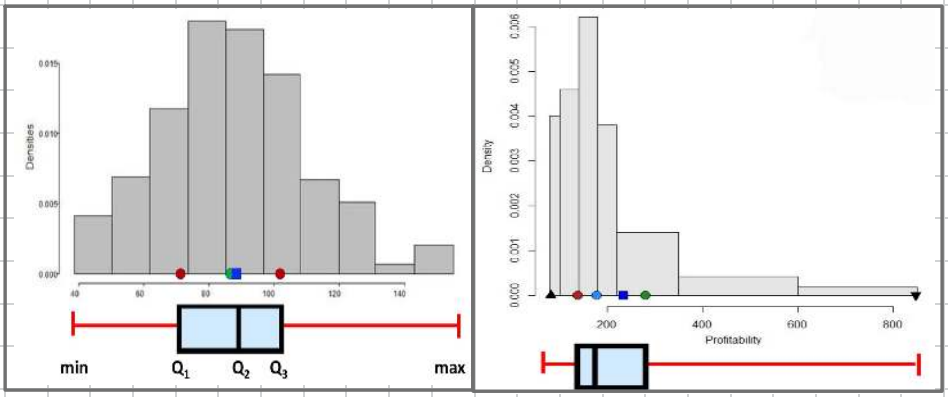
- minimo
- tre quartili
- massimo



scatola → si estende dal primo al terzo quartile e contiene la mediana (secondo quartile)

baffi → collegano la scatola rispettivamente con minimo e massimo valore osservato

il box plot sintetizza le caratteristiche principali di una distribuzione: centro, dispersione locale e globale



in quanto basato su misure di sintesi, esso non varia al variare delle classi di intervallo utilizzate per costruire l'istogramma a partire da dati grezzi

il box plot nella sua versione più elaborata, permette di dare informazioni anche su eventuali valori estremi

questa costruzione si basa sulla distinzione

valori outliers

a seconda che distino dalla scatola più di una volta e mezzo la sua lunghezza o no

valori regolari

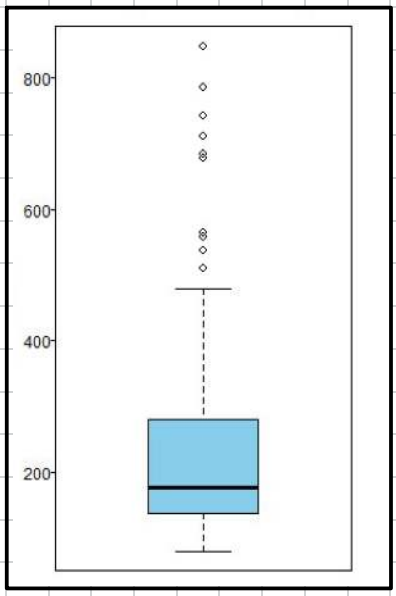
in caso di valori estremi i baffi si fermano ai valori:

sup: $Q_3 + 1.5(Q_3 - Q_1)$ | inf: $Q_1 - 1.5(Q_3 - Q_1)$

in caso di valori solamente regolari i baffi si fermano al massimo e al minimo

nel plot i valori identificati come estremi, che non fanno parte dei baffi, hanno uno specifico simbolo

```
distr.plot.x(x, plot.type="boxplot", data)
```



il box plot è ottenibile solo per variabili numeriche di cui si hanno i dati grezzi (no variabili misurate in classi)

nel caso di variabili rilevate in classi, i quartili, e quindi il box plot, dovranno essere derivati (manualmente) sulla base delle densità delle classi (e in particolare delle aree che sottendono l'isogramma)

a differenza dell'istogramma tramite box plot è possibile apprezzare il collocamento specifico dei valori estremi nelle classi con bassa densità

MISURE DI DISPERSIONE

per confrontare due distribuzioni in termini di dispersione le misure di sintesi non sempre si rivelano le più adatte, tuttavia esistono degli indici che permettono di sintetizzarne efficacemente le differenze

le **misure di dispersione o variabilità** quantificano e sintetizzano la variabilità dei dati osservati

VARIABILITÀ

- campo di variazione
- range interquartile
- varianza
- deviazione standard
- coefficiente di variazione

CAMPO DI VARIAZIONE / RANGE

differenza fra il massimo e il minimo fra i valori osservati

misura non robusta
(molto sensibile a valori estremi)

RANGE / DIFFERENZA INTERQUARTILE

differenza fra primo e terzo quartile, contenente il 50% dei valori centrali

$IQR = Q_3 - Q_1$

esclude il 25% dei dati più piccoli e più grandi

misura robusta
(non influenzata da valori estremi)

VARIANZA

x_i → valore osservato sull'i-esimo caso
$\bar{x}$ → valore della media dei dati osservati

⇒ deviazione di x_i dalla media $(x_i - \bar{x})$ → quantifica la distanza e l'errore di semplificazione fra un dato e la media

la varianza è una misura di dispersione e variabilità dei dati intorno alla media ⇒ nella sua valutazione il focus non è sulla direzione ma sull'entità della dispersione

la **varianza** è la media delle deviazioni dei dati dalla media al quadrato

la varianza assume l'unità di misura dei dati al quadrato (basata sulle deviazioni al quadrato)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

μ → media della popolazione

a differenza della varianza sulla popolazione, la varianza campionaria fa riferimento a una media delle deviazioni non esatta (somma divisa per n-1)

il motivo di questa differenza sarà comprensibile con i concetti di inferenza e stima in particolare

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

$\bar{x}$ → media nel campione

→ la varianza fa riferimento a tutti i valori osservati

→ è basata sulla media quindi è una misura non robusta

è considerabile come l'errore quadratico medio in cui si incorre sostituendo i dati grezzi con la loro media

misura dell'affidabilità della media come sintesi dei dati

FORMULA DI CALCOLO INDIRETTA DELLA VARIANZA CAMPIONARIA

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

la varianza può essere calcolata in funzione della media dei dati al quadrato e del quadrato della media

formula da usare specificatamente in casi di calcolo della varianza a partire da dati raggruppati (variabili rilevate in classi di intervallo)

DIMOSTRAZIONE

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \overset{①}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) \overset{②}{=} \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n}x_i\right] \overset{③}{=}$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 + n\bar{x}^2 - 2n\bar{x}^2\right] \overset{④}{=} \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right] \overset{⑤}{=} \frac{n}{n-1}\left[\sum_{i=1}^{n}\frac{x_i^2}{n} - \bar{x}^2\right] \overset{⑥}{}$$

nel caso di una popolazione lo stesso procedimento porta a

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{i=1}^{N}\frac{x_i^2}{N} - \mu^2$$

DEVIAZIONE STANDARD / SCARTO QUADRATICO MEDIO

lo scarto quadratico medio o deviazione standard (sd) è la radice quadrata della varianza

Campione:

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

varianza ↑ → deviazione standard ↑

dispersione dei dati ↑

è interpretabile come una misura della distanza media dei dati dalla media stessa

Popolazione:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

è possibile calcolare con esattezza la varianza e la deviazione standard a partire dalla tabella di frequenza

- media dei dati (x) → somma dei prodotti fra modalità e frequenze relative
- media dei dati al quadrato ($\bar{x}^2$) → somma dei prodotti fra modalità al quadrato e frequenze relative

⇒ sostituendo i valori nella formula di calcolo indiretta si ricava la varianza e quindi la deviazione standard

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{k=1}^{K}(x_k^* - \bar{x})^2 f_k = \frac{n}{n-1}\sum_{k=1}^{K}(x_k^* - \bar{x})^2 p_k = \frac{n}{n-1}\left[\sum_{k=1}^{K}x_k^{*2}p_k - \bar{x}^2\right]$$

in caso di variabili rilevate in classi le misure di dispersione si possono solo approssimare discretizzando la variabile sui punti centrali degli intervalli ($m_k \cdot p_k$ / $\bar{m}_k \cdot p_k$) e sostituendo i valori approssimati nella formula

COEFFICIENTE DI VARIAZIONE

non è possibile confrontare misure di dispersione (assolute) su variabili in unità di misure diverse

può essere utile misurare la deviazione standard in relazione alla media dei dati

il coefficiente di variazione (cv) coincide con il rapporto fra la deviazione standard e la media

$$CV = \frac{s}{|\bar{x}|} \qquad \bar{x} \neq 0$$

è usato per confrontare la dispersione di variabili con unità di misure diverse (misura adimensionale) o con stessa unità di misura ma medie strutturalmente molto diverse

il coefficiente di variazione non ha un range di valori ben definito → non fornisce informazioni sul livello di dispersione di una singola distribuzione

R STUDIO

```
distr.summary.x(x, stats="dispersion", data)
distr.summary.x(x, stats=c("range","IQrange","sd","var","cv"), data)
```

STATISTICA BIVARIATA

in una moltitudine di applicazioni si è interessati allo studio congiunto di due variabili e delle loro relazioni

le modalità di organizzazione e rappresentazione dei dati dipendono dalle tipologie di variabili in esame:

- entrambe qualitative → in generale con un numero limitato di modalità
- qualitativa-quantitativa → in generale una con poche modalità e una continua
- entrambe quantitative → tipicamente entrambe continue

VARIABILI QUALITATIVE / QUANTITATIVE DISCRETE

le variabili qualitative (o discrete) presentano tipicamente un numero ridotto di modalità

per organizzare i dati in tal caso si considera la distribuzione delle frequenze congiunte, che descrive

- le coppie di modalità (distinte) osservate sulle due variabili,
- la rilevanza (frequenza o percentuale) di ogni coppia di modalità
  - frequenze congiunte assolute → numero di casi che presenta ciascuna coppia di modalità
  - frequenze congiunte relative (corrispondenti percentuali) → proporzione di casi sul totale di ogni coppia di modalità (frequenze congiunte assolute / numero totale di casi)

TABELLE A DOPPIA ENTRATA

la distribuzione delle frequenze congiunte viene organizzata in una tabella a doppia entrata nelle cui celle sono riportate le frequenze congiunte assolute o relative a ogni coppia di modalità (con gli eventuali totali)

	Modalità osservate su Y				
Modalità osservate su X	$X \backslash Y$	y_1^*	y_2^*	...	y_J^*	Totale
	x_1^*	f_{11}	f_{12}	...	f_{1J}	R_1
	x_2^*	f_{21}	f_{22}	...	f_{2J}	R_2
	...	...	...	...	...	...
	x_K^*	f_{K1}	f_{K2}	...	f_{KJ}	R_K
	Totale	C_1	C_2	...	C_J	n

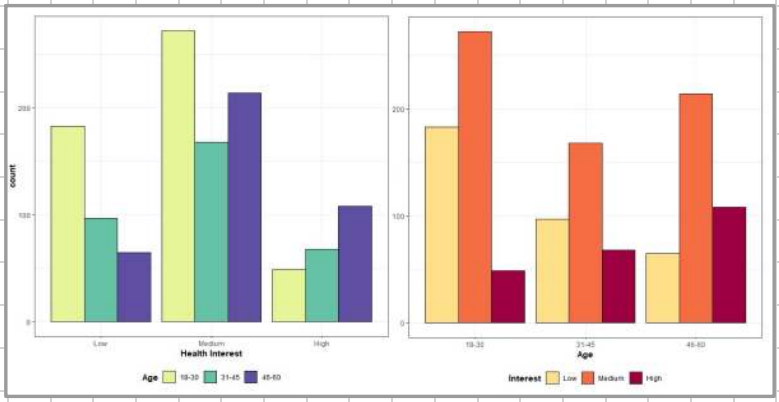
X → variabile sulle righe | Y → variabile sulle colonne

- x/y → modalità delle variabili X/Y
- f_{kj} → frequenze assolute/relative di ogni coppia
- R_k/C_j → distribuzione delle frequenze marginali di X/Y
  - se non ci sono dati mancanti coincidono con la distribuzione univariata delle diverse variabili

per rappresentare graficamente una tabella a doppia entrata si utilizzano in genere i diagrammi a barre

- DIAGRAMMI A BARRE ACCOSTATE
- DIAGRAMMI A BARRE SOVRAPPOSTE

DIAGRAMMA A BARRE ACCOSTATE

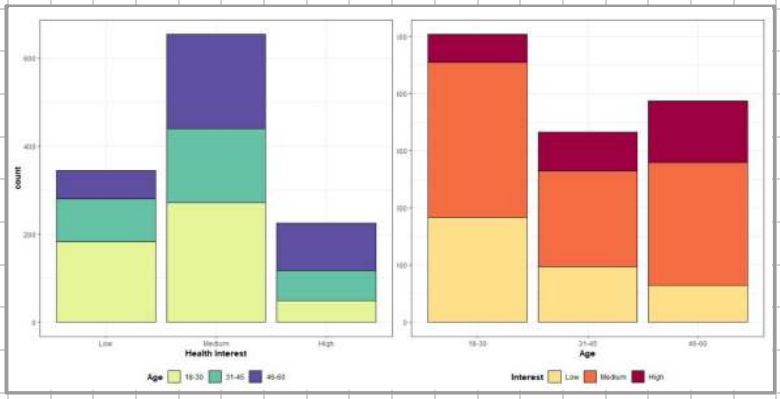


per ogni modalità di una delle variabili viene riportato un insieme di barre di pari ampiezza, una per ogni modalità della seconda variabile

le altezze delle barre sono le frequenze congiunte

l'aspetto non cambia se si considerano le frequenze congiunte relative o assolute

DIAGRAMMA A BARRE SOVRAPPOSTE



per ogni modalità di una delle variabili si riporta una barra costruita sovrapponendo tante barre quante sono le modalità della seconda variabile

le altezze delle barre sono le frequenze congiunte

l'aspetto non cambia se si considerano le frequenze congiunte relative o assolute

---

è possibile costruire grafici, seppur diversi, in funzione di entrambe le variabili
→ quando si incrociano due variabili, una ha sempre maggiore considerazione in funzione dell'altra

tali strumenti non sempre evidenziano al meglio la rilevanza dei valori di una variabile dati quelli dell'altra

per analisi più accurate è utile confrontare le **distribuzioni di frequenza condizionate**

distribuzioni di frequenza di una variabile, nei sottoinsiemi di casi fissati / individuati dalle diverse modalità dell'altra

→ le distribuzioni condizionate si possono costruire per righe o per colonna

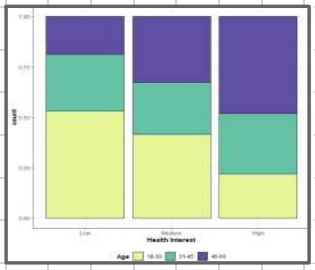
$Y|X = x_k$ → distribuzione di Y condizionata a una modalità di X

$$\text{Freq}\{Y = y_j^* \mid X = x_k^*\} = f_{kj} / R_k \qquad \text{per } j = 1,2,\dots,J$$

$X|Y = y_j$ → distribuzione di X condizionata a una modalità di Y

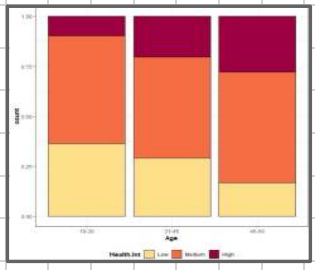
$$\text{Freq}\{X = x_k^* \mid Y = y_j^* \mid\} = f_{kj} / C_j \qquad \text{per } k = 1,2,\dots,K$$

$X \backslash Y$	y_1^*	y_2^*	...	y_J^*	Totale
x_1^*	f_{11}	f_{12}	...	f_{1J}	R_1
x_2^*	f_{21}	f_{22}	...	f_{2J}	R_2
...	...	...	...	...	...
x_K^*	f_{K1}	f_{K2}	...	f_{KJ}	R_K



distribuzioni per riga
(si osserva Y dato X)

$X \backslash Y$	y_1^*	y_2^*	...	y_J^*
x_1^*	f_{11}	f_{12}	...	f_{1J}
x_2^*	f_{21}	f_{22}	...	f_{2J}
...	...	...	...	...
x_K^*	f_{K1}	f_{K2}	...	f_{KJ}
Totale	C_1	C_2	...	C_J



distribuzione per colonna
(si osserva X dato Y)

R STUDIO

- per ottenere le distribuzioni congiunte e/o condizionate si utilizza la **funzione distr. table. xy(...)**

```
distr.table.xy(x,y, freq=c("counts"), freq.type=c("joint"),total=TRUE, data)
```

argomenti
- x e y sono le variabili rispettivamente sulle righe e sulle colonne
- freq indica le frequenze riportate in tabella [assoluta (counts), relativa (prop) o percentuale (perc)]
- freq. type indica quali tipi di frequenze si vuole ottenere [joint, column (x | y) o row (y | x)]
- total specifica se la tabella deve riportare i totali (TRUE) oppure no (FALSE)
- data è il nome del dataframe di riferimento

*è possibile costruire tabelle anche per:
- → variabili numeriche da classificare in intervalli, utilizzando gli argomenti breaks. x e/o breaks. y
- → variabili rilevate in classi di intervallo con gli argomenti interval.x = TRUE e/o interval.y = TRUE

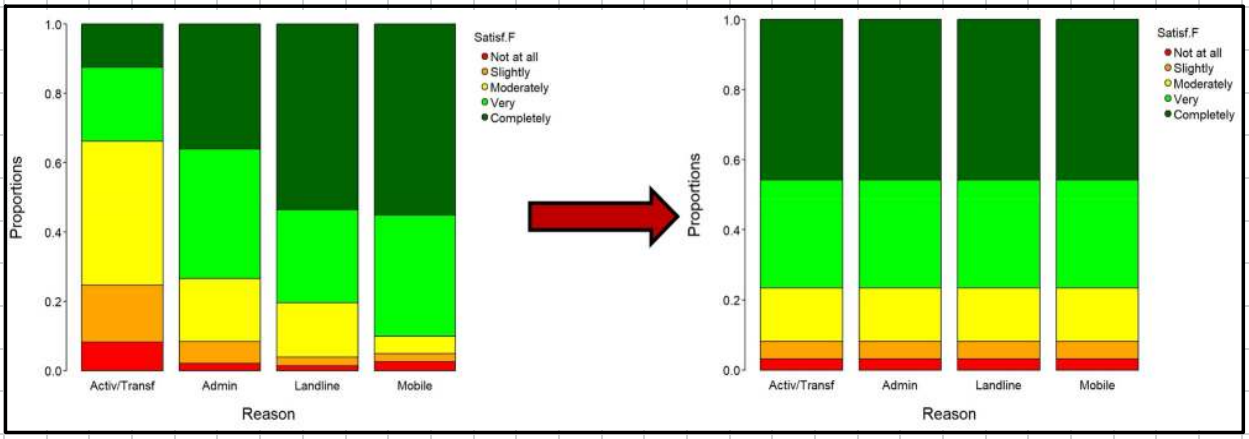
- per rappresentare le distribuzioni congiunte e/o condizionate si utilizza la **funzione distr. plot. xy(...)**

```
distr.plot.xy(x,y,freq="counts",freq.type="joint",plot.type="bars",bar.type="stacked",data)
```

argomenti
- x e y sono le variabili rispettivamente sull'asse orizzontale e sull'asse verticale
- freq e freq. type indicano le frequenze riportate nel grafico [counts,prop, perc / joint,column,row]
- plot. type indica il tipo di grafico da produrre [es. diagramma a barre]
- bar. type specifica il tipo di diagramma a barre [stacked (sovrapposte) o beside (affiancate)]
- data è il nome del dataframe di riferimento

*in questo caso distr. plot. xy() è utilizzato per produrre diagrammi a barre accostate o sovrapposte

*è possibile costruire grafici anche per:
- → variabili numeriche da classificare in intervalli, utilizzando gli argomenti breaks. x e/o breaks. y
- → variabili rilevate in classi di intervallo con gli argomenti interval.x = TRUE e/o interval.y = TRUE



in mancanza di associazione (ossia in caso di indipendenza) le distribuzioni condizionate dovrebbero coincidere tutte con le distribuzioni marginali

grafico: variabile x → motivo del contatto / variabile y → livello di soddisfazione

in caso di indipendenza la distribuzione del livello di soddisfazione dovrebbe essere la stessa a prescindere dal motivo del contatto (tutte le distribuzioni condizionate dovrebbero coincidere con la distribuzione marginale della variabile y)

VARIABILE QUANTITATIVA CONTINUA E VARIABILE QUALITATIVA/DISCRETA

le variabili continue assumono tipicamente un valore diverso per ogni osservazione mentre le variabili discrete assumono valori che possono essere ridotti o molto alti (es. anni di età, numero di transazioni)

⇓ ⇓ ⇓ ⇓ ⇓

strumenti per l'analisi congiunta di due variabili di cui almeno una continua o discreta con molte modalità

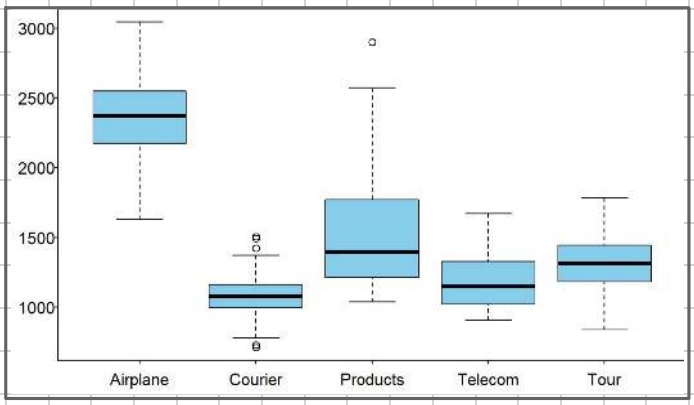
- la tabella a doppia entrata non agevola l'analisi dei dati per via del numero elevato di modalità

→ un primo approccio è classificare in intervalli la variabile con molte modalità (tabella a doppia entrata) e confrontare gli istogrammi che ne esprimono la distribuzione in base alle modalità dell'altra variabile

*tale metodo risulta problematico nel caso in cui le differenze fra distribuzioni condizionate non siano marcate o quando le modalità della variabile condizionante siano relativamente numerose

→ la rappresentazione più efficace per confrontare distribuzioni condizionate è data dai boxplot affiancati

BOXPLOT AFFIANCATI



per ogni modalità della variabile qualitativa si riporta un boxplot che evidenzi i tratti principali della distribuzione della variabile numerica

per confrontare caratteristiche di distribuzione si possono eventualmente utilizzare misure di sintesi della variabile y condizionate alla variabile x

```
distr.plot.xy(x,y, plot.type="boxplot",data)
```

si può costruire solo per variabili numeriche (non è possibile la classificazione in intervalli)

R STUDIO

- la funzione distr. summary. x() consente di ottenere misure di sintesi condizionate specificando le variabili condizionanti (max 2) utilizzando gli argomenti by1 e/o by2

```
distr.summary.x(x, by1, by2, stats, digits=2,f.digits=4, data)
```

→ per classificare in intervalli variabili condizionanti continue si utilizzano gli argomenti breaks. by1 e/o breaks. by2

→ se le variabili condizionanti sono rilevate in classi di intervallo si pone interval.by1 = TRUE e/o interval.by2 = TRUE

stats = "summary" calcola

- minimo [min]
- primo quartile [q1]
- mediana [median/q2]
- media [mean]
- terzo quartile [q3]
- massimo [max]
- deviazione standard [sd]
- varianza [var]

per ogni modalità della variabile condizionante

VARIABILI QUANTITATIVE

in caso di due variabili numeriche continue, la classificazione in intervalli nella tabella a doppia entrata comporta un'eccessiva compressione dei dati pertanto si usa generalmente un diagramma a dispersione

SCATTERPLOT / DIAGRAMMA A DISPERSIONE

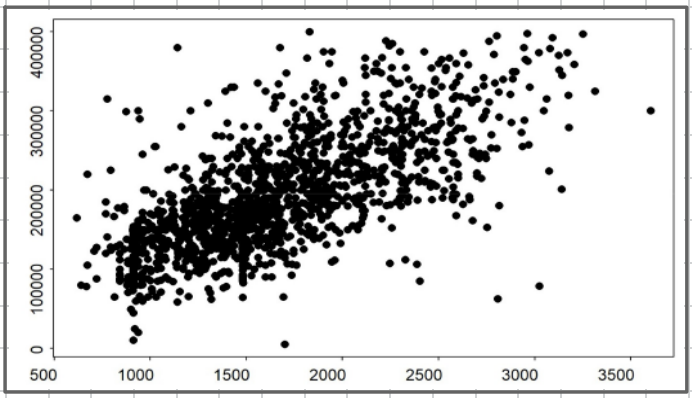
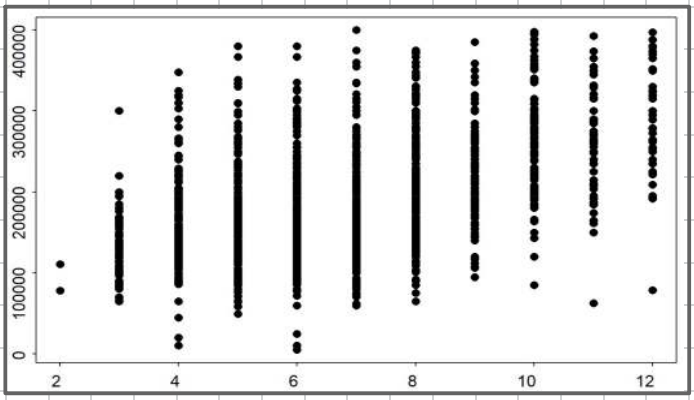


grafico in cui ogni osservazione è individuata da un punto nel piano le cui coordinate coincidono a valori rilevati sulle due variabili riportate sull'asse orizzontale e verticale

lo scatterplot consente di:

- visualizzare la distribuzione congiunta
- identificare eventuali relazioni fra variabili
- identificare eventuali valori anomali/outliers




```
distr.plot.xy(x,y,plot.type="scatter",data)
```

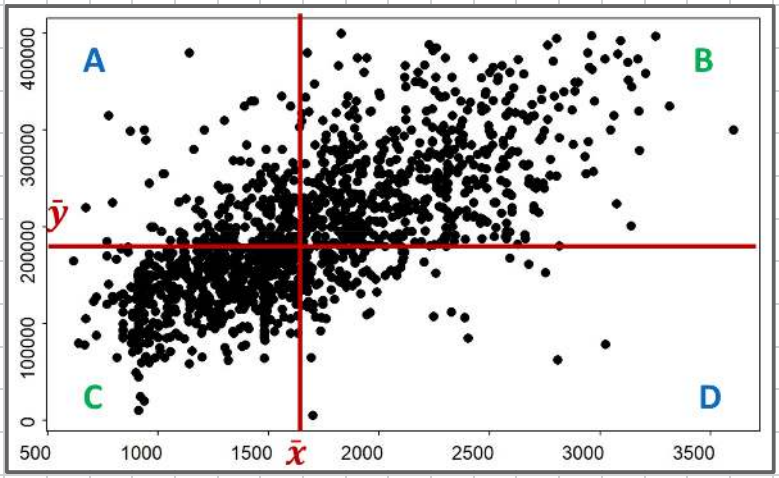
anche nel caso in cui una delle due variabili quantitative sia discreta conviene utilizzare lo scatterplot in quanto ogni osservazione è data da una diversa coppia di valori

la relazione fra le variabili è tanto più debole quanto maggiore è la dispersione dei dati lungo la variabile y

dall'osservazione delle diverse configurazioni che un diagramma a dispersione può assumere, risulta evidente la necessità di individuare dei criteri finalizzati a quantificare l'intensità della relazione fra le due variabili

COPPIE DI VALORI CONCORDANTI E DISCORDANTI

sulla base delle medie delle due variabili è possibile individuare nel plot 4 quadranti



quadranti A-D → coppie di valori discordanti

↳ casi che presentano valori maggiori/minori della media su una variabile e valori minori/maggiori della media sull'altra

quadranti B-C → coppie di valori concordanti

↳ casi che presentano un valore maggiore o minore della media per entrambe le variabili prese in considerazione

due variabili sono
- associate positivamente/in modo diretto → prevalenza di coppie concordanti
- associate negativamente/in modo inverso → prevalenza di coppie discordanti

COVARIANZA

per valutare la forza della relazione va considerata sia la direzione che l'entità delle deviazioni dalle medie

⇒ consideriamo per ogni coppia di valori $(x_i;y_i)$ il prodotto delle deviazioni dalle medie $(x_i - \bar{x})(y_i - \bar{y})$

- osservazioni concordanti → prodotto delle deviazioni positivo — entrambi i valori maggiori o minori della media
- osservazioni discordanti → prodotto delle deviazioni negativo — un valore maggiore, uno minore della media

la covarianza è la media dei prodotti delle deviazioni dei dati dalla media

indice di concordanza (associazione monotona) non di linearità

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

covarianza della popolazione

μ_x e μ_y → medie di X e Y nella popolazione

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

covarianza campionaria

$\bar{x}$ e $\bar{y}$ → medie di X e Y nel campione

come per la varianza, il valore campionario non è del tutto accurato (diviso per n-1 e non n) il motivo di questa differenza sarà comprensibile con i concetti di inferenza e stima in particolare

FORMULA DI CALCOLO INDIRETTA DELLA COVARIANZA CAMPIONARIA

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

la covarianza può essere calcolata in funzione della media dei prodotti dei dati e del prodotto delle medie

formula da usare specificatamente in casi di calcolo della covarianza a partire da dati aggregati (variabili rilevate in classi di intervallo)

DIMOSTRAZIONE

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad ① \quad = \frac{1}{n-1}\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}) \quad ②$$

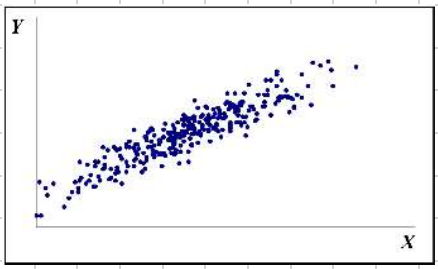
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i y_i) - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}\bar{x}\bar{y}\right] \quad ③$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i y_i) - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right] \quad ④ \quad = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} \quad ⑤$$

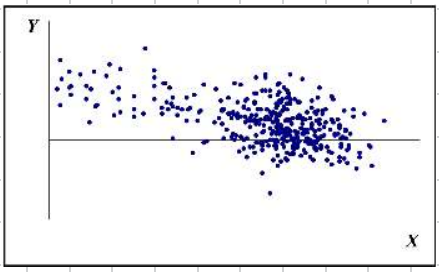
nel caso di una popolazione lo stesso procedimento porta a

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y) = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \mu_X\mu_Y$$

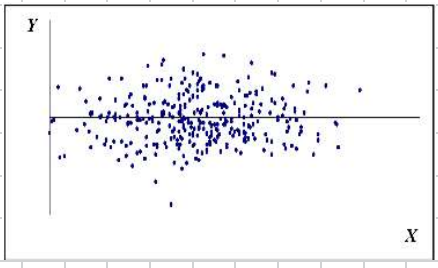
covarianza positiva
(prevalgono coppie concordanti)



covarianza negativa
(prevalgono coppie discordanti)



covarianza prossima a zero
(non si osservano prevalenze)



essendo una misura assoluta e non relativa, dal momento che dipende dalle unità di misura, la covarianza consente di dedurre la direzione ma non la forza della relazione fra due variabili

es: retta = relazione forte e monotona / parabola = relazione forte ma non monotona

COEFFICIENTE DI CORRELAZIONE LINEARE

una misura relativa della forza della relazione lineare fra due variabili è il coefficiente di correlazione lineare

il **coefficiente di correlazione lineare** coincide con il rapporto fra la covarianza e il prodotto degli scarti quadratici medi delle variabili (massimo valore che la covarianza può assumere)

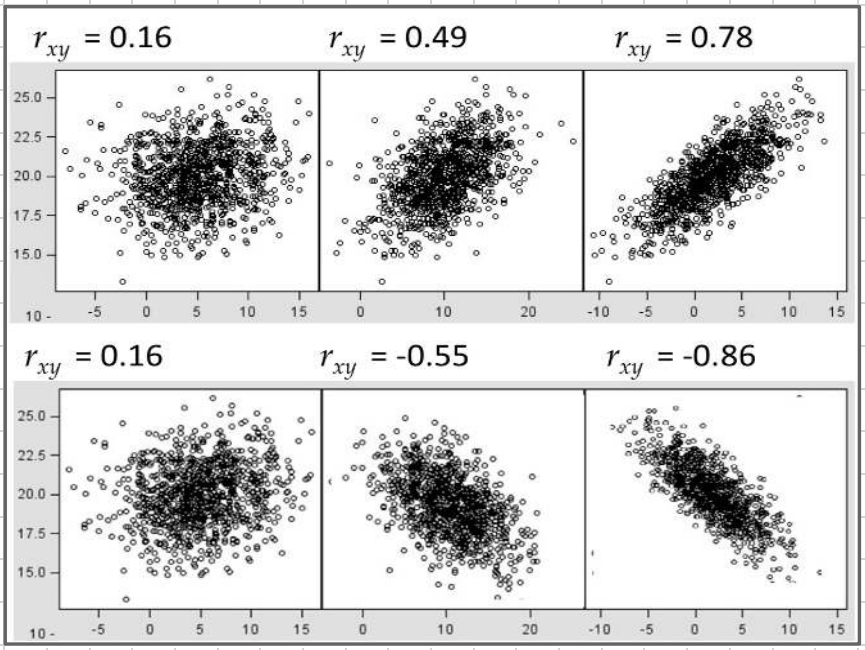
$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$$

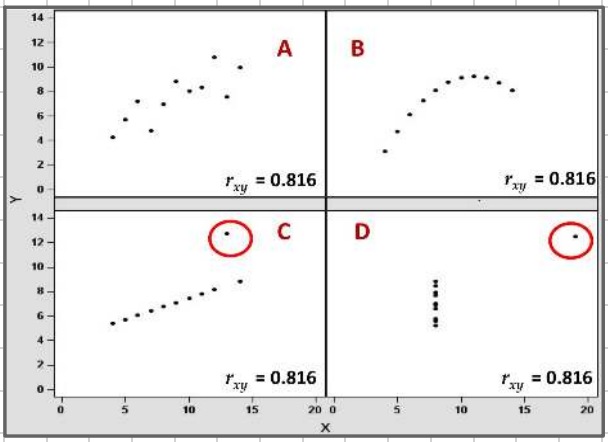
⇛ assume valori compresi fra -1 o +1

- $|r_{xy}| = 1$ → se e solo se la relazione è perfettamente lineare, diretta (+1) o inversa (-1)
- $r_{xy} = 0$ → se le due variabili sono non correlate (non hanno una relazione lineare)

quanto più il coefficiente di correlazione lineare si allontana da 0 verso -1 tanto più i punti tendono a concentrarsi su una retta con pendenza negativa

quanto più il coefficiente di correlazione lineare si allontana da 0 verso +1 tanto più i punti tendono a concentrarsi su una retta con pendenza positiva





in alcuni casi il coefficiente di correlazione non è affidabile

nonostante stessa media, varianza, covarianza e correlazione la relazione fra le variabili è diversa

- A: giusta corrispondenza fra relazione e correlazione
- B: perfetta relazione non lineare (diretta)
- C/D: presenza di valori outliers che alterano r_{xy}

se il coefficiente di correlazione è basso ciò non vuol dire che le variabili non sono relazionate

→ esso misura solo la forza di relazioni lineari

→ il coefficiente può essere molto basso o molto alto anche in caso in cui sussista una relazione non lineare

R STUDIO

- per calcolare covarianza e correlazione lineare tra variabili numeriche si usano le funzioni cov(x, y) e cor(x, y)

argomenti

- x e y sono i vettori delle due variabili analizzate (nome_dataframe$nome_variabile)
- l'argomento use = "complete" permettere di escludere eventuali missing values

FATTORI DI CONFONDIMENTO

talvolta può capitare che delle analisi eseguite su dati grezzi forniscano informazioni completamente o parzialmente diverse rispetto ad analisi effettuate sugli stessi dati aggregati in base ad altre variabili

Paradosso di Simpson — le relazioni tra le frequenze condizionate a specifici gruppi possono scomparire o addirittura invertirsi quando i gruppi sono aggregati

il confronto tra distribuzioni condizionate ha senso solo se queste sono omogenee rispetto a confounding factors (**fattori di confondimento**), che è necessariamente il ricercatore a dover individuare e controllare

es. la frequenza di morte fra vaccinati e non vaccinati è legata all'età (confounding factor) degli individui

	Dead	NoDead	Tot
Vax	3512	281313	284825
NoVax	24364	2676016	2700380
Tot	27876	2957329	2985205

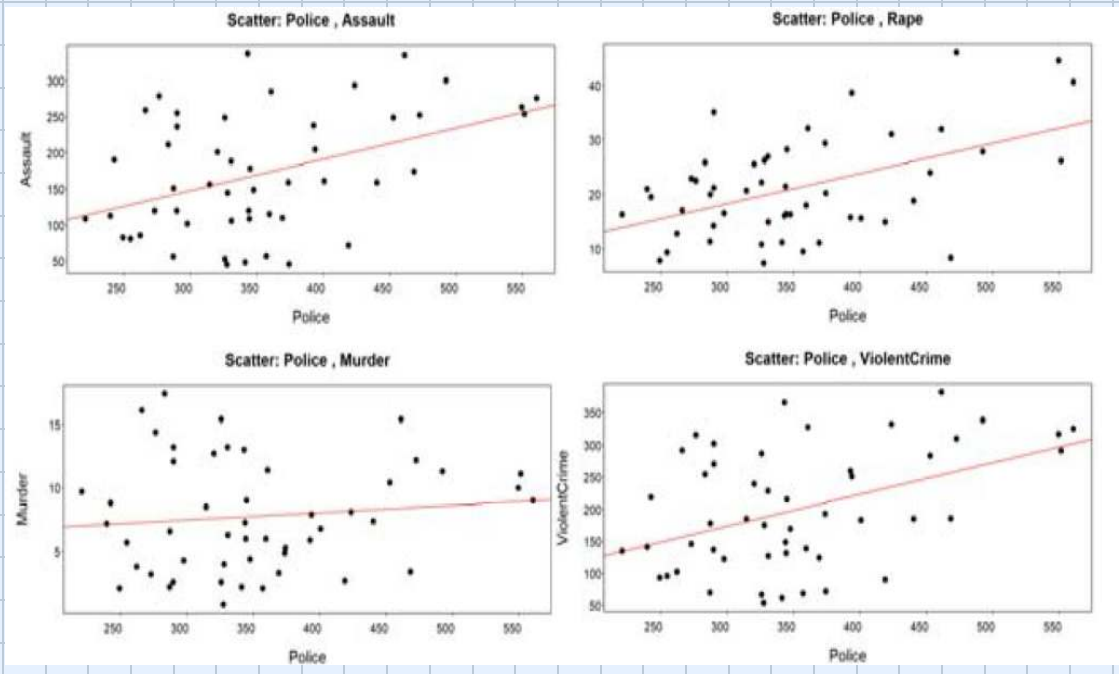
	Dead	NoDead	Tot
Vax	**0.0123**	0.9877	1
NoVax	**0.0090**	0.9910	1

Age	Vax: Dead	Vax: Tot	(Dead|Vax)* 10000	NoVax: Dead	NoVax: Tot	(Dead|NoVax)* 10000
18-49	19	57154	3.32	541	1366914	3.96
50-59	29	31910	9.09	1225	513202	23.87
60-69	76	18981	40.04	2765	424476	65.14
70-79	381	57210	66.60	5901	303782	194.25
80-89	1760	100091	175.84	8679	73510	1180.66
90+	1247	19479	640.18	5253	18496	2840.07
Tot	3512	284825	**123.30**	24364	2700380	**90.22**

---

l'esistenza di un legame lineare non implica nessi di causa-effetto (spesso questo potrebbe essere inverso)

es. tasso di criminalità e risorse investite (l'uno può essere causa dell'altro in base ai fattori considerati)



non si può affermare che il numero di crimini cresce per via del numero di poliziotti

confounding factors:

- tasso demografico
- tasso di denuncia
- concentrazione reddito

STATISTICA INFERENZIALE

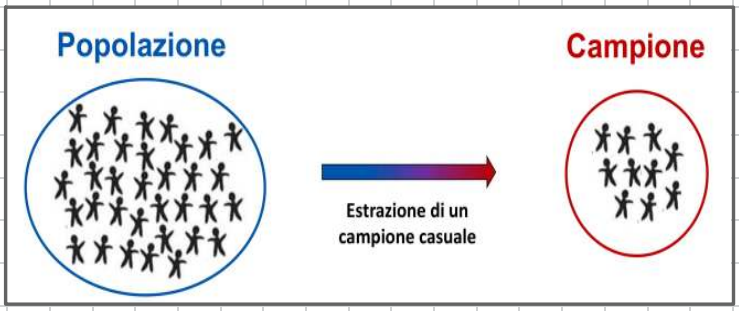
a volte si è interessati a valutare misure che descrivono le caratteristiche di una popolazione (parametri) tuttavia raccoglierne i dati su tutte le unità risulta proibitivo (costi e tempi), difficile o perfino impossibile

→ ciò è quanto si definisce il c.d. **problema inferenziale** ←

in questi casi, è conveniente o necessario rilevare i dati su un campione casuale di unità e fare inferenza sui parametri della popolazione partendo da misure che sintetizzino le caratteristiche del campione: le statistiche

nel fare inferenza è cruciale valutare l'affidabilità dell'estensione, e quindi il rischio ad essa connesso

è necessario valutare la relazione tra parametro di interesse e distribuzione della statistica misurata su tutti i possibili campioni di ampiezza n estraibili



assumendo sia possibile fare assunzioni sulla distribuzione di una variabile nella popolazione si analizzano:

- la distribuzione di statistiche basate su campioni casuali semplici estratti dalla popolazione
- la relazione esistente tra le caratteristiche della distribuzione e il parametro di interesse

per descrivere l'esito dell'estrazione di un campione da una popolazione si usano le **variabili aleatorie** (v.a.)

VARIABILI ALEATORIE

VARIABILI ALEATORIE DISCRETE

una variabile aleatoria discreta è una caratteristica dei casi d'interesse, derivante da un processo di conteggio, di cui non si conosce a priori il valore in corrispondenza di un'unità estratta casualmente

→ si può, tuttavia, valutare la probabilità di ottenere ogni modalità tramite la **funzione di probabilità**

$$P_X(x) = \begin{cases} p_1 & x = x_1 \\ p_2 & x = x_2 \\ p_3 & x = x_3 \\ \vdots & \vdots \\ p_k & x = x_k \\ 0 & \text{altrove} \end{cases}$$

funzione che associa ad ogni modalità x la probabilità p_k che la variabile aleatoria X (per un soggetto estratto in modo casuale) corrisponda a x

es: la probabilità può coincidere con la frequenza relativa di ogni modalità

una variabile aleatoria discreta X può assumere al massimo un insieme numerabile di valori:

- funzione di probabilità → associa a ogni valore x la probabilità che X sia uguale a x

$$P_X(x) = \text{Prob}(X = x)$$

proprietà: $0 \le P_X(x) \le 1$ per ogni x | $\sum_x P_X(x) = 1$

- **funzione di ripartizione** → associa a ogni valore x la probabilità che X sia minore o uguale a x

$$F_X(x) = \text{Prob}(X \le x)$$

VALORE ATTESO E VARIANZA PER V.A. DISCRETE

anche per variabili aleatorie discrete è utile calcolare misure di sintesi come il valore atteso e la varianza

VALORE ATTESO ⇒ somma dei valori possibili (per ogni caso/unità) moltiplicati per le probabilità che si verifichino

$$E(X) = \mu = \sum_x x P_X(x)$$

VARIANZA ⇒ somma degli scostamenti quadratici potenziali moltiplicati per la probabilità che si verifichino

$$Var(X) = \sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2$$

* se la funzione di probabilità riflette esattamente la composizione della popolazione, il valore atteso e la varianza coincidono con la media e la varianza della popolazione

---

DISTRIBUZIONE DI BERNOULLI

la variabile aleatoria di Bernoulli esprime un tipo di distribuzione che descrive se, per un soggetto scelto in maniera casuale nella popolazione, si verifica un evento codificato come successo (X = 1) o insuccesso (X = 0)

$$P_X(x) = \begin{cases} (1-p) & x=0 \\ p & x=1 \\ 0 & altrove \end{cases} = \begin{cases} p^x(1-p)^{1-x} & x=0,1 \\ 0 & altrove \end{cases}$$

valore atteso e varianza di X - Bernoulli

$$E(X) = (1-p)\cdot 0 + p \cdot 1 = p$$

$$Var(X) = (0-p)^2(1-p) + (1-p)^2 p = p - p^2 = p(1-p)$$

VARIABILI ALEATORIE CONTINUE

una variabile aleatoria continua è una caratteristica dei casi d'interesse, derivante da un processo di misurazione, di cui non si conosce a priori il valore in corrispondenza di un'unità estratta casualmente

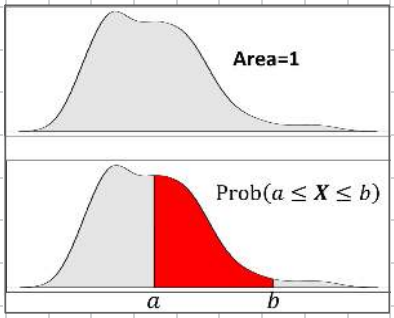
- una variabile aleatoria continua X può assumere qualunque valore in un intervallo
- → si assume che la probabilità che una v.a. continua assuma uno specifico valore x è 0, a prescindere da x

si descrive la popolazione assegnando la probabilità non a ogni singolo valore, ma a ogni possibile intervallo

per descrivere la distribuzione di probabilità di una variabile aleatoria continua X si considera una funzione che consenta di derivare la probabilità che X assuma valori in ogni potenziale intervallo:

- **funzione di densità** → proprietà:
  - $f_X(x) \geq 0$ per ogni x
  - $\int_{-\infty}^{+\infty} f_X(x)dx = 1$

⇒



$f_X(x)$

$$\text{Prob}(a \leq X \leq b) = \int_a^b f_X(x)dx$$

NOTA: solo ed esclusivamente per v.a. continue (no discrete) vale che...

$$\text{Prob}(a \leq X \leq b) = \text{Prob}(a \leq X < b) = \text{Prob}(a < X \leq b) = \text{Prob}(a < X < b)$$

poiché la probabilità assunta in singoli valori specifici è pari a 0

VALORE ATTESO E VARIANZA PER V.A. CONTINUE

la funzione di ripartizione di una v.a. continua X è definita come: $$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

valore atteso e varianza di X si definiscono come nel caso discreto, ma la somma è sostituita dall'integrale

Prob $(a \leq X \leq b) = F(b) - F(a)$ per ogni $a < b$

VALORE ATTESO ⇒ integrale del prodotto fra valori possibili (per ogni caso/unità) e probabilità che si verifichino

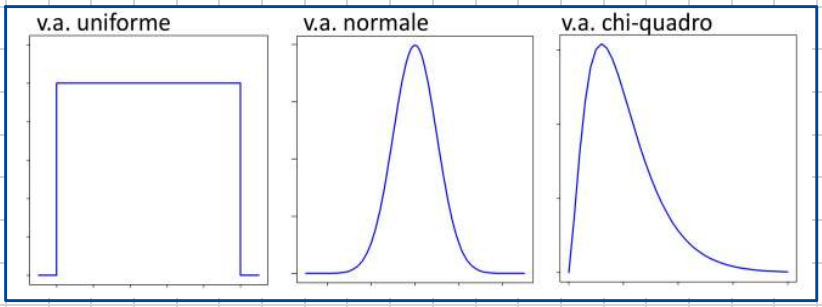
$$E(X) = \mu = \int_{-\infty}^{+\infty} x f_X(x)dx$$

VARIANZA ⇒ integrale del prodotto fra gli scostamenti quadratici e le probabilità che si verifichino

$$Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx$$

DISTRIBUZIONI NOTEVOLI

definire una funzione di densità che sintetizzi in modo adeguato le caratteristiche di una popolazione può risultare complicato pertanto esistono dei modelli funzionali a descrivere determinate situazioni tipiche



distribuzioni notevoli

⇓

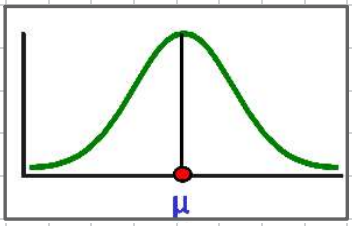
distribuzioni legate a parametri che ne modificano la forma così da adattarla alle caratteristiche ipotizzate del caso

DISTRIBUZIONE NORMALE

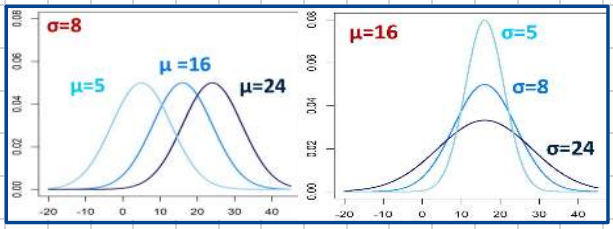
la distribuzione normale è la più importante distribuzione di probabilità, essenziale in ambito inferenziale

una v.a. X ha una distribuzione normale di parametri μ e $\sigma^2 > 0$, che si indica con $X \sim N(\mu,\sigma^2)$ se la sua funzione di densità è : $$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ → valore atteso
σ^2 → varianza



- forma simmetrica e campanulare
- centrata su media e mediana μ
- livello di dispersione legato a σ



i due parametri determinano la forma della distribuzione
- → al variare di (μ) la distribuzione trasla verso destra/sinistra
- → al variare di (σ) la distribuzione si disperde o si concentra

R STUDIO

- per determinare la funzione di ripartizione (probabilità cumulata) e i percentili di una distribuzione normale, dati i parametri μ e σ, si fa riferimento rispettivamente alle **funzioni pnorm() e qnorm()**

`pnorm(q, mean=0, sd=1)`
- **q:** valore in corrispondenza di cui si vuole calcolare la probabilità cumulata, ossia F(q) = Prob (X ≤ q) → è l'area sottesa alla curva di densità fino a q

`qnorm(p, mean=0, sd=1)`
- **p:** valore in corrispondenza del quale la probabilità cumulata è p, ossia l'ordine del percentile x_{1-p} tale che $F(x_{1-p})$ = Prob$(X \leq x_{1-p})$ = p

- mean / sd (con valori di default 0 e 1) consentono di specificare i parametri

APPLICAZIONI UTILI →
- probabilità fra due valori a e b → pnorm(b) - pnorb(a)
- probabilità oltre un dato valore a → 1 - pnorb(a)
- intervallo fra due probabilità → c(qnorm(p_1), qnorm(p_2))

TRASFORMAZIONI LINEARI DI VARIABILI ALEATORIE

data una variabile aleatoria X consideriamo una sua trasformazione lineare → $Y = a + bX$

valore atteso e varianza di Y sono legati al valore atteso e la varianza di X

$$E(Y) = E(a + bX) = a + bE(X) = a + b\mu$$

$$Var(Y) = Var(a + bX) = b^2 Var(X) = b^2\sigma^2 \rightarrow Sd(Y) = |b| \cdot Sd(X) = |b| \cdot \sigma$$

DIMOSTRAZIONE

1. $$E(Y) \overset{①}{=} \sum_x (a + bx)P_X(x) \overset{②}{=} a\sum_x P_X(x) + b\sum_x xP_X(x) \overset{③}{=} a + bE(X) = a + b\mu$$ (dove $\sum_x P_X(x) \rightarrow 1$ e $\sum_x xP_X(x) \rightarrow E(X)$)

2. $$Var(Y) \overset{①}{=} \sum_x [a + bx - (a + b\mu)]^2 P_X(x) \overset{②}{=} \sum_x (bx - b\mu)^2 P_X(x) \overset{③}{=} b^2 \sum_x (x - \mu)^2 P_X(x) \overset{④}{=} b^2 Var(X) = b^2\sigma^2$$

non sempre è possibile determinare la distribuzione di Y a partire dalla distribuzione di X tuttavia vale che:

se una variabile aleatoria ha distribuzione normale, a prescindere dal valore atteso e dalla varianza, qualunque trasformazione lineare ha anch'essa distribuzione normale

$$X \sim \mathcal{N}(\mu, \sigma^2) \rightarrow Y = (a + bX) \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$$

STANDARDIZZAZIONE DI VARIABILI ALEATORIE

la standardizzazione è una particolare trasformazione lineare con specifiche caratteristiche

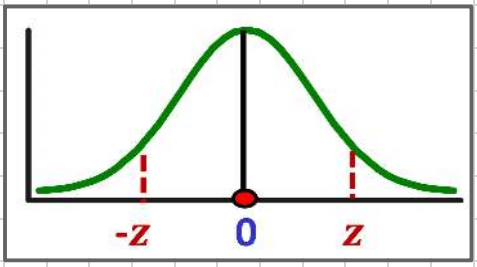
$$Z = \frac{X - \mu}{\sigma}$$

- il valore atteso di una v.a. standardizzata è sempre 0 ⇒ $E(Z) = E\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{E(X)}{\sigma} - \frac{\mu}{\sigma} = 0$
- la varianza di una v.a. standardizzata è sempre 1 ⇒ $Var(Z) = Var\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X) = 1$

standardizzando una variabile aleatoria normale si ottiene la c.d. distribuzione normale standardizzata:

$$X \sim \mathcal{N}(\mu, \sigma^2) \rightarrow Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

v.a. con distribuzione normale, valore atteso pari a 0 e varianza pari a 1



- $\text{Prob}(Z \leq 0) = 0.5 = \text{Prob}(Z \geq 0)$
- se $\text{Prob}(Z \leq z) = p$ e $\text{Prob}(Z \geq z) = 1 - p$ → $\text{Prob}(Z \geq -z) = p$ e $\text{Prob}(Z \leq -z) = 1 - p$

$$X \sim \mathcal{N}(\mu, \sigma^2) \rightarrow \text{Prob}(X < x) = \text{Prob}\left(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}\right) = \text{Prob}\left(Z < \frac{x - \mu}{\sigma}\right)$$

quindi vale che $\rightarrow \text{Prob}(X < x_{1-p}) = p \rightarrow \text{Prob}\left(Z < \frac{x_{1-p} - \mu}{\sigma}\right) = \text{Prob}(Z < z_{1-p}) = p$

quindi poiché $z = (x - \mu) / \sigma$ → $\frac{x_{1-p} - \mu}{\sigma} = z_{1-p}$

$$x_{1-p} = \mu + z_{1-p}\sigma$$
$$x_{1-p} = \mu - z_p\sigma$$

con

- x = qnorm (p, mean = μ, sd=σ)
- z = + qnorm (p, mean = 0, sd=1) = - qnorm (1-p, mean = 0, sd=1)

$$x = \mu + z \cdot \sigma$$ (mean, sd)

COMBINAZIONI LINEARI DI VARIABILI ALEATORIE

per studiare le combinazioni lineari di v.a. bisogna fare riferimento al concetto di **distribuzione congiunta**

⇒ la funzione di probabilità o di densità congiunta di due v.a. X e Y consente di assegnare la probabilità in base a ogni coppia di valori (o intervalli, in caso di variabili continue)

$P_{XY}(x, y) = \text{Prob}(X = x, Y = y)$

$\text{Prob}(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f_{XY}(xy)dxdy$

→ da tali funzioni si possono determinare la covarianza Cov(X,Y) e la correlazione Corr(X,Y) fra due v.a.

COVARIANZA $Cov(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

CORRELAZIONE $Corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

dati valori attesi e varianze di X (μ_x e σ^2_x) e di Y (μ_Y e σ^2_Y)

---

un'importante distribuzione congiunta è la **distribuzione normale bivariata**, per cui vale una proprietà

se X e Y hanno distribuzione congiunta normale → sia X che Y hanno distribuzione normale

non è vero il contrario

---

per l'analisi delle combinazioni lineari di v.a. è di importanza fondamentale il concetto di **indipendenza**

due variabili aleatorie X e Y si definiscono indipendenti se la probabilità di osservare certi valori per una delle due non dipende in alcun modo dai valori assunta dall'altra

la probabilità di osservare congiuntamente valori di X e di Y può essere determinata a partire dalle rispettive distribuzioni (marginali) delle due v.a.

variabili indipendenti

$Cov(X, Y) = Corr(X, Y) = 0$

$P_{XY}(x,y) = \text{Prob}(X = x, Y = y) = \text{Prob}(X = x)\text{Prob}(Y = y) = P_X(x)P_Y(y)$

$f_{XY}(x,y) = f_X(x)f_Y(y)$

---

consideriamo due v.a. X e Y → $E(x) = \mu_x$ / $Var(x) = \sigma^2_x$ / $E(y) = \mu_Y$ / $Var(y) = \sigma^2_Y$ / $Cov(x,y) = \sigma_{XY}$

la loro combinazione lineare (aX + bY) ha valore atteso e varianza:

$E(aX + bY) = aE(X) + bE(Y) = a\mu_x + b\mu_y$

$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y) = a^2\sigma^2_x + b^2\sigma^2_Y + 2ab\sigma_{XY}$

se X e Y sono indipendenti e la covarianza è nulla:

$Var(aX + bY) = a^2\sigma^2_x + b^2\sigma^2_Y$

la distribuzione di (aX + bY) dipende dalla distribuzione congiunta delle due variabili aleatorie

se X e Y hanno distribuzione congiuntamente normale ⇒ $(aX + bY) \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{XY})$

*
FORMULA DI CALCOLO: la covarianza di due variabili è il prodotto del coefficiente di correlazione lineare e le varianze delle variabili stesse

$\sigma_{AB} = \rho_{AB}\sigma_A\sigma_B$

SOMMA E MEDIA DI VARIABILI ALEATORIE I.I.D.

data una v.a. X con valore atteso μ e varianza σ^2, si considerando n v.a. $X_1, X_2, \dots, X_n$ esse si definiscono:

variabili i.i.d. [indipendenti e identicamente distribuite] $\left\{ \begin{array}{l} \bullet \text{ se sono indipendenti (covarianza nulla per ogni coppia)} \\ \bullet \text{ se hanno la medesima distribuzione della v.a. X} \end{array} \right.$

es. caso di n unità estratte a caso da una stessa popolazione: ogni X_i descrive il risultato dell'estrazione

particolari combinazioni lineari delle n v.a. $X_1, X_2, \dots, X_n$ i.i.d sono costituite dalla somma e dalla media:

- **somma**: $S = X_1 + X_2 + \dots + X_n$
- **media**: $\overline{X} = (X_1 + X_2 + \dots + X_n)/n = S/n$

$\Rightarrow$ esse corrispondono a v.a., i cui valori attesi e le cui varianze si possono determinare a partire da μ e σ

somma:

$$\rightarrow E(S) = E(X_1) + \dots + E(X_n) = n\mu$$

$$\rightarrow Var(S) = Var(X_1) + \dots + Var(X_n) = n\sigma^2$$

per via dell'identica distribuzione vale che:

$\left\{ \begin{array}{l} \rightarrow E(X_i) = \mu \\ \rightarrow Var(X_i) = \sigma^2 \end{array} \right.$ hanno tutte stesso valore atteso e stessa varianza

media:

$$\rightarrow E(\overline{X}) = E(X_1/n) + \dots + E(X_n/n) = n \cdot \mu / n = \mu$$

$$\rightarrow Var(\overline{X}) = Var(X_1/n) + \dots + Var(X_n/n) = n \cdot \sigma^2 / n^2 = \sigma^2 / n$$

$\left\{ \begin{array}{l} \rightarrow E(X_i/n) = \mu/n \\ \rightarrow Var(X_i/n) = \sigma^2/n^2 \end{array} \right.$

---

anche se è possibile trovare valore atteso e varianza della somma e della media di variabili i.i.d. a partire dal valore atteso e dalla varianza comune, non è sempre possibile determinarne le funzioni di probabilità

DISTRIBUZIONE NORMALE

la somma e la media di n v.a. $X_1, X_2, \dots, X_n$ i.i.d. secondo una normale hanno anch'esse distribuzione normale, con valore atteso e varianza ottenuti nel medesimo modo

$X_1, X_2, \dots, X_n$ i.i.d. $X \sim \mathcal{N}(\mu, \sigma^2)$ $\Rightarrow$

$$S = (X_1 + X_2 + \dots + X_n) \sim \mathcal{N}(n\mu, n\sigma^2)$$
$$\overline{X} = (X_1 + X_2 + \dots + X_n)/n \sim \mathcal{N}(\mu, \sigma^2/n)$$

TEOREMA DEL LIMITE CENTRALE

la somma e la media di n v.a. $X_1, X_2, \dots, X_n$ i.i.d. come una v.a. X, hanno distribuzione tale che, qualunque sia la distribuzione di X, per n sufficientemente elevato, può essere approssimata dalla distribuzione normale (il valore atteso e la varianza si ottengono conseguentemente)

$\rightarrow$ se n non è sufficientemente elevato, non si può dire nulla su valori attesi e varianze, a meno che non si assuma che le v.a. in considerazione siano distribuite normalmente

DISTRIBUZIONE BERNOULLIANA

date n v.a. $X_1, X_2, \dots, X_n$ i.i.d. $X \sim$ Bernoulli(p), dove il parametro p pari alla probabilità di osservare un successo all'interno di una popolazione, con $E(X) = p$ e $Var(X) = p(1-p)$

- **numero di successi**: $S = X_1 + X_2 + \dots + X_n$
- **proporzione di successi**: $\hat{P} = (X_1 + X_2 + \dots + X_n)/n$

per n sufficientemente elevato (tipicamente n > 30), la distribuzione della somma e della proporzione si può approssimare con una normale $\Rightarrow$

$$S \approx \mathcal{N}(np, np(1-p))$$
$$\hat{P} \approx \mathcal{N}(p, p(1-p)/n)$$

INFERENZA E STIME

la statistica inferenziale riguarda le procedure necessarie per fare estrapolazioni su parametri di una popolazione X a partire da statistiche, calcolate su campioni casuali semplici $X_1, X_2, \dots, X_n$ estratto da X

→ la stima di un parametro può essere:
- puntuale (singolo valore)
- per intervallo (di valori)

→ la verifica di ipotesi su un parametro: consiste nella valutare quale fra due possibili ipotesi è maggiormente supportata dai risultati campionari

parametro (θ) → misura di sintesi della popolazione, con riferimento ad una variabile aleatoria X

esempi: media μ, deviazione standard σ, proporzione campionaria p — misura della popolazione

stimatore ($\hat{\theta}$) → statistica utilizzata per stimare il parametro: funzione dei risultati rilevati sul campione casuale di n estrazioni $X_1, X_2, \dots, X_n$ i.i.d. come X, tale che $\hat{\theta} = f(X_1, \dots, X_n)$

→ è una variabile aleatoria che descrive le realizzazioni (aleatorie) di $\hat{\theta}$ in corrispondenza di tutti i campioni estraibili $X_1, \dots, X_n$ dalla popolazione X — misura campionaria

stima (ϑ) → realizzazione campionaria di un determinato stimatore $\hat{\theta}$ osservata in corrispondenza del campione specifico effettivamente estratto $x_1, \dots, x_n$

esempi: campione X (1, 3, 6, 10), stimatore $\overline{X}$ (media), stima/realizzazione di $\overline{X}$ ($\overline{x} = 5$)

→ l'affidabilità del processo inferenziale si valuta tramite lo stimatore con migliori proprietà, in particolare

è possibile interrogarsi sulla coincidenza fra il valore atteso dello stimatore usato in relazione a un certo parametro di interesse e il parametro stesso ma anche sulla dispersione delle stime intorno al parametro

*è bene ricordare, però, che in corrispondenza di uno specifico campione non si potrà fare alcuna valutazione

PROPRIETÀ DEGLI STIMATORI

CORRETTEZZA / NON DISTORSIONE

la distorsione è data dalla differenza fra valore atteso dello stimatore e il valore del parametro da stimare

$$D(\hat{\theta}_n) = [E(\hat{\theta}_n) - \theta]$$

uno stimatore puntuale $\hat{\theta}_n$, basato su un campione di n unità, si dice corretto per un parametro θ quando il suo valore atteso è pari al parametro da stimare per ogni valore di n e di θ [$E(\hat{\theta}_n) = \theta$ per ogni valore di n e θ → distorsione = 0]

uno stimatore distorto per un parametro θ si asintoticamente non distorto quando, all'aumentare dell'ampiezza campionaria, la distorsione diminuisce

$$\lim_{n\to\infty} D(\hat{\theta}_n) = 0 \iff \lim_{n\to\infty} E(\hat{\theta}_n) = \theta$$

EFFICIENZA

la varianza di un generico stimatore $\hat{\theta}$ non distorto corrisponde alla deviazione al quadrato attesa dal parametro (θ) per una generica realizzazione dello stimatore

la varianza o lo scarto quadratico medio (standard error) di stimatori corretti danno informazioni sulla loro affidabilità

$$E\left[(\hat{\theta} - \theta)^2\right] = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right] = Var(\hat{\theta})$$

⇒ fra gli stimatori corretti per un certo parametro, è preferibile quello con varianza minima, in quanto garantisce maggior concentrazione delle stime intorno al parametro di interesse

STIMATORE DEL VALORE ATTESO

nello studio di una popolazione descritta da una v.a. X si è in genere interessati al parametro dato da:

media della popolazione μ ⇒ uno stimatore valido è la **media campionaria $\bar{X}$** campione di $X_1,\dots,X_n$ i.i.d.

- essendo il valore atteso della media campionaria: $E(\bar{X}) = n \cdot \mu/n = \mu$ → $\bar{X}$ è uno stimatore non distorto per μ
- essendo la varianza della media campionaria: $Var(\bar{X}) = \sigma^2/n$ → n ↑ = Var(X) ↓ (dispersione intorno alle stime ↓)

* dato n, si dimostra che $\bar{X}$ è lo stimatore non distorto con varianza minima (più efficiente)
** inoltre se X ha distribuzione normale, anche la distribuzione di $\bar{X}$ è normale (lo stesso vale qualunque sia la distribuzione di X, se n è abbastanza elevato: teorema del limite centrale)

es. dati n campioni e una deviazione standard σ si calcoli la stima della media e dello standard error

su R studio
- media stimata → Media <- mean (nome_dataframe$nome_colonna)
- standard error → Standard_Error <- σ / sqrt(n)

- non si può concludere che lo scostamento atteso tra la media nella popolazione μ e la stima della media x sia pari allo standard error rilevato, in quanto quest'ultimo vale per una generica stima e non per una specifica
- inoltre notiamo come minore sia standard error dello stimatore, maggiore è la probabilità di stime vicine a μ

STIMATORE DELLA VARIANZA

spesso è ignota la **varianza di una popolazione σ^2** ⇒ uno stimatore valido è la **varianza campionaria S^2**

- si dimostra che $E(S^2) = \sigma$ → S^2 è uno stimatore corretto per σ^2 qualunque sia la distribuzione di X

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

- la varianza campionaria non corretta è caratterizzata da $E(\tilde{S}^2) = (n - 1/n) \cdot \sigma^2$

$$\tilde{S}^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}$$

→ $\tilde{S}^2$ è uno stimatore distorto per σ^2 ma asintoticamente non distorto

es. dati n campioni, ma con deviazione standard σ ignota, si calcoli il valore dello standard error

su R studio
- varianza stimata → Varianza_S2 <- var (nome_dataframe$nome_colonna)
- standard error → Standard_Error <- Varianza_S2 / sqrt(n)

STIMATORE DELLA PROPORZIONE

si può essere interessati alla **proporzione p** di casi in una popolazione che presentano una caratteristica ⇒ si può ricorrere alla **proporzione campionaria $\hat{P}$**

→ le v.a. $X_1,\dots,X_n$ sono i.i.d. come una distribuzione di Bernoulli, pertanto $\hat{P}$ coincide con la media campionaria

- $E(\hat{P}) = E(X) = p$ → $\hat{P}$ è uno stimatore non distorto per p
- $Var(\hat{P}) = Var(X)/n = p(1-p)/n$ → la varianza di $\hat{P}$ è tanto più piccola quanto più ampio è il campione (la varianza non è mai nota ma si stima sostituendo p con la stima $\hat{p}$)

⇒ nel caso in cui n sia sufficientemente grande, la distribuzione di P può essere approssimata da una normale, $P \approx \mathcal{N}(p,\ p(1-p)/n)$

es. si stimi la proporzione p di casi risolti e lo standard error dello stimatore

su R studio
- distribuzione → table (nome_dataframe$nome_colonna) → su n casi, si ottiene Y (successi)
- proporzione campionaria → si ricava p = Y/n
- standard error → Standard_Error <- sqrt(p*(1-p)/n)

STANDARD ERROR E AMPIEZZA CAMPIONARIA

per ottenere uno standard error inferiore, è necessario aumentare l'ampiezza campionaria

⇛ nel caso di uno stimatore $\bar{X}$ (media campionaria), di cui si conosce la deviazione standard σ, se si vuole mantenere lo standard error al di sotto un determinato valore z, bisognerà soddisfare la condizione:

$$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \leq z \rightarrow \sqrt{n} \geq \frac{\sigma}{z} \rightarrow n \geq \frac{\sigma^2}{z^2}$$

*qualora non si conosca il valore della deviazione standard/varianza si può usare quella campionaria, ossia una stima (S^2)

resta il fatto che per via dell'aleatorietà delle estrazioni non si possono fare confronti fra stime specifiche

NOTA: su R studio le funzioni mean() e var() calcolano rispettivamente media e varianza campionaria

STIMATORI PER DIFFERENZA TRA MEDIE

spesso si può essere interessati a valutare le differenze tra le medie di due popolazioni

Popolazione 1 – descritta da una v.a. X	Popolazione 2 – descritta da una v.a. Y
$X_1, \dots, X_{n_X}$ campione di ampiezza n_X iid	$Y_1, \dots, Y_{n_2}$ campione di ampiezza n_Y iid
X con media μ_X e varianza σ_X^2	Y con media μ_Y e varianza σ_Y^2
$\bar{X} = (X_1 + \dots + X_{n_X})/n_X$	$\bar{Y} = (Y_1 + \dots + Y_{n_Y})/n_Y$
$S_X^2 = \sum_{i=1}^{n_X} \frac{(X_i - \bar{X})^2}{n_X - 1}$	$S_Y^2 = \sum_{i=1}^{n_Y} \frac{(Y_i - \bar{Y})^2}{n_Y - 1}$

il parametro di interesse è la differenza fra medie → $\mu_X - \mu_Y$

⇓

si utilizza lo stimatore delle medie campionarie → $\bar{X} - \bar{Y}$

- $E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$ → $(\bar{X} - \bar{Y})$ è quindi uno stimatore non distorto per $\mu_X - \mu_Y$
- la varianza, e quindi lo standard error, dello stimatore dipendono dalle relazioni tra le due popolazioni (e quindi tra i campioni da queste estratte) e dalle ipotesi relative alla distribuzione congiunta di X e Y

⇛ anche la distribuzione di $(\bar{X} - \bar{Y})$ dipende dalla relazione tra X e Y e dalla loro distribuzione congiunta

→ si può genericamente dire che con n sufficientemente elevati la distribuzione delle medie campionarie è approssimabile con una normale (nonostante le varianze non siano note)

CAMPIONI INDIPENDENTI

i campioni vengono estratti indipendentemente l'uno dall'altro (possono avere ampiezza diversa)

⇛ essendo le medie campionarie indipendenti e non correlate:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} \rightarrow SE_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

standard error ($SE_{\bar{X}-\bar{Y}}$)

raramente entrambe le varianze sono note:

1. se si assumono diverse la varianza comune si calcola sostituendo le varianze campionarie
2. se si assumono uguali la varianza comune si calcola con la varianza campionaria corretta pooled

varianza campionaria pooled → media ponderata delle varianze campionarie →

$$S_{Pool}^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

⇛ stima dello standard error ($se_{\bar{X}-\bar{Y}}$)

1. $\sigma_X^2 \neq \sigma_Y^2 \rightarrow \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$
2. $\sigma_X^2 = \sigma_Y^2 \rightarrow \sqrt{\frac{s_{Pool}^2}{n_X} + \frac{s_{Pool}^2}{n_Y}}$

CAMPIONI APPAIATI

i campioni sono relativi a misurazioni effettuate su stesse unità statistiche (hanno necessariamente stessa ampiezza)

per ognuna delle n unità sono disponibili due misurazioni X_i e Y_i, le cui differenze $D_i = X_i - Y_i$ costituiscono le misurazioni della v.a. $D = X - Y$ per $i = 1,\dots,n$ (stimata tramite media campionaria delle differenze $\bar{D} = \bar{X} - \bar{Y}$)

- $E(\bar{D}) = \mu_D = E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$
- $Var(\bar{D}) = Var(D)/n = \sigma_D^2/n$ con

$$Var(D) = Var(X - Y) = Var(X) + Var(Y) - 2Cov(X,Y) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$$

è rarissimo che $Var(D)$ sia nota, in quanto bisognerebbe fare ipotesi sulla varianza delle differenze (σ_D^2) o sulle varianze di X e Y e la loro covarianza

→ per stimare la varianza $\bar{D} = \bar{X} - \bar{Y}$ si utilizza la varianza campionaria corretta delle differenze

$$S_D^2 = \sum_{i=1}^{n} \frac{(D_i - \bar{D})^2}{n-1} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} + \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1} - 2\sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = S_X^2 + S_Y^2 - 2S_{XY}$$

S_D^2 viene sostituita a σ_D^2 per stimare varianza e standard error di $\bar{D} = \bar{X} - \bar{Y}$

⇒ standard error ($SE_{\bar{X}-\bar{Y}}$) $\sqrt{\frac{\sigma_D^2}{n}} = \sqrt{\frac{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}{n}}$

stima dello standard error ($se_{\bar{X}-\bar{Y}}$) $\sqrt{\frac{s_D^2}{n}} = \sqrt{\frac{s_X^2 + s_Y^2 - 2s_{XY}}{n}}$

STIMATORI PER DIFFERENZA TRA PROPORZIONI

spesso si può essere interessati a confrontare le proporzioni di un fenomeno in due popolazioni

Popolazione 1	Popolazione 2
$X_1, \dots, X_{n_X}$ campione di ampiezza n_X iid X distribuita secondo una Bernoulli di parametro p_X (cioè, X assume valore 1 o 0 a seconda che si osservi o meno un successo, e p_X = proporzione di successi nella popolazione), con $E(X) = p_X$ e $Var(X) = p_X(1 - p_X)$	$Y_1, \dots, Y_{n_Y}$ campione di ampiezza n_Y iid Y distribuita secondo una Bernoulli di parametro p_Y (cioè, Y assume valore 1 o 0 a seconda che si osservi o meno un successo, e p_Y = proporzione di successi nella popolazione), con $E(Y) = p_Y$ e $Var(Y) = p_Y(1 - p_Y)$
$\hat{P}_X = (X_1 + \dots + X_{n_X})/n_X$ = proporzione campionaria di successi	$\hat{P}_Y = (Y_1 + \dots + Y_{n_Y})/n_Y$ = proporzione campionaria di successi

il parametro di interesse è: differenza fra proporzioni ⇓ $p_X - p_Y$

si utilizza lo stimatore: proporzioni campionarie ⇓ $\hat{P}_X - \hat{P}_Y$

- $E(\hat{P}_X - \hat{P}_Y) = E(\hat{P}_X) - E(\hat{P}_Y) = p_X - p_Y$ → $(\hat{P}_X - \hat{P}_Y)$ è quindi uno stimatore non distorto per $p_X - p_Y$
- la varianza, e quindi lo standard error, dello stimatore dipendono dalle relazioni tra le due popolazioni

CAMPIONI INDIPENDENTI

$$Var(\hat{P}_X - \hat{P}_Y) = Var(\hat{P}_X) + Var(\hat{P}_Y) = \frac{Var(X)}{n_X} + \frac{Var(Y)}{n_Y} = \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}$$

non è mai nota, ma si può stimare sostituendo a p_X e p_Y le relative stime

⇒ standard error ($SE_{\hat{P}_X-\hat{P}_Y}$) $\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}$

stima dello standard error ($se_{\hat{P}_X-\hat{P}_Y}$) $\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}$

RIASSUNTO STIMATORI E STANDARD ERROR

Parametro	Stimatore	Stima	Campioni	Standard error, *SE*	Stima standard error, *se*
μ	$\overline{X}$	$\overline{x}$		$\sigma/\sqrt{n}$	$s/\sqrt{n}$
p	$\widehat{P}$	$\widehat{p}$		$\sqrt{p(1-p)/n}$	$\sqrt{\widehat{p}(1-\widehat{p})/n}$
$\mu_X - \mu_Y$	$\overline{X} - \overline{Y}$	$\overline{x} - \overline{y}$	Indipendenti	$\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$	$\sigma_X^2 = \sigma_Y^2 \rightarrow \sqrt{\frac{s_{Pool}^2}{n_X} + \frac{s_{Pool}^2}{n_Y}}$ $\sigma_X^2 \neq \sigma_Y^2 \rightarrow \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$
			Appaiati	$\sqrt{\frac{\sigma_D^2}{n}} = \sqrt{\frac{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}{n}}$	$\sqrt{\frac{s_D^2}{n}} = \sqrt{\frac{s_X^2 + s_Y^2 - 2s_{XY}}{n}}$
$p_X - p_Y$	$\widehat{P}_X - \widehat{P}_Y$	$\widehat{p}_X - \widehat{p}_Y$	Indipendenti	$\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}$	$\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n_X} + \frac{\widehat{p}_Y(1-\widehat{p}_Y)}{n_Y}}$

PER DUBBI O SUGGERIMENTI SULLA DISPENSA

MARCO FORMISANO

marco.formisano@studbocconi.it

@marco_formisano__

+39 3313433934

PER INFO SULL'AREA DIDATTICA

MARCO FORMISANO

marco.formisano@studbocconi.it

@marco_formisano__

+39 3313433934

ELENA CACIOLI

elena.cacioli@studbocconi.it

@elenacaciolii_

+39 3928931605

TEACHING DIVISION

I NOSTRI PARTNERS

70C+ CLUB

ETHAN SUSTAINABILITY

DELIVERY VALLEY
NO GENDER KITCHEN

LA PIADINERIA

LA PIADINERIA dal 1994