Brought to you by:

ASTRA

STATISTICA

2° ANNO CLEAM / CLEF

Written by
Matteo Cordaro

2022-2023 Edition

Find more at:

astrabocconi.it

Ciao! Sono Matteo, studente CLEAM del terzo anno.
Questa dispensa è innanzi tutto il mio metodo di studio ed in questo modo preparo tutti i miei esami: spero che possa essere utile anche a voi per preparare i vostri.
Vi segnalo che sono disponibile per ripetizioni nelle materie di cui realizzo le dispense.
Se siete interessati, qui trovate il mio Instagram:
@_matteocordaro.

Buono studio!

<table>
<tr><th colspan="2">CONCETTI DI BASE DELLA STATISTICA</th></tr>
<tr><td>Terminologia di base</td><td>

Popolazione (N) → numero di unità statistiche nella popolazione
Parametri → caratteristiche della popolazione
Campione (n) → sottoinsieme di unità statistiche selezionate dalla popolazione la cui ampiezza è inferiore rispetto a quella della popolazione (n << N)
Statistica → caratteristiche del campione
Variabilità campionaria → il valore di una statistica dipende dal campione selezionato; essa assume un valore diverso/varia da campione a campione → incertezza nelle conclusioni/decisioni
Statistica descrittiva: banca della statistica che consiste in un insieme di strumenti (tabelle, grafici, indici, misure di sintesi) volti a sintetizzare ed organizzare i dati grezzi (dataset) al fine di farne emergere caratteristiche rilevanti per il fenomeno e/o il problema decisionale affrontato
Statistica inferenziale: consiste in un insieme di logiche e procedure (stimatori, intervalli di confidenza, verifiche di ipotesi) volte, in primo luogo, ad estendere e generalizzare le conclusioni tratte dal campione all’intera popolazione e, in secondo luogo, a misurare e controllare l’incertezza nelle conclusioni, dovuta alla variabilità campionaria, ovvero l’errore statistico
Variabile statistica: aspetto, caratteristica delle unità statistiche della popolazione
Modalità: uno dei possibili valori (non necessariamente numerico) che può assumere una variabile
Tipi di variabili statistiche:

Qualitative/Categoriche → esprimono una qualità posseduta dall’unità, l’appartenenza ad una categoria; modalità: nomi, aggettivi

Ordinali: modalità ordinabili in modo oggettivo (livello di gradimento alto, medio, basso)
Nominali: non c’è un ordinamento logico

Quantitative/Numeriche → frutto della rilevazione di una quantità sull’unità; modalità: valori numerici

Discrete → con poche modalità distinte e derivante da un processo di conteggio→ numero di esami superati, numero di componenti della famiglia
Continue → con molte modalità distinte (spesso le quantità monetarie) → qualsiasi numero reale in un intervallo → tempo impiegato per arrivare all’università

Scale di misurazione (da aggiungere se richiesto il tipo di variabile):

Scala nominale
Scala ordinale
Scala a livello di intervallo → importanza delle distanze tra le modalità: l’origine è arbitraria (zero in senso arbitrario)
Scala a livello di rapporto → importanza delle distanze e dei rapporti tra le modalità: l’origine è fissa (zero in senso assoluto)

</td></tr>
<tr><th colspan="2">STATISTICA DESCRITTIVA: ANALISI UNIVARIATA DI VARIABILI CATEGORICHE E NUMERICHE DISCRETE</th></tr>
<tr><td>Tipologie di analisi</td><td>

Analisi univariata → focus su singola variabile
Analisi bivariata → focus su una coppia di variabili (relazione tra variabili)
Analisi multivariata → focus su un gruppo di variabili (relazione di dipendenza/interdipendenza in un gruppo di variabili)

</td></tr>
</table>

Distribuzione di frequenza

Sia X una variabile statistica, e siano $x_1, x_2, x_i, \dots, x_k$ le k modalità distinte di X osservate negli n dati ($k < n$). Si chiama **distribuzione di frequenza** una tabella che permette di organizzare i dati a seconda delle **modalità**
Si definiscono, $\forall i = 1, \dots, k$:

1. **FREQUENZA ASSOLUTA** della i-esima modalità, che indichiamo con f_i, il **numero di volte per cui x_i si ripete**, o alla stessa maniera, il numero di unità nei dati che presentano $X = x_i$
2. **FREQUENZA RELATIVA** della i-esima modalità, che indichiamo con p_i la **proporzione di unità per cui x_i si ripete nei dati**:

$$p_i = Fr\{X = x_i\} = \frac{f_i}{n}$$

Chiameremo **FREQUENZA RELATIVA PERCENTUALE** della i-esima modalità: $p_i\% = 100 * p_i$

Frequenza cumulata

FREQUENZA RELATIVA CUMULATA DI X_i

$$F_i = Fr\{X \leq x_i\} = \sum_{j=1}^{1} p_j = p_1 + p_2 + \cdots + p_i$$

Ed è, quindi, il peso congiunto delle prime *i* modalità

Proprietà delle frequenze

FREQUENZE ASSOLUTE
- $0 \leq f_i \leq n$
- $\sum_{i=i}^{k} f_i = f_1 + \cdots + f_i + \cdots + f_k = n$

FREQUENZE RELATIVE
- $0 \leq p_i \leq 1$
- $\sum_{i=i}^{k} p_i = p_1 + \cdots + p_i + \cdots + p_k = 1$

FREQUANZE RELATIVE CUMULATE (SU SCALA ORDINALE O SUPERIORE)
- $0 \leq F_i \leq 1$
- $F_0 = 0, F_k = 1$
- $F_{i-1} \leq F_i$ → NON DECRESCENTI
- $p_i = F_i - F_{i-1}$ → dalle frequenze cumulate alle frequenze relative

Tabella di distribuzione delle frequenze

MODALITÀ X_i	FREQUENZA ASSOLUTA	FREQUENZA RELATIVA	FREQUENZA CUMULATA
[a, b)	f_1	p_1	p_1
[c, d)	f_2	p_2	p_1+p_2
...	...	...	$p_1+p_2+...$
[m, k]	f_k	p_k	$p_1+p_2+...+p_k = 1$

Rappresentazioni grafiche

DIAGRAMMA A TORTA

Per **variabili qualitative nominali** con al più 5 modalità distinte

Calcolo dei gradi degli angoli al centro
% : 100 = x : 360

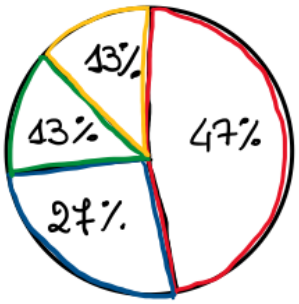
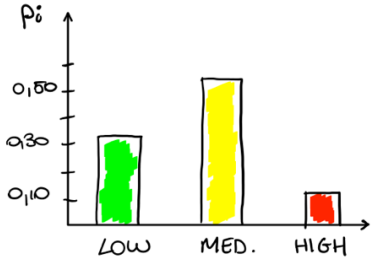
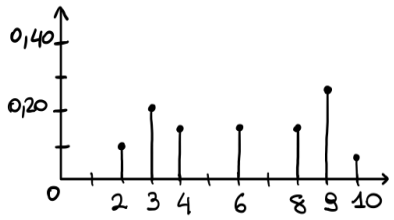
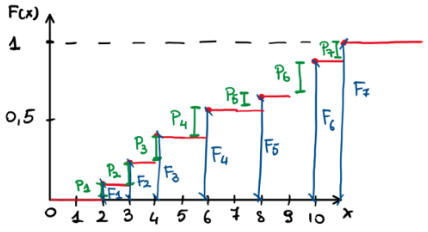


DIAGRAMMA A BARRE

Per **variabili qualitative ordinali**.
Sull'asse verticale sono riportate le frequenze; sull'asse orizzontale le modalità.
Le barre sono equidistante e non è orientato



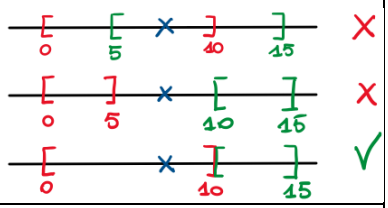
<table>
<tr><td rowspan="2"></td><td>DIAGRAMMA AD ASTE</td><td>Per variabile quantitativa discreta. L'asta ha altezza proporzionale alla frequenza relativa, osservabile sull'asse verticale. Sull'asse orizzontale, orientato, vengono riportate i valori numerici osservati e riportati con la loro posizione corretta. Le aste sono più sottili delle barre così da permettere di mantenere la distanza tra unità</td><td> 
</td></tr>
<tr><td>CURVA DELLE FREQUENZE CUMULATE</td><td>Permette di rappresentare la funzione cumulata di una variabile quantitativa discreta.
È un grafico non decrescente, continuo da destra, con limite inferiore pari a 0 e limite superiore di 1
Rappresenta la relativa funzione di ripartizione
La funzione è costante a tratti con punti di salto corrispondenti al valore della variabile e ampiezza di salto corrispondente alla frequenza relativa nel punto di salto</td><td> 
</td></tr>
<tr><td>Funzione della curva delle frequenze cumulate</td><td colspan="2">È la funzione di ripartizione di equazione generica:
$$F(x) = \begin{cases} 0 & x < x_1 \\ F_i & x_i \leq x < x_{i+1} \\ 1 & x \geq x_k \end{cases}$$</td></tr>
<tr><td>Calcolo delle frequenze cumulate dalla funzione di ripartizione</td><td colspan="2">• $\boldsymbol{Fr\{X \leq a\} = F(a)}$
• $Fr\{X < a\} \rightarrow$ somma le frequenze relative fino ad a escluso
• $Fr\{X > a\} = 1 - Fr\{X \leq a\} = 1 - F(a)$
• $Fr\{a < x \leq b\} = F(b) - F(a) = p(a) + \cdots + p(b)$
• $Fr\{a \leq x \leq b\} \rightarrow$ somme le frequenze relative da a fino a b incluso
• N.B.: $Fr\{a \leq x \leq b\} \neq Fr\{a < x \leq b\}$
• In ogni caso, posso sempre sommare le frequenze relative; attento nell'uso della differenza con la funzione delle frequenze cumulate!</td></tr>
<tr><td colspan="3">STATISTICA DESCRITTIVA: ANALISI UNIVARIATA DI VARIABILI DISCRETE CONTINUE</td></tr>
</table>

<table>
<tr><td>Classi e Frequenza</td><td>

Classi/Intervalli di valori → poiché le variabili rilevate sono molte, le modalità vengono raggruppate in range di valori, pur perdendo parte del dettaglio per favorire la leggibilità
Frequenza → a ciascun intervallo viene corrisposta la sua frequenza assoluta (= numero di unità per le quali il carattere in esame ha una modalità che cade nell'intervallo) e relativa (= la frequenza assoluta divisa per la dimensione del campione)

</td></tr>
<tr><td>Criteri di classificazione</td><td>
Esistono due criteri da rispettare per definire le classi:

DISGIUNTE → ogni valore deve appartenere al più ad una sola classe (no doppi conteggi)
ESAUSTIVE → ogni valore deve appartenere ad almeno una classe (no mancati conteggi)



</td></tr>
<tr><td>Notazione</td><td>

k → numero di classi
$x_0, x_1 \dots, x_i \dots, x_k$ → estremi delle classi

</td></tr>
<tr><td>Definizione degli estremi (ampiezza costante)</td><td>
Classi di uguale ampiezza

Scelta di k, numero di classi dell'intervallo
Ampiezza costante a tutte le classi:

$$w = \frac{max - min}{k} \rightarrow \text{arrotondata per eccesso}$$
Dove min → minimo valore osservato; max → massimo valore osservato

Determinazione degli estremi:

$x_0 = min$
$x_1 = min + w$
$x_i = min + i * w$
$x_k = x_{k-1} + w = min + k * w > max$

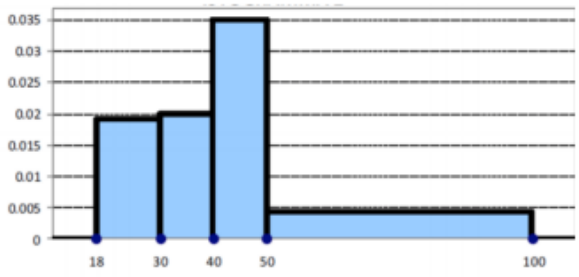
</td></tr>
<tr><td>Distribuzione di frequenza in classi</td><td>

f_i → frequenza assoluta della i-esima classe = numero di unità con valore X nell'intervallo $[x_{i-1}, x_i)$
$p_i = F\{X \in [x_{i-1}, x_i)\} = \frac{f_i}{n}$ → frequenza relativa della i-esima classe = proporzione di unità con valore X nell'intervallo $[x_{i-1}, x_i)$
$F_i = Fr\{X < x_i\} = \sum_{j=1}^{i} p_j$ → frequenza cumulata = peso congiunto delle prime i classi

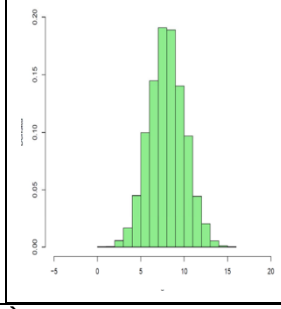
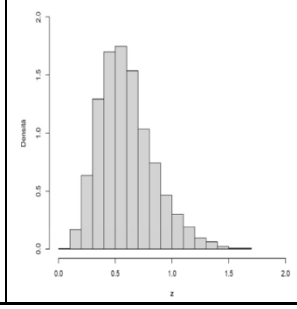
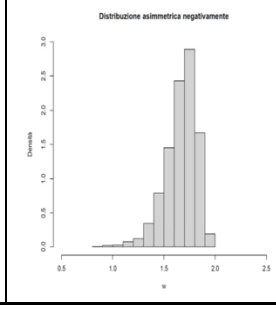
</td></tr>
<tr><td>Tabella</td><td>
<table>
<tr><td>Intervallo</td><td>f_i</td><td>p_i</td><td>c_i</td><td>w_i</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
</table>
</td></tr>
<tr><td>Istogramma</td><td>
Serie di rettangoli contigui, tali che l'i-esimo rettangolo ha:

Come base, l'i-esima classe $[x_{i-1}, x_i) = w_i$
Area pari alla frequenza relativa della i-esima classe p_i
Altezza pari alla densità di frequenza c_i

$$c_i = \frac{p_i}{w_i}$$
DENSITÀ DI FRQUENZA = frequenza relativa contenuta in ogni intervallo unitario nell'i-esima classe
</td></tr>
</table>

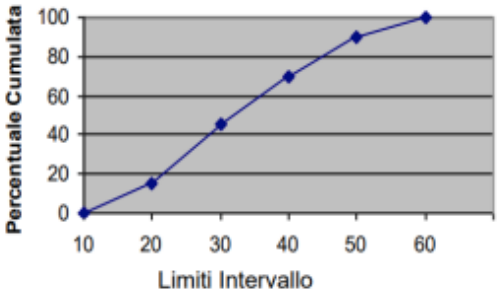
<table>
<tr><td></td><td> 

Come è possibile osservare, l'istogramma può avere anche intervalli di diversa ampiezza per aumentare il grado di dettaglio.
Il fatto che sull'asse verticale vengano usata la densità di frequenza, al posto delle frequenze assolute, risiede nella necessità di non "falsare" l'analisi dal punto di vista grafico
L'unico caso in cui la rappresentazione con frequenze assolute equivale a quella con densità di frequenza e il caso in cui le ampiezze delle classi siano uguali</td></tr>
<tr><td>Calcolo delle frequenze a partire dall'istogramma</td><td>L'area descritta dall'istogramma lungo un certo intervallo descrive la frequenza relativa per quell'intervallo</td></tr>
<tr><td>Forma della distribuzione</td><td>Gli istogrammi, a seconda della forma assunta dalla distribuzione, si classificano in:
Ricorda: la simmetria si valuta osservando la densità ed osservando globalmente tutti gli intervalli

<table>
<tr><th rowspan="2">DISTRIBUZIONE SIMMETRICA</th><th colspan="2">DISTRIBUZIONE ASIMMETRICA</th></tr>
<tr><th>POSITIVAMENTE</th><th>NEGATIVAMENTE</th></tr>
<tr><td>  </td><td>  </td><td> 
</td></tr>
</table></td></tr>
<tr><td>Curva delle frequenze cumulate</td><td>È la funzione che, per ogni x sull'asse, ha espressione

$F(x) = Fr\{X \leq x\} =$ Frequenza relativa di valori osservati inferiori ad x

Proprietà
• $0 \leq F(x) \leq 1$
• $F(x) = 0\ se\ x < x_0$
• $F(x) = 1\ se\ x \geq 1$
• $F(x) \leq F(y)\ se\ x < y \rightarrow$ non decrescente
• È una funzione lineare a tratti
• $Fr(a < X \leq b) = F(b) - F(a)$</td></tr>
<tr><td>Ogiva</td><td>Per variabili quantitative continue, è la rappresentazione della funzione cumulata
È un grafico non decrescente, continuo da destra con limite inferiore pari a 0 e limite superiore pari ad 1
La pendenza di ogni tratto corrisponde alla densità della classe</td></tr>
</table>



→ Tipologia di esercizi: può essere chiesto il valore di *x* per cui la frequenza cumulata sia pari ad un valore *a*. Per farlo occorre:

- Individuare l'intervallo in cui cade il valore di *x* richiesto
- Calcolare la retta, parte dell'ogiva, passante tra i due punti dell'intervallo: prima calcolare il coefficiente angolare; poi imporre il passaggio per uno dei due estremi; esplicitare l'equazione della retta
- Sostituire il valore *a* al posto della *y*
- Risolvere l'equazione in *x*

STATISTICA DESCRITTIVA: ANLISI BIVARIATA DI VARIABILI CATEGORICHE

Tabella di contingenza e frequenza assoluta congiunta

X \ Y	Y_1	…	y_j	…	y_c	**TOTALE**
x_1	f_{11}	…	f_{1j}	…	f_{c1}	R_1
…	…	…	…	…	…	…
x_i	f_{i1}	…	f_{ij}	…	f_{ic}	R_i
…	…	…	…	…	…	…
x_r	f_{1r}	…	f_{rj}	…	f_{rc}	R_r
TOTALE	C_1	…		…	C_c	n

Con:

- r = # modalità distinte di X
- c = # modalità distinte di Y
- i = indice per le modalità di X righe
- j = indice per le modalità di Y colonne

La tabella di contingenza è ottenuta:

- Classificando le unità statistiche sulla base della coppia di modalità per variabili X e Y che esibiscono (x_i, y_j)
- **DISTRIBUZIONE CONGIUNTA**. Contando il numero di unità che esibiscono ciascuna coppia, e determinando così per ogni i = 1, …, r e per ogni j = 1, …, c

  f_{ij} = **FREQUENZA ASSOLUTA CONGIUNTA** della coppia (x_i, y_j)

  Essa rappresenta il **numero di unità per cui congiuntamente X = x_i e Y = y_j**

  PROPRIETÀ
  - $0 \leq F_{ij} \leq n$
  - $\sum_{ij} f_{ij} = \sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij}$

- **DISTRIBUZIONI MARGINALI**. Nell'ultima colonna e nell'ultima riportiamo le distribuzioni marginali, così dette poiché a margine della tabella
  - R_i = **TOTALE DELLA i − ESIMA RIGA** = #unità che presentano X = x_i → **Frequenza Assoluta** della modalità x_i
  - C_j = **TOTALE DELLA j − ESIMA COLONNA** = #unità che presentano Y = Y_i → **Frequenza Assoluta** della modalità Y_i

  PROPRIETÀ
  - $0 \leq R_i = f_{i1} + f_{i2} + \cdots + f_{ic} \leq n$
  - $0 < C_j = c_{1j} + c_{2j} + \cdots + c_{ij} \leq n$

Quindi, dalla tabella di contingenza riesco a trovare le distribuzioni univariate di X e di Y

Frequenza relativa congiunta

Esiste anche la possibilità di calcolare le **FREQUENZE RELATIVE CONGIUNTE**: basta dividere le frequenze assolute congiunte per il totale *n*

$$p_{ij} = \frac{f_{iJ}}{n} = Fr\{X = x_i, Y = y_j\}$$

In questo caso, si avrà che *n = 1*

X \ Y	y_1	...	y_j	...	y_c	TOTALE
x_1	p_{11}	...	p_{1j}	...	p_{1c}	R_1/n
...	...	...	...	...	...	...
x_i	p_{i1}	...	$p_{ij} = f_{ij}/n$	...	p_{ic}	R_i/n
...	...	...	...	...	...	...
x_r	p_{r1}	...	p_{rj}	...	p_{rc}	R_r/n
TOTALE	C_1/n	...	C_j/n	...	C_c/n	1

Frequenza subordinata (o condizionata) e distribuzioni subordinate di frequenze

Ci permettono di calcolare le relative frequenze percentuali per ogni modalità della variabile:

1. **FREQUENZA SUBORDINATA DI Y DATA X**

$$\frac{f_{ij}}{R_i} = Fr\{Y = y_j | X = x_i\}$$

Ovvero, la proporzione di unità che presentano Y = y_j nel sottogruppo/categoria di quelle con X = x_i

X \ Y	y_1	...	y_j	...	y_c	TOTALE
x_1	f_{11}/R_1	...	f_{1j}/R_1	...	f_{1c}/R_1	1
...	...	...	...	...	...	1
x_i	f_{i1}/R_i	...	f_{ij}/R_i	...	f_{ic}/R_i	1
...	...	...	...	...	...	...
x_r	f_{r1}/R_r	...	f_{rj}/R_r	...	f_{rc}/R_r	1
TOTALE	−	−	−	−	−	1

Abbiamo eseguito una **normalizzazione per riga**: fatto 100 la singola x_i abbiamo analizzato il contributo delle singole y_i

2. FREQUENZA SUBORDINATA DI X DATA Y

$$\frac{f_{ij}}{C_j} = Fr\{X = x_i | Y = y_j\}$$

X \ Y	y_1	...	y_j	...	y_c	TOTALE
x_1	f_{11}/C_1	...	f_{1j}/C_j	...	f_{1c}/C_c	–
...	...	...	...	...	...	–
x_i	f_{i1}/C_1	...	f_{ij}/C_j	...	f_{ic}/C_c	–
...	...	...	...	...	...	–
x_r	f_{r1}/C_1	...	f_{rj}/C_j	...	f_{rc}/C_c	–
TOTALE	1	...	1	...	1	–

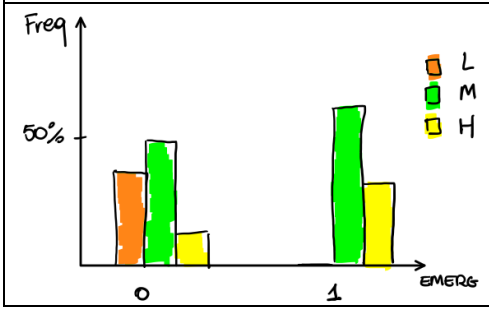
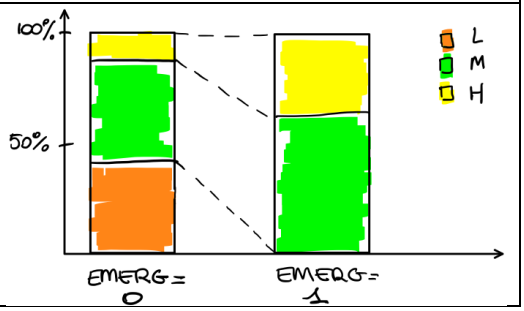
Abbiamo eseguito una **normalizzazione per colonna**: fatto 100 la singola y_i, abbiamo analizzato il contributo della singola x_i

Dipendenza ed indipendenza statistica

- Le variabili X ed Y si dicono **statisticamente indipendenti** se le distribuzioni di ***frequenza subordinata*** di Y data X (o, equivalentemente di X data Y) ***sono tutte uguali tra loro***
  - → Ciò significa che, con riguardo alla variabile Y, le unità *si comportano allo stesso modo*, all'interno di ogni sottogruppo o sottopopolazione definito da una modalità di X
- Se ciò non accade, ovvero se ***almeno due frequenze subordinate sono diverse tra loro***, allora X ed Y si dicono **statisticamente dipendenti**

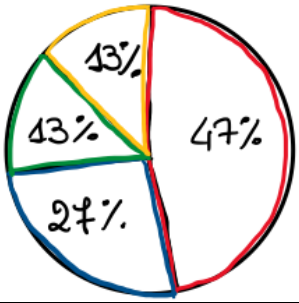
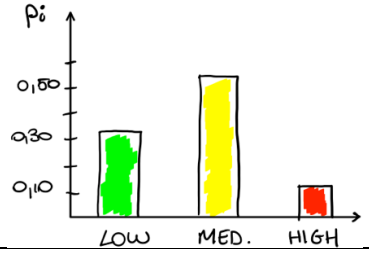
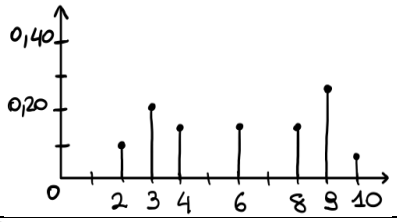
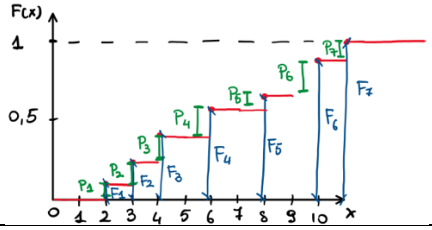
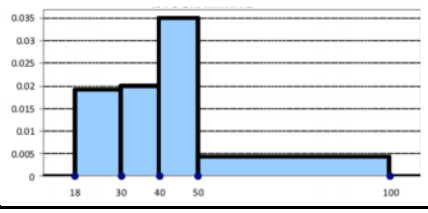
Diagramma a barre accostate e diagramma a barre sovrapposte

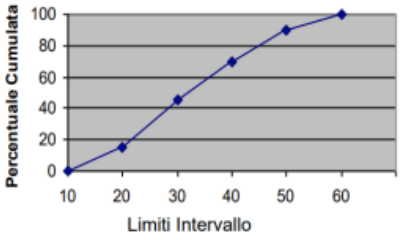
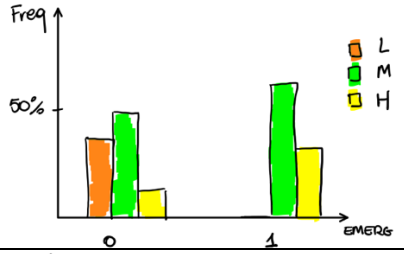
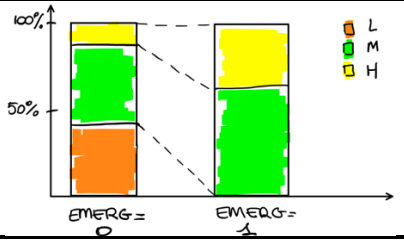
Per rappresentare graficamente un'analisi descrittiva bivariata possiamo utilizzare alternativamente:

Diagramma a barre accostate	Diagramma a barre sovrapposte
Asse X → modalità della variabile presa in considerazione come condizionale Asse Y → frequenza subordinata della variabile condizionata	Fatta cento la frequenza subordinata, viene impiegata una sola colonna ripartita nelle diverse modalità, espresse in percentuale
	

Paradosso di Simpson

Il **paradosso di Simpson** indica una situazione in cui ***una relazione tra due fenomeni appare modificata, o perfino invertita, a causa di variabili non presi in considerazione nell'analisi iniziale***

<table>
<tr><td></td><td>Il paradosso di Simpson permette dunque, a certe condizioni, il verificarsi di situazioni in cui il comportamento di sottogruppi è diverso dal comportamento complessivo.
Il paradosso di Simpson avviene quando non viene inclusa nell'analisi bivariata una variabile essenziale, creando risultati fuorvianti, dettati da una errata analisi delle frequenze.</td></tr>
<tr><td>Calcolo di frequenze</td><td>Dato un campione di n elementi, prendiamo due variabili A e B per le quali abbiamo costruito una tabella di contingenza. Prestiamo attenzione alle richieste fatte: Percentuale di campione che è A e B → $Fr\{A \cap B\} = \frac{A \cap B}{n}$ Percentuale di campione che è A tra quelli che sono B → $Fr\{A|B\}$ Supponendo che B < A, percentuale di campione che in totale ha fatto almeno A → $Fr\{X \leq A\} = \frac{A+B}{n}$</td></tr>
<tr><td>Grafici appropriati per tipologia di variabile</td><td>
<table>
<tr><th>Analisi</th><th>Variabile</th><th colspan="2">Grafico</th></tr>
<tr><td rowspan="4">Descrittiva univariata</td><td>Qualitativa nominale</td><td>Grafico a torta</td><td> 
</td></tr>
<tr><td>Qualitativa ordinale</td><td>Diagramma a barre</td><td> 
</td></tr>
<tr><td rowspan="2">Quantitativa discreta</td><td>Diagramma ad aste</td><td> 
</td></tr>
<tr><td>Curva delle frequenze cumulate</td><td> 
</td></tr>
<tr><td>Quantitativa continua</td><td>Istogramma</td><td> 
</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td rowspan="4"></td><td></td><td></td><td>Ogiva</td><td> 
</td></tr>
<tr><td rowspan="2">Descrittiva bivariata</td><td rowspan="2">Condizionate</td><td>Diagramma a barre accostate</td><td> 
</td></tr>
<tr><td>Diagramma a barre sovrapposte</td><td> 
</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td colspan="5">MISURA DI TENDENZA CENTRALE</td></tr>
<tr><td>Definizione</td><td colspan="4">Sono dei valori assunti dalle modalità che vengono calcolati con lo scopo di sintetizzare la distribuzione di una variabile mediante un’unica modalità. Sono indicatori il più possibile rappresentativi della tendenza della variabile. Le misure di tendenza centrale (e non centrale) sono modalità, non frequenze.</td></tr>
<tr><td>Media</td><td colspan="4">La media aritmetica rappresenta il baricentro della distribuzione: le distanze dalla media dei valori alla sua sinistra e di quelli alla sua destra si compensano, dando luogo ad una somma di tali distanze nulla.
Si utilizza solo per le variabili quantitative. È impossibile da calcolare per quelle qualitative, perché, essendo una somma, richiede necessariamente dei valori numerici.

<u>PER DATI GREZZI (o DISAGGREGATI)</u>
1) Media aritmetica semplice
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
Proprietà
• La media minimizza la funzione di errore statistico:
$$L(\bar{x}) < L(a) \rightarrow \sum_i (x_i - \bar{x})^2 < \sum_i (x_i - a)^2$$
• La somma delle deviazioni $(x_i - \bar{x})$ della media è nulla
$$\sum_{+} (x_i - \bar{x}) + \sum_{-} (x_i - \bar{x}) = 0$$
• Preserva il totale $\rightarrow n\bar{x} = \sum_i x_i$
• È interna $\rightarrow min < \bar{x} < max$
• Media e trasformazioni lineari</td></tr>
</table>

	$\forall\, y_i = a + bx_i \; con \; a \; e \; b \; arbitrari \rightarrow \; \bar{y} = a + b\bar{x}$ • Monotonia: se x_i aumenta/diminuisce anche $\bar{x}$ aumenta/diminuisce $$\frac{d\bar{x}}{dx_i} = \frac{1}{n} > 0$$ **2) Media aritmetica ponderata** Dati x_i valori, con w_i pesi, avremo che: $$\overline{x_w} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$ **<u>PER DISTRIBUZIONE DI FREQUENZE</u>** **1) Variabile numerica discreta** Con le **frequenze assolute**: $$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$ Con **frequenze relative**: $$\bar{x} = \frac{\sum_{i=1}^{k} p_i x_i}{\sum_{i=1}^{k} p_i} \rightarrow \bar{x} = \sum_{i=1}^{k} p_i x_i$$ **2) Variabile numerica in classi** Riunendo in classi, abbiamo *perso* i veri valori del dataset. Per procedere al calcolo della media dobbiamo supporre che l'elemento rappresentativo interno di ogni classe, che chiamiamo $\boldsymbol{m_i}$**, ovvero il punto medio della classe** $$\boldsymbol{m_i = \frac{x_{i-1} + x_i}{2}}$$ Da cui: $$\bar{x} = \frac{\sum_{i=1}^{k} f_i \boldsymbol{m_i}}{n} = \sum_{i=1}^{k} p_i \boldsymbol{m_i}$$ N.B.: *Proprietà* ***La media preserva il totale***: dati n valori, se al posto dei valori x_1, x_2, …, x_n, gli n individui presentassero tutti il medesimo valore $\bar{x}$, media della variabile X, allora il totale del carattere nel gruppo resterebbe invariato
Moda	È la **modalità alla quale è associata la frequenza più elevata**. La moda non tiene conto dell'aspetto numerico o meno delle modalità, ma si basa esclusivamente sulle loro frequenze Attenzione: • Va bene sia per **variabili qualitative** che **quantitative** • È la modalità che esibisce la frequenza più elevata, non la frequenza più elevata • Per distribuzioni in classi si impiega la **classe modale**, ovvero la classe con la **<u>densità di frequenza più elevata</u>**
Mediana	La **mediana** degli n (o N) valori osservati della variabile X $(x_1, x_2, \dots, x_i, \dots, x_n)$ è un qualunque valore *Me*, che **lascia alla sua sinistra e alla sua destra esattamente**

il 50% del totale delle osservazioni; in altri termini, *Me* deve soddisfare la disuguaglianza:

$$\boldsymbol{Fr\{X \leq Me\} \leq 0.5 \ \wedge Fr\{X \geq Me\} \geq 0.50}$$

Me è quel valore che *spacca in due la distribuzione*, ovvero 50% dei dati sono a destra e 50% dei dati sono a sinistra

Per dati disaggregati

Denotando con $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ la sequenza degli n valori osservati della variabile X, posti **in ordine crescente**, *Me* è l'osservazione di posto 0.5*(n+1):

$$Me = \begin{cases} x_{(0.5*(n+1))} & se\ n\ dispari \\ \dfrac{x_{(0.5*n)} + x_{(0.5*(n+1))}}{2} & se\ n\ pari \end{cases}$$

La logica sottostante è la seguente:

- Se n è dispari → scelgo l'osservazione centrale
- S n è pari → non c'è un'osservazione centrale, bensì un intervallo tra *due* osservazioni centrali → scelgo il valor medio di quell'intervallo

Per calcolarla procedo così:

- Calcolo la posizione della mediana nella sequenza ordinata:

$$Pos(Me) = 0.5(n+1)$$

- Determino che tale posizione è $x_{(Pos(Me))} = x_{(0.5(n+1))}$
- Se n è dispari, il valore con posizione *Pos(Me)* è proprio la mediana; Se n è pari, allora *Pos(Me)* sarà un valore decimale indicante che la mediana si trova tra due valori ed, in particolare, è il punto medio tra questi
- Nel caso in cui la posizione della Me ha decimali 0,50, allora è il punto medio tra i due valori

Per distribuzione di frequenza

- Variabile numerica discreta
  Me è il primo valore le cui frequenze cumulate, F_i, raggiungono o superano un valore pari a 0.5
  Qualora il valore sia esattamente 0,5 allora occorre eseguire una media tra il valore che registra 0,5 e quello immediatamente successivo
- Variabile numerica continua
  Me è il valore (o i valori) che soddisfano l'equazione $F(Me) = 0.50$; in particolare, se $F_{i-1} < 0.5 \leq F_i$ (le frequenze cumulate superano il valore 0.50 nella i-esima classe), *Me* può essere determinata risolvendo l'equazione:

$$0.5 = F_{i-1} + (Me - x_{i-1}) * c_i$$

$$Me = x_{i-1} + \frac{0.5 - F_{i-1}}{c_i}$$

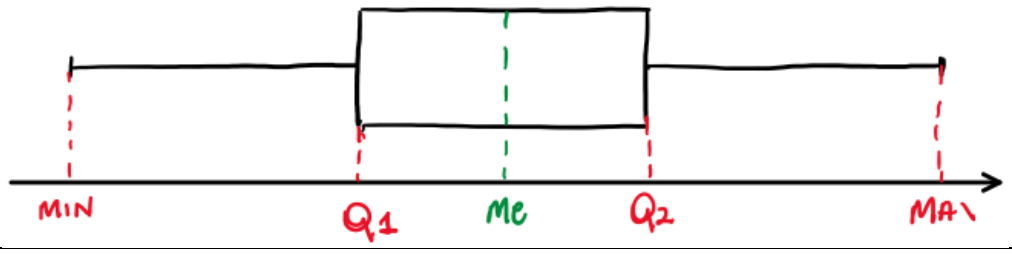
Quindi, prima di tutto, identifico la classe di cui fa parte la mediana (la classe che incorpora una frequenza cumulata superiore a 0.5); in seconda battuta applico la formula, ricordando che $x_{(i-1)}$ rappresenta l'estremo inferiore della classe scelta e F_{i-1} la frequenza cumulata della classe precedente

Opportuni indicatori di tendenza centrale	• **Media**: solo per variabili **quantitative discrete o continue** • **Moda**: sia per variabili **qualitative che quantitative**; tuttavia, è un indicatore eccessivamente povero di informazioni • **Mediana**: sia per variabili **quantitative che qualitative**
MISURE DI TENDENZA NON CENTRALE	
Definizione	Sono valori che **dividono la distribuzione in due parti non uguali e che non giacciono al centro di essa**; sono estensioni dell'idea di mediana
Quantile di ordine p	Fissato p, con $0 \le p \le 1$, un **quantile di ordine p** dei valori osservati $x_1, x_2, \ldots, x_n$ è un qualunque valore x_p, tale che: $$Fr\{X \le x_p\} \le p$$ $$Fr\{X \ge x_p\} \ge 1 - p$$ Quindi il **P-esimo quantile (o, analogamente, percentile) è quel valore che lascia alla sua sinistra (eventualmente includendo lo stesso valore) approssimativamente il P% di osservazioni** Nella sequenza degli *n* valori osservati della variabile *X*, posti in **ordine crescente**, x_p è l'osservazione di posto $p*(n+1)$, ovvero $$x_p = \begin{cases} x_{(h)} & con\ h\ l'intero\ più\ vicino\ a\ p^*(n+1) \\ \frac{x_{(l)} + x_{(l+1)}}{2} & se\ p*(n+1)\ ha\ decimali\ 0.5 \\ & l\ è\ p(n+1)\ arrotondato\ per\ difetto \end{cases}$$ Procedo così: • Calcolo da posizione del quartile → $Pos(x_p) = p*(n+1)$ • Arrotondo all'intero più vicini (per eccesso o per difetto); ma attenzione al caso in cui siamo esattamente a decimali 0.5 • Individuo x_p con la formula usata sopra e facendo attenzione ai decimali **Quantile di ordine p per distribuzioni di frequenza** • <u>Variabile numerica discreta</u> x_p è il primo valore le cui frequenze cumulate F_i raggiungono o superano il valore p • <u>Variabile numerica in classi</u> x_p è il valore (o i valori) che soddisfano l'equazione $F(x_p) = p$; in particolare, se $F_{i-1} < p \le F_i$ (ovvero le frequenze cumulate superano il valore p nell'i-esima classe), x_p può essere determinato così: $$p = F_{i-1} + (x_p - x_{i-1}) * c_i$$ $$x_p = x_{i-1} + \frac{p - F_{i-1}}{c_i}$$ Nel caso in cui la classe riporti esattamente l'ordine p del percentile come frequenza cumulata, allora l'estremo superiore è esattamente il percentile ricercato **Quartili fondamentali** Sono quantili particolari:

<table>
<tr><td></td><td>
1. Primo quartile o 25esimo percentile (Q_1) → p = 0.25; Pos(Q_1)=0.25(n+1)
2. Secondo quartile o 50esimo percentile = Mediana (Q_2=Me) → p = 0.50; Pos(Q_1)=0.50(n+1)
3. Terzo quartile o 75esimo percentile (Q_3) → p = 0.75; Pos(Q_1)=0.75(n+1)

Osservazione:
Nota che tutte le relazioni possono essere espresse in funzione del primo quartile:

$$Pos(Q_1) = 0{,}25(n+1) \rightarrow Pos(Me) = 2 * 0{,}25(n+1) = 2 * Pos(Q_1)$$
</td></tr>
<tr><td>Percentili</td><td>
I percentili sono 99 punti che dividono la distribuzione in 100 parti con la stessa frequenza (pari a 0.01 o 1%)
• L'$n-esimo$ percentile è il quantile di ordine $p = n\%$
• Si calcolano allo stesso modo dei quartili e i fondamentali sono esattamente quelli già citati
• Ha senso calcolarli per variabili quantitative (discrete e continue) e per le qualitative ordinali; non ha senso calcolarli per variabili qualitative nominali

Il significato del quantile/percentile è il seguente:
Ordinando in senso crescente un dataset, trovare i migliori x% risultati, vuol dire calcolare il (1-x)-esimo percentile, ovvero il quantile di ordine p = (1-x)
Quindi se viene fornita una scala di valori crescenti, e viene richiesto di trovare il 70% dei valori migliori, allora dobbiamo calcolare il quartile di ordine 0,30 (cioè, il 30esimo percentile): infatti, tale valore identifica il valore a sinistra del quale si trovano il 30% dei valori totali (quelli più bassi, i peggiori) e, riflessivamente, a destra del quale si trova il 70% dei valori (quelli più alti, i migliori)
Una seconda interpretazione ci permette di dire che il valore corrispondente al 30esimo percentile è quel valore che, nel 70% dei casi, viene superato.
</td></tr>
<tr><td>Cinque numeri di sintesi</td><td>
<table>
<tr><th>min</th><th>Q_1</th><th>Me</th><th>Q_3</th><th>max</th></tr>
<tr><td>Valora più piccolo rilevato per la variabile campionaria</td><td>Primo quartile $p = 0{,}25$</td><td>Mediana $p = 0{,}50$</td><td>Terzo quartile $p = 0{,}75$</td><td>Valora più grande rilevato per la variabile campionaria</td></tr>
</table>
</td></tr>
<tr><td>Box Plot</td><td>
Il box plot o diagramma a scatola e baffi è la rappresentazione grafica dei cinque numeri di sintesi


</td></tr>
</table>

Forma della distribuzione

I cinque numeri di sintesi ed il Box Plot consentono una valutazione della simmetria o asimmetria della distribuzione dei dati, mediante i confronti di:

Simmetria	Numeri di sintesi	Grafico
Distribuzione approssimativamente simmetriche	$\begin{cases} Q_1 - min \approx max - Q_3 \\ Me - Q_1 \approx Q_3 - Me \\ Me \approx \bar{x} \end{cases}$	
Distribuzione positivamente asimmetrica (obliqua a destra)	$\begin{cases} Q_1 - min < \max - Q_3 \\ Me - Q_1 < Q_3 - Me \\ Me < \bar{x} \end{cases}$	
Distribuzione negativamente asimmetrica (obliqua a sinistra)	$\begin{cases} Q_1 - min > \max - Q_3 \\ Me - Q_1 > Q_3 - Me \\ Me > \bar{x} \end{cases}$	

Outlier e loro identificazione

<u>OUTLIER</u>: osservazione per cui il valore di X si discosta notevolmente dal resto dei dati, ovvero dal centro della distribuzione, influenzando significativamente l'analisi

Cause:
- Errori di rilevazione
- Errori di trascrizione
- Manifestazione raramente osservata del fenomeno

Identificazione:
Per prima cosa identifichiamo gli estremi oltre i quali cadono gli outlier

$$\bar{L} = Q_3 + 1{,}5(Q_3 - Q_1) = Q_1 - 1{,}5IQR$$

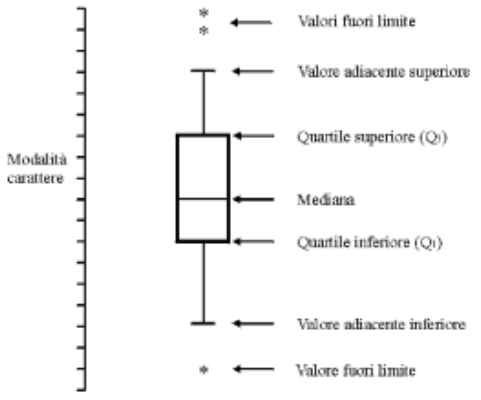
$$\underline{L} = Q_1 - 1{,}5(Q_3 - Q_1) = Q_3 - 1{,}5IQR$$

Dove, **IQR = Inter Quartile Range**
Quindi, **tutti i valori x_i che ricadono al di fuori dell'intervallo $[\bar{L}, \underline{L}]$ sono considerati outlier**

Robustezza

Una misura di sintesi si dice **robusta** se ***non risente in modo significativo dei cambiamenti in una porzione limitata di valori del dataset***

- La MODA e la MEDIANA sono misure ROBUSTE
- La MEDIA NON è ROBUSTA



MISURE DI VARIABILITÀ E DISPERSIONE PER VARIABILI QUANTITATIVE UNIVARIATE

Campo di variazione (Range)

È l'ampiezza dell'intervallo che contiene il 100% dei dati:

$$R = max - min$$

- È una misura **non sensibile** in quanto, al variare della variabilità, il range può non cambiare

<table>
<tr><td></td><td>

- È una misura **non robusta** poiché funzione unicamente del max e del min

</td></tr>
<tr><td>Differenza interquartile (IQR)</td><td>

È l'ampiezza dell'intervallo che contiene il 50% centrale dei dati:

$$IQR = Q_3 - Q_1$$

- IQR è una misura **robusta** poiché non risente eccessivamente della dispersione

</td></tr>
<tr><td>Varianza</td><td>

Sia X una variabile quantitativa univariata e si conosca la sua media e il suo scarto quadratico medio. Si chiama **varianza** degli N (n) valori osservati la **somma delle differenze, al quadrato, tra ciascuna osservazione e la media della popolazione (campione), divisa per la dimensione della popolazione N (ampiezza del campione *n*)**
Essa ci permette di misurare la **variabilità dei dati** ma presenta una **<u>unità di misura elevata al quadrato</u>** rispetto a quella originale.

Per **dati grezzi**

<table>
<tr><th>POPOLAZIONE</th><th>CAMPIONE
(VARIANZA CAMPIONARIA)</th></tr>
<tr><td>

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

<u>Formula ridotta</u> (Devianza)

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

</td><td>

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

<u>Formula ridotta</u> (Devianza)

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

</td></tr>
</table>

Per **distribuzioni di frequenza**

<table>
<tr><th>POPOLAZIONE</th><th>CAMPIONE
(VARIANZA CAMPIONARIA)</th></tr>
<tr><td>

Distribuzione discreta

$$\sigma^2 = \frac{1}{N}\sum_{I=1}^{k} f_I(x_I - \mu)^2 = \sum_{i=1}^{k} p_i(x_i - \mu)^2$$

<u>Formula ridotta</u>

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i} - \mu^2 = \sum_{i=1}^{k} p_i x_i^2 - \mu^2$$

Distribuzione continua

$$\sigma^2 = \frac{\sum_{I=1}^{k} f_I(m_I - \mu)^2}{N} = \sum_{i=1}^{k} p_i(m_i - \mu)^2$$

</td><td>

Distribuzione discreta

$$s^2 = \frac{1}{n-1}\sum_{I=1}^{k} f_I(x_I - \overline{x})^2 = \frac{n}{n-1}\sum_{i=1}^{k} p_i(x_i - \overline{x})^2$$

<u>Formula ridotta</u>

$$s^2 = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} f_i - 1}\left[\frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i} - \overline{x}^2\right] = \frac{n}{n-1}\sum_{i=1}^{k} p_i x_i^2 - \overline{x}^2$$

Distribuzione continua

$$s^2 = \frac{\sum_{I=1}^{k} f_i(m_i - \overline{x})^2}{n-1}$$

</td></tr>
</table>

</td></tr>
</table>

<table>
<tr>
<td></td>
<td>
<table>
<tr>
<td>Formula ridotta
$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i m_i^2}{\sum_{i=1}^{k} f_i} - \mu^2$$
$$= \sum_{i=1}^{k} p_i m_i^2 - \mu^2$$</td>
<td>$$= \frac{n}{n-1} \sum_{i=1}^{k} p_i (m_i - \overline{x})^2$$
Formula ridotta
$$s^2 = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} f_i - 1} \left[\frac{\sum_{i=1}^{k} f_i m_i^2}{\sum_{i=1}^{k} f_i} - \overline{x}^2 \right]$$
$$= \frac{n}{n-1} \sum_{i=1}^{k} p_i m_i^2 - \overline{x}^2$$</td>
</tr>
</table>
La ***varianza è una misura di dispersione della media aritmetica***; è espressa in **un'unità di misura** che risulta il **quadrato dell'unità di misura della variabile in esame**.
Essa è **tanto più grande quanto più è alta la dispersione di ciascun valore della media aritmetica**
</td>
</tr>
<tr>
<td>Scarto quadratico medio o deviazione standard</td>
<td>Sia X una variabile quantitativa univariata e si conosca la sua media e il suo scarto quadratico medio. Si chiama scarto quadratico medio o deviazione standard la radice quadrata della varianza, che permette di valutare dispersione delle singole osservazioni rispetto alla media, con <u>unità di misura uguale a quella osservata</u>
<table>
<tr><th>POPOLAZIONE</th><th>CAMPIONE</th></tr>
<tr>
<td>$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{I=1}^{N} (x_i - \mu)^2}{N}}$$</td>
<td>$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$</td>
</tr>
</table>
</td>
</tr>
<tr>
<td>Proprietà della varianza e dello scarto quadratico medio</td>
<td>

$\sigma^2 \geq 0$ e $\sigma \geq 0$
$\sigma^2 = 0 \Leftrightarrow x_i = \mu, \forall i$
Varianza e trasformazioni lineari:
Se per ogni i, $y_i = a + bx_i$, con a e b valori reali arbitrari, allora:
$$\sigma_Y^2 = b^2 * \sigma_X^2$$
$$\sigma_Y = |b| \sigma_x$$
La varianza/scarto quadratico medio è una misura di dispersione; a trasla semplicemente la distribuzione → non modifica i due valori
Sono due misure NON ROBUSTE

</td>
</tr>
</table>

Coefficiente di Variazione

Sia X una variabile quantitativa univariata e si conosca la sua media e il suo scarto quadratico medio. Si chiama **coefficiente di variazione l'indice che misura la variabilità di una variabile, relativamente alla media**

POPOLAZIONE	CAMPIONE
$CV = \frac{\sigma}{\lvert\mu\rvert}$	$CV = \frac{s}{\lvert\bar{x}\rvert}$

Indica la **fluttuazione (percentuale) di una variabile** rispetto alla media; permette di apprezzare in termini relativi, rispetto alla media, la variazione rispetto al centro della distribuzione
Questa misura:

- È appropriata per **effettuare confronti di variabilità tra gruppi o variabili differenti**
- μ deve essere diverso da 0; in generale, valori di $\mu \to 0$ provocano oscillazioni molto forti
- CV è un **numero puro**, non ha unità di misura
- CV **non è** una misura **robusta**

N.B.: è l'**indice opportuno** per rispondere alla domanda: **qual è la variabile con variabilità maggiore?**

Disuguaglianza di Chebychev

Siano μ e σ la media e la deviazione standard di una popolazione qualsiasi. Fissato $k > 1$, allora:

$$Fr\{\mu - k\,\sigma \leq X \leq \mu + k\,\sigma\} \geq 1 - \frac{1}{k^2}$$

La disuguaglianza ci permette di **calcolare la distribuzione approssimata (ma sottostimata) dei dati in uno specifico intervallo, avendo a disposizione unicamente la media e la varianza**. In particolare, permette di definire:

- La frequenza **almeno** osservabile in un **intervallo** di valori
- La frequenza **al più** osservabile sulle **code**

Per ricavare k basta, avendo media, deviazione standard e gli estremi dell'**intervallo [a,b]**

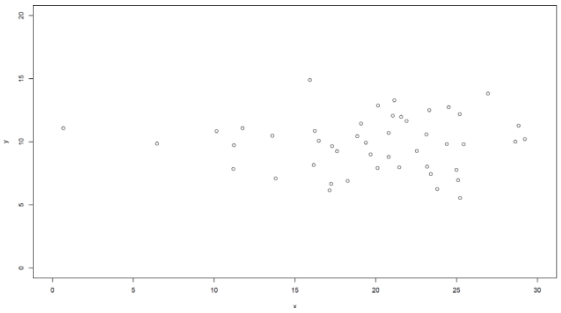
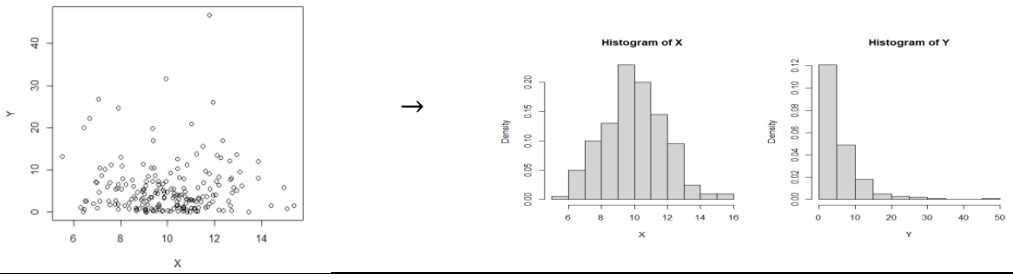
$$k = \frac{\mu - a}{\sigma}$$

N.B.: l'intervallo usato da Chebychev è **centrato nella media** ovvero, l'intervallo dato, supponiamo [a,b], deve essere tale che $\frac{a+b}{2} = \bar{x}$.
Se **l'intervallo non fosse centrato nella media** allora avremmo dovuto cercare **l'intervallo più grande centrato nella media**, **contenuto nell'intervallo dato**: saremo certi che **la frequenza dell'intervallo dato è maggiore (o al massimo uguale) a quello centrato**

N.B.: attento a controllare tutto quello che si ha! Potrebbe capitare che l'intervallo dato corrisponda a specifici punti forniti (es.: differenza tra Max e Q_1*): se ciò accade abbiamo una misura esatta e non approssimata (*0,75 = 75%)

Attenzione: in generale **non è possibile trovare un limite inferiore né un limite superiore per la probabilità usando il teorema dei Chebychev**

Regola empirica dei tre σ
Se la distribuzione della variabile X è di **forma campanulare e simmetrica**, allora:

	$Fr\{\mu - \sigma \leq X \leq \mu + \sigma\} \approx 0.68$ $Fr\{\mu - 2\sigma \leq X \leq \mu + 2\sigma\} \approx 0.95$ $Fr\{\mu - 3\sigma \leq X \leq \mu + 3\sigma\} \approx 0.9973$ **N.B.: prima di applicare Chebychev controlla se è possibile usare la legge empirica** 1. Verifica se la distribuzione è simmetrica, controllando la distanza tra i cinque punti di sintesi oppure la relazione tra media e mediana 2. Controllo se è campanulare, ovvero se la densità si concentra principalmente nel centro; per farlo, osserva se la distanza tra Min e Q_1 è maggiore alla distanza tra Q_1 e Me (oppure la distanza tra Q_3 e Max maggiore di Me e Q_3)
MISURE DI RELAZIONE LINEARE PER VARIBILE QUANTITATIVE BIVARIATE	
Forma dei dati grezzi	Per le analisi bivariate impiegheremo due variabili numeriche X e Y raggruppate in n coppie: (x_n, y_n)
Diagramma a dispersione	Il diagramma a dispersione, detto anche **scatterplot**, rappresenta la distribuzione delle coppie di variabili in un piano cartesiano di assi X e Y L'i-esima unità del dataset è rappresentata da **un punto di coordinate (x_i, y_i) nel piano cartesiano di assi X e Y**  **Simmetria e correlazione con istogramma** Posso ricondurre lo scatterplot a due istogrammi (per X e per Y) tracciando le proiezioni dei punti sugli assi e osservando la simmetria 
Covarianza	Siano X ed Y due variabili quantitative bivariate e sia N la popolosità di una popolazione mentre *n* sia l'ampiezza di un campione. Si chiama **covarianza** l'**indice** di **misura dell'<u>esistenza</u> della relazione lineare** tra le due variabili X ed Y. Essa si calcola come segue: **POPOLAZIONE** | **CAMPIONE**

$\sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$	$s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
$\sigma_{XY} = Cov(X,Y) = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \mu_X \mu_Y$	$s_{XY} = \frac{n}{n-1}\left[\frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x} \cdot \bar{y}\right]$

In particolare:

- **Cov > 0** → x e y tendono a muoversi nella stessa direzione (**dipendenza lineare positiva o diretta**)
- **Cov < 0** → x e y tendono a muoversi in direzioni opposte (**dipendenza lineare negativa o inversa**)
- **Cov = 0** → x e y sono non correlate, non c'è dipendenza lineare (**assenza di relazione lineare** oppure **relazione forte ma non monotona**)

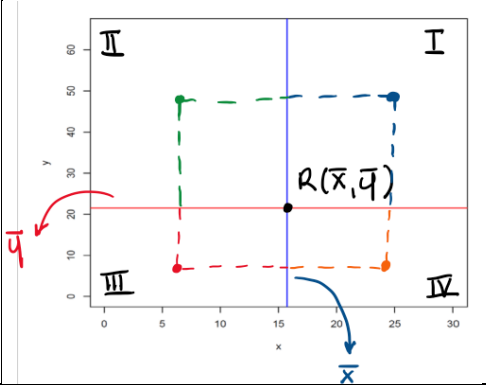
Osservazioni:

- Difficile valutare la covarianza in termini assoluti…
- Covarianza e trasformazioni lineari
  Se $x_i^* = a + bx_i$ e $y_i^* = c + dy_i$ con a, b, c e d valori reali arbitrari, allora:

$$\sigma_{X^*Y^*} = bd * \sigma_{XY}$$

- Si osserva che nella trasformazione:
  - **Le variabili additive non hanno effetto**
  - **Le variabili moltiplicative hanno effetto**
- La covarianza è influenzata dalla dispersione individuale (scarto quadratico) di ciascuna variabile. Infatti, se la variabilità σ_x è alta, allora le deviazioni $x_i - \bar{x}$; conseguentemente tenderanno ad essere elevati i prodotti incrociati $(x_i - \bar{x})(y_i - \bar{y})$ e, quindi, la covarianza; quindi, basta anche solo la variabilità elevata di una variabile, inflaziona l'intero indice…
- La covarianza non è una misura robusta

Interpretazione della covarianza



QUADRANTE	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
I	+	+ =>	⊕
II	−	+ =>	⊖
III	−	− =>	⊕
IV	+	− =>	⊖

Osservando i valori nei quadranti posso stimare il segno della Covarianza, in quanto posso determinare il segno di $(x_i - \bar{x})(y_i - \bar{y})$

Coefficiente di correlazione lineare (di Pearson)

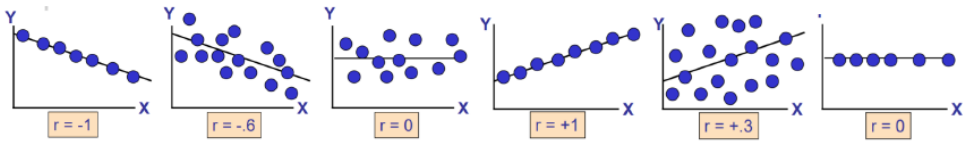
Il valore della covarianza dipende dall'unità di misura e quindi, non si tratta di un indice adeguato a valutare l'intensità della relazione lineare tra due variabili.

Siano X ed Y due variabili quantitative bivariate e si conoscano la covarianza e gli scarti quadratici medi delle due variabili. Si chiama **coefficiente di correlazione lineare di Pearson** un **indice che misura l'esistenza e l'intensità della relazione lineare** tra le due variabili X ed Y. Esso si calcola nel seguente modo:

POPOLAZIONE	CAMPIONE
$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$	$r_{XY} = \frac{s_{XY}}{s_X s_Y}$

Proprietà

- $-1 \leq \rho_{XY} \leq 1$
- $\boldsymbol{\rho_{XY} > 0 \rightarrow}$ **correlazione lineare positiva**
- $\boldsymbol{\rho_{XY} < 0 \rightarrow}$ **correlazione lineare negativa**
- $\boldsymbol{|\rho_{XY}| = 1 \rightarrow}$ **correlazione lineare esatta**
  - Ovvero $y_i = a + bx_i$ con b > 0 se $\rho_{XY} = 1$; con b < 0 se $\rho_{XY} = -1$
- $\boldsymbol{\rho_{XY} \approx 0 \rightarrow}$ **assenza di correlazione lineare**, **ma non di qualsiasi altro tipo di relazione** (ad esempio, potrebbe esserci una dipendenza quadratica)



Teorema

Indipendenza statistica ⇒ NON correlazione lineare

Regola pratica
Si può assumere l'esistenza di una relazione lineare se:

$$|r| = \frac{2}{\sqrt{n}}$$

ELEMENTI DI PROBABILITÀ E DISTRIBUZIONI NOTEVOLI

Funzioni aleatorie

- **Funzione di probabilità**: funzione che associa ad ogni numero reale la probabilità con cui una variabile assume quel numero reale stesso

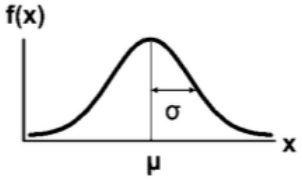
$$\boldsymbol{p(x) = p(X = x)}$$

  Gode delle seguenti proprietà:
  - $p(x) \geq 0$
  - $\sum_{i=1}^{n} p(x_n) = 1$

- **Funzione di ripartizione**: funzione che associa ad ogni numero reale la probabilità con cui una variabile assume tutti i valori precedenti a tale numero, esso compreso

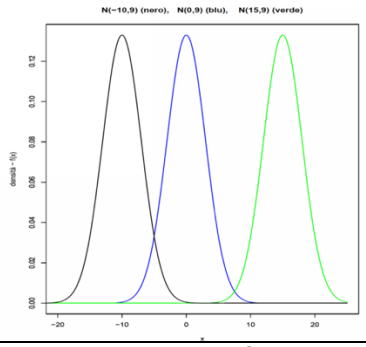
$$\boldsymbol{F(x) = P(X \leq x) = \sum_{u \leq x} p(u)}$$

	Gode delle seguenti proprietà: ○ $0 \leq F(x) \leq 1$ ○ $F(x)$ non decrescente ○ $F(x)$ costante a tratti (discontinua)
Valore atteso	Il valore atteso E(x) è la **media dei valori ed è il baricentro** (punto centrale) **della distribuzione della variabile aleatoria** Si calcola come la media di X ponderata per le sue modalità: $$E[x] = \mu = \sum xp(x)$$ **Interpretazione** Il valore atteso E[x] è approssimativamente la media aritmetica dei risultati che si ottengono replicando *all'infinito* l'esperimento
Varianza e scarto quadratico medio	Secondo la definizione la **varianza** è il valore atteso della differenza, al quadrato, tra X e la media: $$Var(x) = \sigma^2 = E[(x-\mu)^2] = \sum (x-\mu)^2 p(x)$$ Ma applichiamo la **formula per il calcolo** $$\boldsymbol{Var(x) = \sigma^2 = E[x^2] - \mu^2 = \sum [x^2 p(x)] - \mu^2}$$ Infine, lo **scarto quadratico medio** $$\boldsymbol{\sigma = \sqrt{Var(x)}}$$
Covarianza	La **covarianza** è la differenza tra il valore atteso del prodotto delle variabili aleatorie ed il prodotto dei valori attesi: $$Cov(X,Y) = E[XY] - E[X]E[Y]$$
Trasformazioni lineari di variabili aleatorie	**Trasformazione lineare semplice** Consideriamo una variabile aleatoria X e la sua trasformazione lineare $$\boldsymbol{Y = a + bX}$$ Allora: • $E(Y) = a + bE(x)$ • $Var(Y) = b^2 Var(x)$ • $\delta(Y) = \lvert b \rvert \delta(x)$ **Trasformazione di combinazione lineare** Consideriamo una variabile aleatoria X e Y e la sua combinazione lineare $$\boldsymbol{W = aX + bY}$$ Risulta: ○ $E(W) = aE(X) + bE(Y)$ ○ $Var(W) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y)$ N.B.: Se avessimo $$\boldsymbol{W = aX - bY}$$

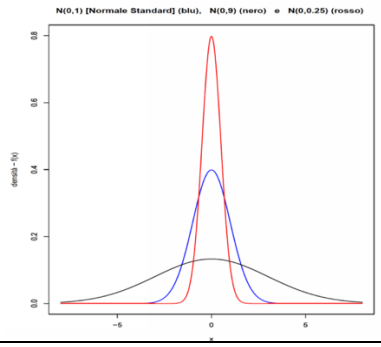
	Risulterebbe ○ $E(W) = aE(X) - bE(Y)$ ○ $Var(W) = a^2Var(X) + b^2Var(Y) - 2abCov(X,Y)$ Ricorda che: $$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$ Se X e Y fossero **indipendenti**, cioè $Cov(X,Y) = 0$, risulterebbe: $$Var(W) = a^2Var(X) + b^2Var(Y)$$
Standardizzazione	È la trasformazione lineare di una variabile aleatoria che si ottiene sottraendo μ e dividendo per δ $$Z = \frac{X - \mu}{\delta}$$ Essa trasforma una qualsiasi variabile aleatoria X in una nuova variabile Z, detta standardizzata, tale che E(Z) = 0 e Var(Z) = 1
Distribuzione binomiale (variabili discrete)	Riporta qual è la **probabilità di avere x successi su n prove** $$X \sim Bin(n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,2,3,\dots,n \\ 0 & altrove \end{cases}$$ Con $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ **Proprietà**: • $E(x) = np$ • $Var(x) = np(1-p)$
Distribuzione bernoulliana (variabili discrete)	Una variabile si distribuisce secondo una bernoulliana allorquando si fronteggia un **esperimento dicotomico**, ovvero un esperimento che può avere **due soli esiti possibili** (**successo, fallimento**). È un particolare caso di una distribuzione **binomiale con n = 1** $$X \sim Bern(p) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$$ **Proprietà** • $E(x) = p$ • $Var(x) = p(1-p)$
Distribuzione normale o gaussiana (variabili continue)	Una variabile aleatori si distribuisce come una normale/gaussiana di media μ e varianza δ^2 quando: $$X \sim N(\mu, \sigma^2)$$  • La distribuzione ha andamento **simmetrico e campanulare** • Al centro della simmetria la distribuzione è $\mu = E(x)$ • Al crescere di δ^2 la densità si appiattisce e le code divengono più pesanti: questo perché, la varianza indica la dispersione della distribuzione rispetto al centro; tanto più la varianza è grande tanto più i valori saranno dispersi • $Z \sim N(0,1)$ si chiama **normale standard**

- $X \sim N(\mu, \delta^2)$ e $Y = a + bX$ allora $Y \sim (a + b\mu, b^2\delta^2)$
- $\boldsymbol{Media = Mediana = Moda}$

Parametri e forma della distribuzione



Le curve hanno stessa δ^2 ma risulta: $\mu_{nera} < \mu_{blu} < \mu_{verde}$



Le curve hanno stessa μ ma risulta: $\delta^2_{rossa} < \delta^2_{blu} < \delta^2_{nera}$

Calcolo della probabilità cumulata e standardizzazione

L'area sottesa dalla curva rappresenta la probabilità cumulata; per eseguirne il calcolo si procede così:

1. Definisco la curva per il problema in oggetto
2. Standardizzo traducendo X in $Z = \frac{X-\mu}{\sigma}$
3. Riscrivo la probabilità assegnata:

$$P(X < a) = P\left(Z < \frac{a-\mu}{\sigma}\right)$$

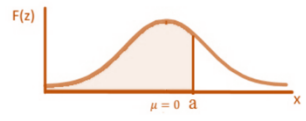
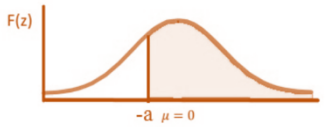
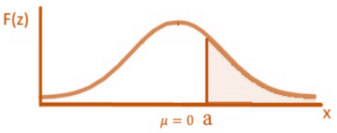
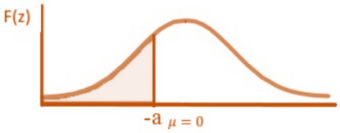
4. Uso le tavole per ritrovare il risultato

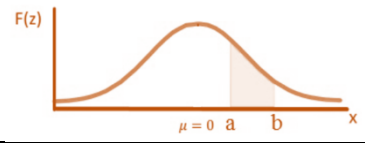
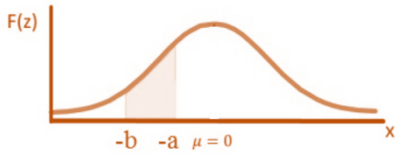
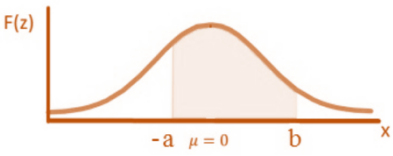
Simmetria e campanularità

- Poiché la distribuzione è simmetrica la media è uguale alla mediana e alla media interquartile

$$\mu = Me = \frac{Q_3 + Q_1}{2}$$

Casi di calcolo di probabilità con la normale

1	$P(Z < a) = F_z(a)$	
2	$P(Z > -a) = P(Z < a) = F_z(a)$	
3	$P(Z > a) = 1 - F_z(a)$	
4	$P(Z < -a) = P(Z > a) = 1 - F_z(a)$	

	5	$P(a < Z < b) = F_z(b) - F_z(a)$	
	6	$P(-b < Z < -a) = F_z(b) - F_z(a)$	
	7	$P(-a < Z < b) = F_z(b) + F_z(a) - 1$	
Chebichev e normale	Se i dati forniti applicando Chebichev, su un intervallo centrato sulla media, fornita con la deviazione standard, suggerissero una forma **simmetrica** e **campanulare**, allora è possibile approssimare più precisamente la frequenza **normalizzando** e ricorrendo alle tavole di sintesi numerica		
Approssimazione della distribuzione binomiale con la normale	La **distribuzione binomiale** si può **approssimare alla distribuzione normale** solo quando **n è abbastanza grande** ovvero quando $\boldsymbol{np(1-p) > 9}$ In questo caso: • $E(x) = np$ • $Var(x) = np(1-p)$ E la distribuzione binomiale $\boldsymbol{X \sim Bin(n,p)}$ può essere trasformata **approssimativamente** in una normale: $$\boldsymbol{X \approx N(np, np(1-p))}$$ Per cui, la standardizzazione diviene: $$Z = \frac{X - E(x)}{Var(x)} = \frac{X - np}{\sqrt{np(1-p)}}$$		

STATISTICA INFERENZIALE	
Introduzione	La statistica inferenziale ci permette di trarre conclusioni/prendere decisioni riguardanti una popolazione sulla base dei risultati di un singolo campione
Metodi di campionamento probabilistici	**Campionamento casuale semplice (CCS)**: ogni unità della popolazione ha la stessa probabilità di **essere** incluso nel campione; si procede così: 1. Si selezione un'unità della popolazione 2. Si rileva il valore d'interesse 3. Si esegue una scelta: a. Si reinserisce l'unità (**campionamento con reimmissione**) b. Non si reinserisce l'unità (**campionamento senza reimmissione**) 4. Si ripetono le fasi finché non si selezionano *n* unità Proprietà della scelta • **CCS con reimmissione** → le probabilità osservata da ciascuno dei due sono le medesime e, in particolare, sono uguali a quelle che caratterizzano la popolazione → **V.A. INDIPENDENTI E IDENTICAMENTE DISTRIBUITE:** $\boldsymbol{X_1, X_2, \ldots \sim^{iid} f_x}$ → $Cov(X,Y) = 0$

	• **CCS senza reimmissione con *n* piccolo →** le probabilità osservata da ciascuno dei due sono diverse ma si può dimostrare che sono uguali a quelle che caratterizzano la popolazione → **V.A. DIPENDENTI E IDENTICAMENTE DISTRIBUITE** - $X_1, X_2, \dots \sim^{id} f_x \rightarrow Cov(X,Y) \neq 0$ • **CCS senza reimmissione con *n grande* →** le probabilità osservata da ciascuno dei due sono *minimamente diverse* e dipendono *debolmente* da quanto precedentemente osservato → **V.A. INDIPENDENTI E IDENTICAMENTE DISTRIBUITE:** $X_1, X_2, \dots \approx^{id} f_x \rightarrow Cov(X,Y) = 0$
Media campionaria	Si consideri una campione causale di ampiezza n da una popolazione con media μ e varianza δ^2; sia tale campione IID. Allora si chiama **media campionaria** la media di *n* variabili aleatorie: $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ Proprietà: 1. Il valore atteso della media campionaria è uguale alla media della popolazione (**la media campionaria è uno stimatore non distorto per la media della popolazione**, *vedi oltre*) $$E[\bar{X}] = \mu$$ Dimostrazione: $$E[\bar{X}] = E\left[\frac{1}{n}\sum x_i\right] = \frac{1}{n}E\left[\sum x_i\right] = \frac{1}{n}\sum E[x_i] = \frac{1}{n} * n\mu = \mu$$ 2. La varianza della media campionaria è la varianza della popolazione divisa per la popolosità del campione: $$Var[\bar{X}] = \frac{\sigma^2}{n}$$ Dimostrazione: $$Var[\bar{X}] = Var\left[\frac{1}{n}\sum x_i\right] = \left(\frac{1}{n}\right)^2 Var\left[\sum x_i\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$ 3. La deviazione standard risulta: $$\sigma = \frac{\sigma}{\sqrt{n}}$$ N.B.: se il campionamento è **SENZA REIMMISSIONE** e risulta $n > 0.05N$ allora: $$Var[\bar{X}] = \frac{\sigma^2}{n} * \frac{N-n}{N-1}$$ Dove $\frac{N-n}{N-1}$ è il **fattore di correzione per popolazione finita** **Distribuzione della media campionaria** 1. **POPOLAZIONE NORMALE** **Teorema**:

Sia X un V.A. IID che si distribuisce come una normale $X \sim N(\mu, \sigma^2)$. Sotto queste ipotesi, **se la popolazione ha distribuzione normale allora il campione ha distribuzione ancora normale**

a. CAMPIONE **CON REIMMISSIONE**
   - $E(\bar{X}) = \mu$
   - $Var(\bar{X}) = \frac{\sigma^2}{n}$
   - $\delta(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ (**standard error**)
   - Distribuzione:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

   - Standardizzazione $\rightarrow Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

b. CAMPIONE **SENZA REIMMISSIONE** (con $n > 0.05N \rightarrow$ fattore di correzione per popolazioni finite)
   - $E(\bar{X}) = \mu$
   - $Var(\bar{X}) = \frac{\sigma^2}{n} * \frac{N-n}{N-1}$
   - $\delta(\bar{X}) = \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$ (**standard error**)
   - Distribuzione:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} * \frac{N-n}{N-1}\right)$$

   - Standardizzazione $\rightarrow Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}}$

N.B.: è possibile ridurre l'errore standard (deviazione) aumentando n.

2. **POPOLAZIONE QUALSIASI**

Teorema centrale del limite
Sia X un campione casuale di ampiezza *n* da una popolazione arbitraria con $E(x) = \mu$ e $Var(X) = \sigma^2$; sia *n* sufficientemente elevato (approssimativamente n > 25).
Allora la media (o la somma) di variabili aleatorie IID si distribuisce approssimativamente come una normale con parametri:

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \sum_i X_i \approx N(n\mu, n\sigma^2)$$

Standardizzando:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1) \Leftrightarrow \frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \approx N(0,1)$$

3. **DISTRIBUZIONE BERNOULLIANA**

	Si consideri un campione di una popolazione che si distribuisce secondo una Bernoulliana di parametro p, quindi con media μ e varianza $p(1-p)$ Se risulta $\boldsymbol{np(1-p)>9}$ allora è possibile applicare il **teorema centrale del limite** e P si distribuisce **approssimativamente**: $$\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right)$$
Distribuzione della proporzione campionaria	Data una distribuzione **Bernoulliana**. Sia p la proporzione della popolazione che possiede la caratteristica oggetto di studio. Si dice **proporzione campionaria $\boldsymbol{\hat{P}}$** la **stima della proporzione del successo de carattere osservato *p*, sul totale del campione in esame**: $$\boldsymbol{\hat{P} = \frac{\#popolazione\ successo}{dimensione\ campione} = \frac{\sum_{i=1}^{n} x_i}{n}}$$ Con $\boldsymbol{0 \leq \hat{P} \leq 1}$ La $\boldsymbol{\hat{P}}$ ha lo stesso significato di $\bar{X}$, ma si riferisce al caso particolare della **bernoulliana**; in particolare, per n sufficientemente grande $[\boldsymbol{np(1-p)>9}]$ risulta che: $$\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right)$$ Standardizzazione: $$Z = \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}}$$ E deviazione standard (standard error) $$\sigma = \sqrt{\frac{p(1-p)}{n}}$$ Un modo alternativo, ma del tutto equivalente, per affrontare questo tipo di problemi, è quello di valutare la corrispondente distribuzione binomiale. Data una distribuzione bernoulliana $X \sim Bern(p)$, **la somma di *n* distribuzioni bernoulliane si distribuiscono come una binomiale** $\boldsymbol{X \sim Bin(n,p)}$. La binomiale rappresenta la ripetizione di n esperimenti bernoulliani. Se $\boldsymbol{np(1-p)>9}$, può essere approssimata ad una normale con media e varianza pari a: $$X \approx N(np, np(1-p))$$ X rappresenta il totale di elementi considerati non sotto forma di proporzione. Da cui segue la standardizzazione: $$Z = \frac{X-np}{\sqrt{np(1-p)}}$$

Esercizi tipici sulla media e proporzione campionaria	1. <u>Calcolo dell'errore standard della media o della proporzione campionaria</u> Sono le rispettive deviazioni standard 2. <u>Calcolo della probabilità con media campionaria</u> Possono riguardare diverse tipologie, ma occorre partire sempre: ○ Verificare l'ampiezza del campione e controllare se si necessita del fattore di correzione (campionamento senza reimmissione con n > 0.05N; se N non è fornito supporremo la popolazione infinita e non si necessiterà del fattore di correzione ○ Scrivere la distribuzione della media campionaria secondo la normale: il suo valore atteso è uguale a quello della popolazione; la sua varianza dipende dalla presenza o meno del fattore di correzione ○ Non *spaventarsi* se non vengono forniti *n* o *Var(X)* o E(X): in questi casi l'esercizio chiede, probabilmente, di determinare proprio uno di quei valori; scriviamo, i parametri della normale mantenendo incogniti i valori non dati e ricerchiamoli 3. <u>Calcolo dei percentili</u> Viene fornito il percentile (p) o il suo complementare (1-p); in questi casi occorre prima standardizzare e poi cercare sulle tavole della normale *all'incontrario*: individuo il valor di probabilità che si avvicina di più a quello dato (se intermedio, eseguo una media) e risalgo al valore corrispondente 4. <u>Considerazioni circa la probabilità al variare di n</u> Un risultato fondamentale afferma che **all'aumentare di n, la varianza diminuisce e viceversa**. La varianza è una misura della dispersione dei dati del campione; se stiamo considerando un intervallo che comprende la media, allora una riduzione della varianza e la conseguente concentrazione maggiore dei dati attorno alla media, provocherà un aumento della probabilità; se, viceversa, stiamo considerando un intervallo che include le *code* della normale, una riduzione della varianza, provocherà un minore appiattimento della curva e, quindi, code più sottili: la probabilità diminuisce 5. <u>Calcolo della probabilità con la proporzione campionaria</u> La proporzione campionaria indica un rapporto (e quindi una ***percentuale***) dei successi sul totale della popolazione. Uso la proporzione campionaria quando viene richiesto la probabilità circa una determinata percentuale; uso la distribuzione della somma dei successi, nel caso venga richiesto una probabilità di numero. Stesse considerazioni circa le variazioni di n, valgono anche per la proporzione campionaria 6. <u>Superamento e variazioni della media della popolazione da parte della media campionaria</u> In questo tipo di esercizi viene fornita unicamente la deviazione standard (o la varianza) ma non la media (inutile); distinguiamo due casi modello (possono esserci piccole variazioni ma il ragionamento è il medesimo): ○ Qual è la probabilità che la media campionaria **superi** la media della popolazione per **più** di a $$P(\bar{X} > \mu + a) = P\left(Z > \frac{\mu + a - \mu}{\sigma}\right) = P\left(Z > \frac{a}{\sigma}\right)$$ Ora calcolo normalmente con le tavole Analogamente alla richiesta di quale sia la probabilità che la media campionaria sia **inferiore** alla media della popolazione per **più** di a si risponde come segue: $$P = P(\bar{X} < \mu - a) = P\left(Z < \frac{\mu - a - \mu}{\sigma}\right) = P\left(Z < -\frac{a}{\sigma}\right)$$

<table>
<tr>
<td></td>
<td>

Qual è la probabilità che la media campionaria differisca dalla media della popolazione per più di a
$$P(|\bar{X}-\mu|>a)=P\left(\left|\frac{\bar{X}-\mu}{\sigma}\right|>\frac{a}{\sigma}\right)=P\left(|Z|>\frac{a}{\sigma}\right)$$
Scompongo il valore assoluto
$$P\left(Z>\frac{a}{\sigma}\right)+P\left(Z<-\frac{a}{\sigma}\right)=2\left[1-F\left(\frac{a}{\sigma}\right)\right]$$
Qualora fosse stato richiesta la probabilità che la media campionaria differisse per meno di a allora:
$$P(|\bar{X}-\mu|<a)=P(\mu-a<\bar{X}<\mu+a)$$

N.B.: in un esercizio può capitare che sia richiesto di calcolare la probabilità di un parametro per la n-esima unità specifica estratta dal campione. In questo caso NON dobbiamo valutare la media campionaria ma la distribuzione nel suo complesso poiché la specifica unità è parte della popolazione intera. Infatti, la singola unità del campione ha la stessa distribuzione del campione stesso
</td>
</tr>
<tr>
<td>Esempio riepilogativo</td>
<td>
Si consideri un campione casuale di 50 incassi settimanali (tra loro indipendenti) per un ristorante. Sapendo che l'incasso medio settimanale sia pari a 25 e che la varianza è pari a 4, quale distribuzione è una ragionevole descrizione per l'incasso complessivo delle 50 settimane?

Sia X l'incasso settimanale della i-esima settimana; l'incasso complessivo delle 50 settimane considerate è dato dalla somma dei 50 incassi settimanali, (indipendenti e identicamente distribuiti) ovvero da $\sum_{i=1}^{50} X_i$
Poiché il campione è sufficientemente ampio, possiamo applicare teorema centrale del limite: la somma di variabili aleatorie IID si distribuisce approssimativamente come una normale con i seguenti parametri:

$$\sum_i X_i \approx N(n\mu, n\sigma^2) \rightarrow \sum_i X_i \approx N(25*50, 50*4) \rightarrow \sum_i X_i \approx N(1250, 200)$$
Sia 30 il primo quartile. Supponete che in una qualunque settimana il ristorante considerato raggiunga il pareggio solo se l'incasso è superiore a 30. In quante settimane su 50 ci si può attendere che il ristorante sia in pareggio?

$$P(Incasso\ settimanale \geq 30) = 0.25$$

Ogni settimana rappresenta una prova bernoulliana Y_i, con probabilità di raggiungere il pareggio pari a p = 0.25
Il numero di settimane su 50 in cui il pareggio sarà raggiunto corrisponde alla ripetizione per 50 volte dell'esperimento bernoulliano e quindi ad una binomiale con i seguenti parametri:

$$Y_i \sim Bern(p=0.25) \rightarrow \sum_i Y_i \sim Bin(n=50; p=0.25)$$

Da cui segue che:

$$E\left[\sum_i Y_i\right] = 50*p = 50*0.25 = 12.5$$

Il pareggio è raggiunto in 12.5 settimane su 50
</td>
</tr>
</table>

	Supponiamo che la probabilità che il ristorante raggiunga il pareggio è paria a p=0.40, con quale probabilità il ristorante raggiungerà il pareggio almeno in 15 settimane su 50? Ripartiamo dalla distribuzione binomiale: poiché *np(1-p)>9*, il teorema centrale del limite ci consente di approssimare la binomiale ad una normale; quindi, il numero di settimane in cui verrà raggiunto il pareggio su 50 ha distribuzione: $$\sum_i Y_i \approx N(np; np(1-p)) = N(20,12)$$ Possiamo calcolare la probabilità che il ristorante raggiunga il pareggio in almeno 15 settimane su 50: $$P\left(\sum_i Y_i \geq 15\right) = P\left(Z \geq \frac{15-20}{\sqrt{12}}\right) = P(Z > -1.44) = P(Z < 1.44) = 0.9251$$

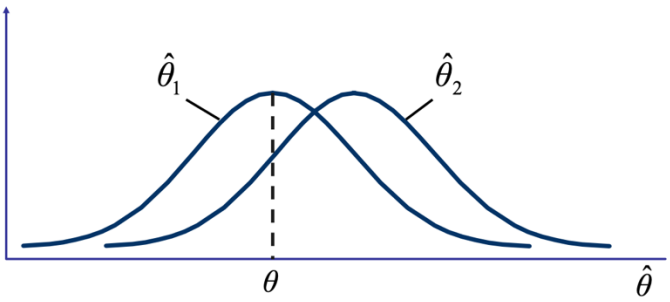
GLI STIMATORI PUNTUALI	
Definizioni	Uno **stimatore** di un parametro $\boldsymbol{\theta}$ è: • Una **variabile aleatoria** che dipende dall'informazione contenuta nella popolazione • Il **cui valore fornisce un'approssimazione del valore sconosciuto del parametro** (a posteriori, dopo l'estrazione del campione) • Uno **specifico valore** dallo **stimatore** viene chiamato **stima puntuale** Un **intervallo di confidenza** fornisce ulteriori informazioni circa la variabilità della stima ed indica un *intervallo di valori plausibili* per il parametro delle popolazioni 
Stime puntuali	Come possiamo trovare uno stimatore per un parametro d'interesse? • **STIME PER ANALOGIE** **Stima** / **Parametro** $\bar{X}$ / μ S^2 / σ^2 S / σ S_{XY} / σ_{XY} r_{XY} / ρ_{XY} • **STIMATORI DEI MINIMI QUADRATI** • **STIMATORI B.L.U.E**
Criteri di valutazione degli stimatori	**1. NON DISTORSIONE** Uno **stimatore puntuale T per un parametro $\boldsymbol{\theta}$ si dirà non distorto se**: $E[T] = \theta$

Se T fosse distorto, allora è possibile misurare la **distorsione dello stimatore D(T)**:

$$D(T) = E[T] - \theta$$

Infine, uno stimatore è non distorto se e solo se ha distorsione nulla:

$$T\ non\ distorto \iff D(T) = 0$$



Nella figura, lo stimatore $\widehat{\theta_1}$ è non distorto; $\widehat{\theta_2}$ è distorto
Da qui segue che:

- La media campionaria è uno stimatore non distorto per la media
- La varianza campionaria è uno stimatore non distorto per la varianza
- La proporzione campionaria è uno stimatore non distorto per la proporzione

Sovrastima, sottostima e non distorsione

- $E[T] > \theta\ o\ D(T) > 0 \Rightarrow$ Sovrastima
- $E[T] < \theta\ o\ D(T) < 0 \Rightarrow$ Sottostima
- $E[T] = \theta\ o\ D(T) = 0 \Rightarrow$ Non distorsione (correttezza)

Più in generale, dato uno stimatore nella forma:

$$T_w = w_1X_1 + w_2X_2 + \cdots + w_nX_n$$

Con $w_n \in R$, risulta che:

- $\sum w_i > 1 \Rightarrow$ Sovrastima
- $\sum w_i < 1 \Rightarrow$ Sottostima
- $\sum w_i = 1 \Rightarrow$ Non distorsione

Osservazioni utili alla **verifica** della **distorsione** di uno stimatore

- $T_1 = \frac{\sum x_i}{n} \rightarrow w_i = \frac{1}{n}$
- $E\left[\frac{\sum x_i}{n}\right] = \frac{1}{n}E[\sum x_i]$
- $E[X] = \mu$
- Per verificare la distorsione di uno stimatore T del parametro θ occorre valutare il suo E[T] e verificare che questo sia pari esattamente a θ

2. NON DISTORSIONE ASINTOTICA

Dato un campione, uno stimatore T si dirà **asintoticamente non distorto** per un parametro θ se $E[T] \rightarrow \theta$ quando $n \rightarrow +\infty$ o, equivalentemente, se $D(T) \rightarrow 0$

3. EFFICIENZA

Quale stimatore scegliere se entrambi sono non distorti? Lo stimatore **più efficiente** è lo stimatore **con la varianza più piccola**:

$$Var[T_1] < Var[T_2] \rightarrow T_1 \text{ più efficiente di } T_2$$

È possibile anche definire l'**efficienza relativa di T_1 rispetto a T_2** come:

$$ER[T_1|T_2] = \frac{Var[T_2]}{Var[T_1]}$$

E T_1 si dirà più efficiente di T_2 quando $ER[T_1|T_2] > 1$

Stimatore più efficiente (stimatore BLUE)

Uno stimatore T* non distorto per un parametro θ si dirà il più efficiente se:

$$Var[T^*] < Var[T]$$

Per qualunque altro stimatore non distorto T

4. STANDARD ERROR

Se T è uno stimatore non distorto per $\theta \rightarrow$ il suo standard error è dato:

$$Std.Error = \sqrt{Var[T]} = \sqrt{E[T - \theta\,]^2}$$

Esso rappresenta la deviazione standard della distribuzione campionario dello stimatore: ovvero, l'errore di stima

Esempi notevoli:

- **Standard error per media campionaria** come stimatore non distorto della media (n<0,05N)

$$Std.Error(\bar{X}) = \begin{cases} \frac{\sigma}{\sqrt{n}} & se\ \sigma \text{ è } nota \\ \frac{s}{\sqrt{n}} & se\ \sigma \text{ è } incognita \end{cases}$$

  Nel secondo caso ho **stimato la varianza con la varianza campionaria**

- Standard error per proporzione campionaria

$$Std.Error(\bar{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

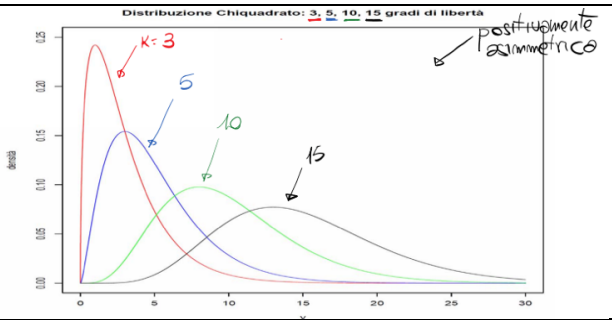
5. ERRORE QUADRATICO MEDIO

L'errore quadratico medio di uno stimatore T (potenzialmente distorto) è dato da:

$$EQM[T] = E[(T - \theta] = Var[T] + D(T)^2$$

Quindi, se **T è non distorto e D(T)=0, risulta**:

$$EQM[T] = Var[T]$$

	Dati due stimatori, lo stimatore con minor EQM risulta essere quello maggiormente preciso ed è, quindi, da preferire
Distribuzione della media campionaria	**Proprietà** La media campionaria è uno **stimatore non distorto** per la media $$E[\bar{X}] = \mu$$ Dimostrazione: $$E[\bar{X}] = E\left[\frac{1}{n}\sum x_i\right] = \frac{1}{n}E\left[\sum x_i\right] = \frac{1}{n}\sum E[x_i] = \frac{1}{n} * n\mu = \mu$$ Distribuzione: *Vedi sopra (parte relativa al primo parziale)*
Distribuzione della proporzione campionaria	*Vedi sopra (parte relativa al primo parziale)*
Distribuzione della varianza campionaria	Si consideri un campione casuale di ampiezza n da una popolazione con media μ e varianza σ^2. La varianza campionaria: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ **Proprietà**: • Per una **qualsiasi popolazione arbitraria** si dimostra che la varianza campionaria è uno stimatore non distorto per la varianza → $E[S^2] = \sigma^2$ • Per una **popolazione normale**, in particolare, si può dimostrare che la varianza della varianza campionaria risulta: $$Var[S^2] = \frac{2\sigma^4}{n-1}$$ È possibile, inoltre, ricondurre la varianza campionaria ad una distribuzione chi-quadrato con (n-1) gradi di libertà nel modo seguente $$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$ E quindi, la varianza campionaria si distribuisce: $$S^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{(n-1)}$$
Distribuzione chi-quadrato	Se V ha una distribuzione chi-quadrato con k gradi di libertà, ovvero $V \sim \chi^2_{(k)}$ allora: ▪ $E[V] = k$ ▪ $Var[V] = 2k$ 

<table>
<tr><th colspan="2">INTERVALLI DI CONFIDENZA E STIMATORI PER INTERVALLO</th></tr>
<tr><td>Stimatore per intervallo</td><td>Uno stimatore per intervallo con livello di confidenza $1-\alpha$ per un parametro θ della popolazione è una coppia di variabili aleatorie, A e B, funzioni del campione casuale, tali che:

▪ $A < B$
▪ $P(A < \theta < B) = 1-\alpha \in (0,1)\ \forall\theta$

$1-\alpha$ è detto livello di confidenza
Un intervallo di confidenza è un intervallo numerico (a,b) calcolato in corrispondenza di una specifica realizzazione campionaria

$$IC_{1-\alpha}(\theta) = (a,b)$$

Interpretazione frequentista
$1-\alpha$ è la frequenza di lungo periodo con cui parametro viene osservato nell'intervallo (a,b) in un gran numero di prove dell'esperimento

Interpretazione dell'intervallo di confidenza
Dire che, ad esempio per la media, $IC_{1-\alpha}(\mu) = (a,b)$ vuol dire che siamo confidenti al $1-\alpha * 100$ (%) che il parametro θ cada in un range di valori compreso tra a e b (ATTENZIONE: è <u>scorretto</u> parlare di probabilità)
Inoltre, NON siamo certi che il parametro cada all'interno dell'intervallo (ma abbiamo solo un certo grado di confidenza che ciò accada); siamo, invece, certi che la media campionaria cada nell'intervallo poiché è su essa centrato

Esiste una relazione diretta tra livello di confidenza e ampiezza dell'intervallo: tanto maggiore è il livello di confidenza, tanto maggiore è l'ampiezza dell'intervallo</td></tr>
<tr><td>Metodo delle quantità pivotali</td><td>Il procedimento utilizzato per la costruzione di un intervallo di confidenza si basa su una quantità detta quantità pivotale o pivot
Dato un campione iid da una distribuzione f_x con parametro θ, una quantità pivotale V è una variabile aleatoria:
▪ Funzione $V = g(X_1, X_2 \dots, X_n;\ \theta)$ del campione e del parametro θ
▪ La distribuzione di V non dipende dal parametro incognito θ
▪ $V = g(X_1, X_2 \dots, X_n;\ \theta)$ invertibile</td></tr>
<tr><th colspan="2">INTERVALLI DI CONFIDENZA PER LA MEDIA DELLA POPOLAZIONE</th></tr>
<tr><td>Intervalli di confidenza per la media della popolazione</td><td>Analizzeremo ora gli intervalli di confidenza per la media a partire da:
▪ POPOLAZIONE NORMALE E…
○ VARIANZA NOTA
○ VARIANZA INCOGNITA (sempre, a meno che diversamente segnalato)
▪ POPOLAZIONE ARBITRARIA
▪ POPOLAZIONE BERNOULLIANA</td></tr>
<tr><td>Intervalli di confidenza per la media di una popolazione normale e varianza nota</td><td>Consideriamo un campione casuale semplice da una popolazione normale e varianza nota σ^2

$$X \sim N(\mu, \sigma^2)$$
Avremo che:
▪ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
▪ Pivot: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$IC_{1-\alpha}(\mu) = \left(\bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = (LCL, UCL)$$</td></tr>
</table>

	▪ Ampiezza: $w = UCL - LCL = 2z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}$ ▪ Margine di errore: $ME = \frac{w}{2} = z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}$ N.B.: ricorda di calcolare correttamente α; $z_{\frac{a}{2}}$ è il quantile della N(0,1) di ordine $1 - \frac{a}{2}$ Proprietà dell'intervallo: ▪ **Affidabilità** → $1 - \alpha$ rappresenta il livello di confidenza, ovvero quanto siamo confidenti di trovare μ in quell'intervallo. Attenzione: siamo certi di trovare la media campionaria (è il centro dell'intervallo) ma non la media della popolazione (per la quale siamo solo confidenti al livello $1 - \alpha$ di trovarla) ▪ **Precisione** → quanto sono precise le informazioni fornite dall'intervallo circa la media? Ci viene detto dal margine d'errore e dall'ampiezza dell'intervallo; ○ Nel caso di **varianza nota**, **la precisione aumenta all'aumentare di *n* del campione** e, quindi, l'**intervallo diviene più piccolo** ○ Nel caso di **varianza incognita**, questo non è più vero poiché, al variare di *n*, varia anche la varianza campionaria (e non sappiamo in che direzione): non vi è certezza ○ N.B.: anche per la bernoulliana non è vero dal momento che lo *standard error* dipende da P cappuccio Le due condizioni rappresentano un **trade off**: all'aumentare dell'uno si riduce necessariamente l'altro Fattori che influenzano il margine d'errore: $$ME = z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}$$ ▪ Maggiore è σ (cioè la variabilità di $\bar{X}$) maggiore sarà Me ▪ Maggiore è n, minore è ME ▪ Maggiore è $1 - a$, maggiore è ME (minore precisione) **Determinazione dell'ampiezza campionaria per ottenere ME desiderato** Per un fissato livello di confidenza $1 - \alpha$ quale valore di n garantisce che si raggiunga un ME* desiderato $$ME = z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}} \leq ME^*$$ Basta risolvere l'equazione per n e si trova facilmente: $$n \geq \frac{\left(z_{\frac{a}{2}}\right)^2 \sigma^2}{ME^{*2}}$$
Intervalli di confidenza per la media di una	Consideriamo un campione casuale semplice da una **popolazione normale varianza non nota σ^2** $$X \sim N(\mu, \sigma^2)$$

<table>
<tr>
<td>popolazione normale e varianza non nota</td>
<td>Avremo che:
▪ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
▪ Pivot: $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T_{(n-1)}$
$$IC_{1-\alpha}(\mu) = \left(\bar{X} \pm t_{(n-1),\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) = (LCL, UCL)$$
▪ Ampiezza: $w = UCL - LCL = 2t_{(n-1),\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
▪ Margine di errore: $ME = \frac{w}{2} = t_{(n-1),\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$T_{(n-1)}$ è la distribuzione T di Student con $n-1$ gradi di libertà

Proprietà: se T ha una distribuzione T di Student con k gradi di libertà, e scriveremo $T \sim T_{(k)}$, allora:
$$E[T] = 0 \; se \; k > 1 \quad e \quad Var[T] = \frac{k}{k-2} \; se \; k > 2$$
N.B.: nelle tavole della T di Student:
• α indica i quantili della distribuzione verso destra (indica la probabilità delle code)
• υ indicano i gradi di libertà

È L'UNICO CASO IN CUI NASCE LA T DI STUDENT</td>
</tr>
<tr>
<td>Intervalli di confidenza per la media di una popolazione arbitraria</td>
<td>Consideriamo un campione casuale semplice da una popolazione arbitraria e varianza non nota. Sia $n \geq 30$. Avremo:
▪ $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ in base al <u>teorema centrale del limite</u>
▪ Pivot: $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \approx N(0,1)$
▪ Intervallo di confidenza di livello (approssimativamente) $1-\alpha$
$$IC_{1-\alpha}(\mu) = \left(\bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) = (LCL, UCL)$$
▪ Ampiezza: $w = UCL - LCL = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
▪ Margine di errore: $ME = \frac{w}{2} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Intervalli di confidenza per la proporzione campionaria</td>
<td>Consideriamo un campione casuale semplice da una popolazione bernoulliana di parametro $p \in [0,1]$. Sia $n \geq 30$ (sufficientemente elevato). Avremo:
▪ $\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right)$ in base al <u>teorema centrale del limite</u>
▪ Pivot: $Z = \frac{\hat{P}-p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \approx N(0,1)$
▪ Intervallo di confidenza di livello (approssimativamente) $1-\alpha$
$$IC_{1-\alpha}(p) = \left(\hat{P} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) = (LCL, UCL)$$</td>
</tr>
</table>

<table>
<tr><td></td><td>

- Ampiezza: $w = UCL - LCL = 2z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$
- Margine di errore: $ME = \frac{w}{2} = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$

Determinazione dell'ampiezza campionaria per ottenere ME desiderato
Per un fissato livello di confidenza $1-\alpha$ quale valore di n garantisce che si raggiunga un ME* desiderato

$$ME = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq ME^*$$

Si dimostra

$$n \geq \frac{\left(z_{\frac{\alpha}{2}}\right)^2 0.25}{ME^{*2}}$$

(da arrotondare sempre per eccesso)

0,25 rappresenta valore massimo che $\hat{P}(1-\hat{P})$ può raggiungere: nè il punto di massimo della parabola $\hat{P}(1-\hat{P}) = 0$ nel piano $[\hat{P}, \hat{P}(1-\hat{P})]$
</td></tr>
<tr><td>Intervalli di confidenza con assunzione di normalità</td><td>

Consideriamo un campione casuale semplice da una **popolazione arbitraria**. Sia $\boldsymbol{n < 30}$: n non è abbastanza grande per usare il teorema centrale del limite
Per poter procedere occorre **assumere la normalità del campione**; così facendo ci riconduciamo al caso di **popolazione normale** e varianza nota/non nota a seconda del caso richiesto:

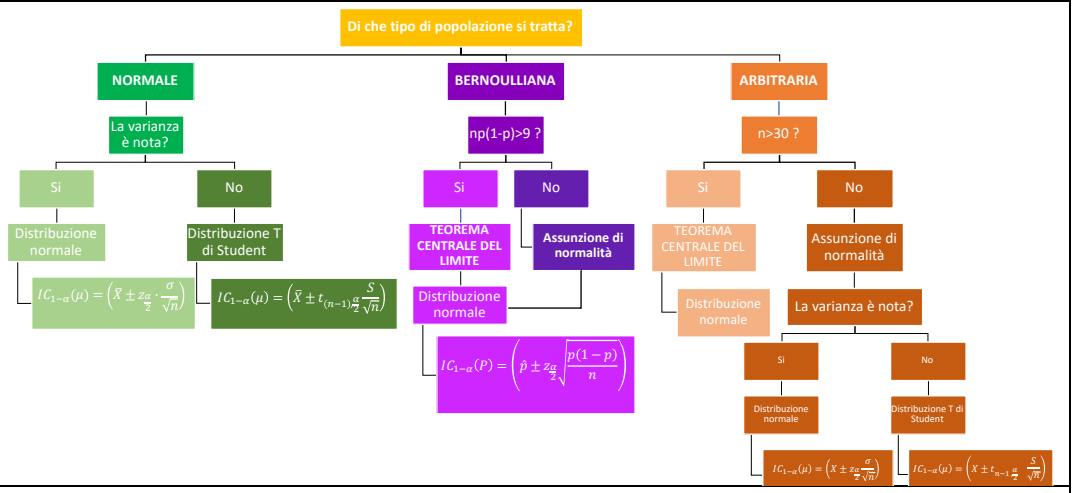
<table>
<tr><td>Varianza nota</td><td>$IC_{1-\alpha}(\mu) = \left(\bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$</td></tr>
<tr><td>Varianza non nota</td><td>$IC_{1-\alpha}(\mu) = \left(\bar{X} \pm t_{(n-1),\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right)$</td></tr>
</table>

</td></tr>
<tr><td>Fattore di correzione per popolazioni finte</td><td>

Occorre ricordare che qualora la **popolazione** presentasse un numero di **individui finito N** occorre applicare allo standard errore il fattore di correzione

$$n > 0.05N \Rightarrow SD * \frac{N-n}{N-1}$$

Questo risultato vale sempre ed è bene ricordarlo!
</td></tr>
</table>

Riassunto	
INTERVALLI DI CONFIDENZA PER LA DIFFERENZA TRA MEDIE DI DUE POPOLAZIONI	
Intervalli di confidenza per la **differenza tra le medie di due popolazioni**	Consideriamo due popolazioni; vogliamo confrontarle in termini di media di una variabile di interesse • POPOLAZIONE 1 $$X_1, \dots, X_{n_x} \sim^{iid} f_x \quad E[X] = \mu_x \quad Var[X] = \sigma_x^2$$ • POPOLAZIONE 2 $$Y_1, \dots, Y_{n_x} \sim^{iid} f_y \quad E[Y] = \mu_y \quad Var[X] = \sigma_y^2$$ Diamo due importanti definizioni: • **CAMPIONI DIPENDENTI**: le **<u>stesse unità</u>** vengono analizzate prima e dopo il trattamento → $Cov(X_i, Y_i) \neq 0$ • **CAMPIONI INDIEPNDENTI**: **<u>unità (potenzialmente) differenti</u>** vengono analizzate prima e dopo il trattamento → $Cov(X_i, Y_i) = 0$ Analizzeremo i seguenti casi: • **CAMPIONI DIPENDENTI** (varianza campionaria S_D^2) ○ POPOLAZIONE NORMALE E VARIANZE NON NOTE E DIVERSE ○ POPOLAZIONE ARBITRARIA (n>30) E VARIANZE NON NOTE MA ASSUNTE UGUALI ○ POPOLAZIONE ARBITRARIA (n>30) E VARIANZE NON NOTE E DIVERSE (SOFTWARE R) • **CAMPIONI INDIPENDENTI** (varianza pooled S_P^2) ○ POPOLAZIONE NORMALE ▪ VARIANZE INCOGNITE MA ASSUNTE UGUALI ▪ VARIANZE INCOGNITE E DIVERSE (SOFTWARE R)
Intervalli di confidenza per la differenza tra le medie di due **popolazioni normali dipendenti** (**varianze non note**)	Consideriamo due campioni casuali semplici **dipendenti** $(Cov(X_i, Y_i) \neq 0)$ da **popolazioni normali** **(alternativamente, assumiamo la normalità)** **varianza non nota** σ^2 Definito $D_i = X_i - Y_i$ avremo che: $$D_1, \dots, D_n \sim N(\mu_D, \sigma_D^2) \; con$$ $$\mu_D = \mu_X - \mu_Y$$ $$S_D^2 = S_X^2 + S_Y^2 + (-2r_{xy})$$ • Stimatore di $\mu_D \rightarrow \bar{D} = \frac{1}{n}\sum D_i \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$

<table>
<tr><td></td><td>

- Pivot $\rightarrow T = \frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} \sim T_{(n-1)}$
- Intervallo di confidenza di livello $1-\alpha$

$$IC_{1-\alpha}(\mu_D) = IC_{1-\alpha}(\mu_X - \mu_Y) = \left(\bar{D} \pm t_{(n-1),\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}\right)$$

Proprietà di D_i

- $E[D_i] = E[X_i - Y_i] = E[X_i] - E[Y_i] = \mu_X - \mu_Y = \mu_D$
- $Var[D_i] = Var[X_i - Y_i] = Var[X_i] + Var[Y_i] - 2Cov[X_i, Y_i]$

Proprietà dello stimatore $\bar{D}$

- $E[\bar{D}] = \mu_D$
- $Var[\bar{D}] = \frac{\sigma_D^2}{n} = \frac{\sigma_X^2+\sigma_Y^2-2\sigma_{XY}}{n}$

</td></tr>
<tr><td>Intervalli di confidenza per la differenza tra le medie di due popolazioni normali indipendenti (varianze note)</td><td>

Consideriamo due campioni casuali semplici **indipendenti** $(\boldsymbol{Cov(X_i, Y_i) = 0})$ da **popolazioni normali** e **varianza nota** $\boldsymbol{\sigma^2}$

- Stimatore di $\Delta\mu = \mu_X - \mu_Y \rightarrow \bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$
- Pivot $\rightarrow Z = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \sim N(0,1)$
- Intervallo di confidenza di livello $1-\alpha$

$$IC_{1-\alpha}(\mu_X - \mu_Y) = \left((\bar{X} - \bar{Y}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right)$$

Proprietà di $\bar{X} - \bar{Y}$

- $E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}]$
- $Var[\bar{X} - \bar{Y}] = Var[\bar{X}] - Var[\bar{Y}]$

N.B.: se una popolazione fosse arbitraria e n non abbastanza grande per usare il teorema centrale del limite, **dobbiamo assumere la normalità**

</td></tr>
<tr><td>Intervalli di confidenza per la differenza tra le medie di due popolazioni normali indipendenti (varianze non note ma assunte uguali)</td><td>

Consideriamo due campioni casuali semplici **indipendenti** $(\boldsymbol{Cov(X_i, Y_i) = 0})$ da **popolazioni normali con media μ non nota e varianze non note ma assunte uguali**

- Considerando che $\sigma_X^2 = \sigma_Y^2$, le due varianze campionarie sono stimatori naturali e non distorti; tuttavia, si dimostra che la **varianza pooled** è uno stimatore non distorto e più efficiente delle varianze campionarie *semplici*

$$S_P^2 = \frac{(n_x - 1)S_X^2 + (n_y - 1)S_Y^2}{n_X + n_Y - 2}$$

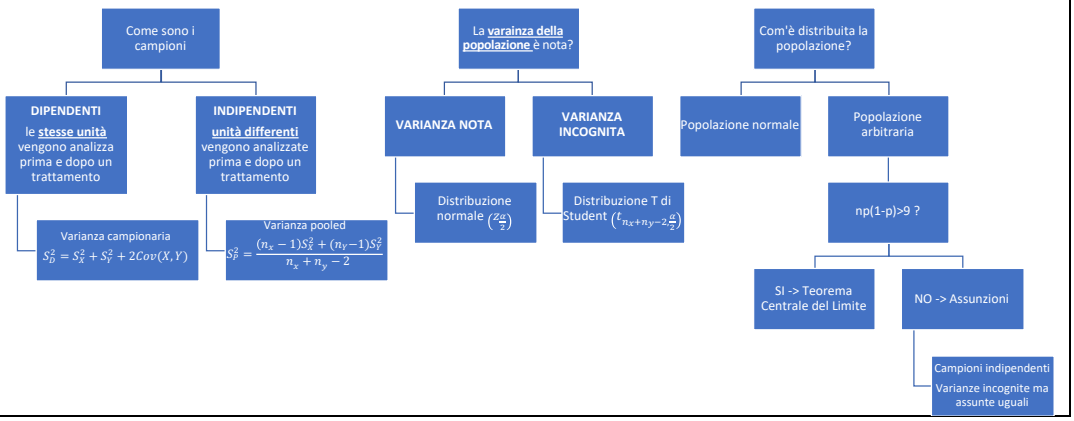
- Possiamo, quindi, stimare la varianza di $\bar{X} - \bar{Y}$:

$$Var[\bar{X} - \bar{Y}]_{stima} = S_P^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)$$

- Pivot $\rightarrow T = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_P\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim T_{(n_x+n_y-2)}$
- Intervallo di confidenza di livello $1-\alpha$

</td></tr>
</table>

	$$IC_{1-\alpha}(\mu_X-\mu_Y)=\left((\bar{X}-\bar{Y})\pm t_{n_X+n_Y-2,\frac{\alpha}{2}}S_P\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}\right)$$
Intervalli di confidenza per la differenza tra le medie di due **popolazioni normali indipendenti (varianze non note e differenti)**	Consideriamo due campioni casuali semplici **indipendenti** $(Cov(X_i,Y_i)=0)$ da **popolazioni normali** **con media μ non nota e varianze non note e differenti** • Possiamo, quindi, stimare la varianza di $\bar{X}-\bar{Y}$: $$Var[\bar{X}-\bar{Y}]_{stima}=\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}$$ • Pivot $\rightarrow Z=\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}}\approx T_{(v)}$ • Intervallo di confidenza di livello approssimativamente pari a $1-\alpha$ $$IC_{1-\alpha}(\mu_X-\mu_Y)=\left((\bar{X}-\bar{Y})\pm t_{v,\frac{\alpha}{2}}\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}\right)$$
Intervalli di confidenza di **popolazioni arbitrarie dipendenti (grandi campioni)** con **varianze non note ma uguali**	Consideriamo due campioni sufficientemente grandi ($n\geq 30$) casuali semplici **dipendenti** $(Cov(X_i,Y_i)\neq 0)$ da **popolazioni arbitrarie** **e varianze non note ma uguali** • $\bar{D}=\frac{1}{n}\sum D_i\sim N\left(\mu_D,\frac{\sigma_D^2}{n}\right)$ • $Z=\frac{\bar{D}-\mu_D}{S_D/\sqrt{n}}\approx N(0,1)$ • Intervallo di confidenza di livello $1-\alpha$ $$IC_{1-\alpha}(\mu_D)=IC_{1-\alpha}(\mu_X-\mu_Y)=\left(\bar{D}\pm z_{\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}}\right)$$
Intervalli di confidenza di **popolazioni arbitrarie indipendenti (grandi campioni)** con **varianze non note e differenti**	Consideriamo due campioni sufficientemente grandi ($n\geq 30$) casuali semplici **indipendenti** $(Cov(X_i,Y_i)=0)$ da **popolazioni arbitrarie** **e varianze non note e differenti** • Pivot $\rightarrow Z=\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}}\approx N(0,1)$ • Intervallo di confidenza di livello approssimativamente pari a $1-\alpha$ $$IC_{1-\alpha}(\mu_X-\mu_Y)=\left((\bar{X}-\bar{Y})\pm z_{\frac{\alpha}{2}}\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}\right)$$
Intervalli di confidenza di **popolazioni arbitrarie** (campioni non sufficientemente grandi)	Assunzioni: • Campioni indipendenti • Varianze (non note) ma assunte uguali • Normalità della popolazione Mi riconduco al caso di **campioni indipendenti con varianze non note ma assunte uguali** → calcolo della varianza pooled

	$$IC_{1-\alpha}(\mu_X - \mu_Y) = \left((\bar{X} - \bar{Y}) \pm t_{n_X+n_Y-2,\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right)$$
Riassunto	

TEST D'IPOTESI	

Concetti di base dell'ipotesi statistica	Si definisce **ipotesi statistica un'affermazione** circa il **valore di un parametro incognito della popolazione**
	
	Tipi di ipotesi:
	• **Semplice**: specifica un singolo valore per il paramento della popolazione considerato
	• **Composta**: specifica uno o più intervalli di valori per il parametro della popolazione considerato
	  ○ **Unilaterale**: considera tutti i possibili valori del paramento della popolazione a destra o a sinistra rispetto all'ipotesi fatta
	  ○ **Bilaterale**: considera tutti i possibili valori diversi dall'ipotesi nulla,
	
	Ipotesi nulla e alternativa
	• **IPOTESI NULLA H_0** → ipotesi da considerarsi vera a meno di ottenere prove evidenti della validità del suo contrario (**rappresenta lo status quo**)
	• **IPOTESI ALTERNATIVA H_1** → ipotesi contro la quale viene verificata l'ipotesi nulla e che viene considerata vera se l'ipotesi nulla è considerata falsa (**affermazione da porre a verifica e per la quale si ricerca l'evidenza**)
	
	Relazione tra test d'ipotesi bilaterali e intervalli di confidenza
	Un intervallo di confidenza identifica una coppia di valori a e b che con una confidenza del 100(1 – α)% contiene il valore del parametro θ.
	Se l'intervallo include il 100(1 – α)% di valori più plausibili per θ, esso esclude il 100α% di campioni meno plausibili.
	Se il valore posto in ipotesi nulla θ_0 è fuori dall'intervallo perché inferiore ad a o superiore a b sarà considerato un valore non verosimile rispetto alla realizzazione campionaria

Decisione, statistica test ed errori	Il processo decisionale usa una **statistica test** costruita a partire dagli **stimatori puntuali del parametro oggetto di studio**; essa **assume una distribuzione campionaria nota**: questo ci permette di **determinare i valori della statistica test che avrebbero una bassa probabilità di verificarsi** se l'***ipotesi nulla fosse vera***. Se la statistica test assume uno tra questi valori, **rifiutiamo** l'ipotesi nulla; altrimenti **non rifiutiamo** l'ipotesi nulla

Per scegliere tra ipotesi nulla e ipotesi alternativa dobbiamo sviluppare una *regola di decisione* chiamata **test** per verificare un'ipotesi contro l'altra
Tuttavia, possiamo anche commettere degli **errori:**

	IPOTESI VERA	
	H_0	H_1
RIFIUTO H_0	ERRORE DI PRIMA SPECIE	NESSUN ERRORE
NON RIFIUTO H_0	NESSUN ERRORE	ERRORE DI SECONDA SPECIE

Quindi diremo di aver commesso:

- **Errore di I specie: rifiuto H_0 quando è vera H_0**
  - La nostra regola di decisione sarà definita in modo da ottenere una **probabilità di commettere l'errore di primo tipo** pari ad α e detta **livello di significatività**

$$\alpha = P(rifiutare\ H_0 | H_0 vera = h_0)$$

  - I livelli di significatività usuali sono $\alpha = (0.05, 0.01, 0.10)$

- **Errore di II specie: rifiuto H_1 quando è vera H_1**
  - La probabilità che si verifichi l'errore di secondo tipo è indicata con β
  - $1 - \beta$ viene detta **potenza del test** ed è la **probabilità di non rifiutare H_1 quando H_1 è vera, ovvero di commettere l'errore di secondo tipo**

$$\beta = P(non\ rifiutare\ H_0 | H_1 vera = h_1)$$

Esiste un **trade off** tra queste due probabilità: **ogni riduzione della probabilità dell'errore di primo tipo α implica un aumento della probabilità dell'errore di secondo tipo β, e viceversa**

Problemi tipo nella verifica di ipotesi per la media di una popolazione

Sia μ_0 il valore *noto* della media che vogliamo testare. Distinguiamo tre casi:

CASO I	CASO II	CASO III
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$

p-value

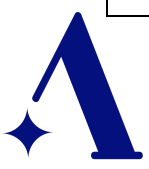
Definiamo **livello di significatività osservato** o ***p-value*** la **probabilità** di **ottenere un valore della statistica test più estremo del valore osservato**, ovvero quel livello di significatività al quale l'ipotesi nulla può essere rifiutata, dato il valore osservato della statistica campionaria
L'**importanza** del p-value è legata al fatto che esso fornisce l'esatta probabilità di rifiuto dell'ipotesi nulla (**livello di significatività al quale rifiutiamo l'ipotesi nulla**) derivante dall'osservazione dei dati campionari. Il p-value ci dice *quanto lontani o vicini siamo dal rifiutare l'ipotesi nulla*

Regione di rifiuto e regione di accettazione

La regione di rifiuto, complementare a quella di accettazione, è una regione indicabile con:

$$R = \{x_1, x_2, \dots, x_n : regola\ di\ \boldsymbol{rifiuto\ di\ H_0}\}$$

E, analogamente, la regione di accettazione si indica con:



	$R = \{x_1, x_2, \dots, x_n : regola\ di\ \boldsymbol{non\ rifiuto\ di\ H_0}\}$

VERIFICA DI IPOTESI PER LA **MEDA DI UNA POPOLAZIONE**		
Verifica di ipotesi sulla **media** di una **popolazione normale** con **varianza nota** → **distribuzione normale**	Statistica test	$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim^{\mu=\mu_0} N(0,1)$
	CASO I $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	
	Regola di rifiuto	• $\bar{X} < \mu_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ oppure $\bar{X} > \mu_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ • $\lvert Z \rvert = \left\lvert \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\rvert > z_{\frac{\alpha}{2}}$ • $p - value = 2P(Z > \lvert z_{oss} \rvert) < \alpha$
	CASO II $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	
	Regola di rifiuto	• $\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ • $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$ • $p - value = P(Z > z_{oss}) < \alpha$
	CASO III $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	
	Regola di rifiuto	• $\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$ • $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < -z_\alpha$ $p - value = P(Z < z_{oss}) < \alpha$
Verifica di ipotesi sulla **media** di una **popolazione normale** con **varianza non nota** → **distribuzione T di Student**	Statistica test	$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim^{\mu=\mu_0} T_{(n-1)}$
	CASO I $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	
	Regola di rifiuto	• $\bar{X} < \mu_0 - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ oppure $\bar{X} > \mu_0 + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ • $\lvert T \rvert = \left\lvert \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \right\rvert > t_{n-1,\frac{\alpha}{2}}$ • $p - value = 2P(T > \lvert t_{oss} \rvert) < \alpha$
	CASO II $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	
	Regola di rifiuto	• $\bar{X} > \mu_0 + t_{n-1,\alpha} \frac{S}{\sqrt{n}}$ • $Z = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} > t_{n-1\alpha}$ • $p - value = P(T > t_{oss}) < \alpha$
	CASO III $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	

<table>
<tr><td></td><td></td><td>Regola di rifiuto</td><td>• $\bar{X} < \mu_0 - t_{n-1,\alpha}\frac{S}{\sqrt{n}}$
• $Z = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} < -t_{n-1,\alpha}$
• $p - value = P(T < t_{oss}) < \alpha$</td><td></td></tr>
<tr><td>Verifica di ipotesi sulla media di una popolazione arbitraria con varianza non nota → grandi campioni (n>30)</td><td colspan="3">Per n≥30 applichiamo il Teorema centrale del limite e impieghiamo una distribuzione normale
<table>
<tr><td>Statistica test</td><td>$Z = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} \approx^{\mu=\mu_0} N(0,1)$</td></tr>
<tr><td colspan="2">CASO I
$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>• $\bar{X} < \mu_0 - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$ oppure $\bar{X} > \mu_0 + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$
• $|Z| = \left|\frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}\right| > z_{\frac{\alpha}{2}}$
• $p - value = 2P(Z > \lvert z_{oss}\rvert) < \alpha$</td></tr>
<tr><td colspan="2">CASO II
$H_0: \mu \leq \mu_0$
$H_1: \mu > \mu_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>• $\bar{X} > \mu_0 + z_\alpha\frac{S}{\sqrt{n}}$
• $Z = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} > z_\alpha$
• $p - value = P(Z > z_{oss}) < \alpha$</td></tr>
<tr><td colspan="2">CASO III
$H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>• $\bar{X} < \mu_0 - z_\alpha\frac{S}{\sqrt{n}}$
• $Z = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} < -z_\alpha$
$p - value = P(Z < z_{oss}) < \alpha$</td></tr>
</table></td><td></td></tr>
</table>

<table>
<tr>
<td>Verifica di ipotesi sulla media di una proporzione con varianza non nota → grandi campioni (n>30)</td>
<td>
Se n è estratto da una popolazione grande a sufficienza da poter affermare che $np(1-p) > 9$ allora vale il teorema centrale del limite e applichiamo una distribuzione normale
<table>
<tr><td>Statistica test</td><td>$Z = \dfrac{\bar{P} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \approx^{p=p_0} N(0,1)$</td></tr>
<tr><td colspan="2">CASO I
$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>
• $\bar{P} < p_0 - z_{\frac{\alpha}{2}}\sqrt{\dfrac{p_0(1-p_0)}{n}}$ oppure $\bar{X} > p_0 + z_{\frac{\alpha}{2}}\sqrt{\dfrac{p_0(1-p_0)}{n}}$
• $|Z| = \left|\dfrac{\bar{P}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}\right| > z_{\frac{\alpha}{2}}$
• $p - value = 2P(Z > |z_{oss}|) < \alpha$
</td></tr>
<tr><td colspan="2">CASO II
$H_0: \mu \leq \mu_0$
$H_1: \mu > \mu_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>
• $\bar{P} > p_0 + z_\alpha \dfrac{\bar{P}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
• $Z = \dfrac{\bar{P}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} > z_\alpha$
• $p - value = P(Z > z_{oss}) < \alpha$
</td></tr>
<tr><td colspan="2">CASO III
$H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>
• $\bar{P} > p_0 - z_\alpha \dfrac{\bar{P}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
• $Z = \dfrac{\bar{P}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} < -z_\alpha$
• $p - value = P(Z < z_{oss}) < \alpha$
</td></tr>
</table>
</td>
</tr>
<tr>
<td>Potenza di un test</td>
<td>
Quando rifiutiamo l'ipotesi nulla deduciamo che vi sia una forte evidenza empirica a favore della nostra conclusione
Quando non rifiutiamo l'ipotesi nulla è perché:

L'ipotesi nulla è vera
L'ipotesi nulla è falsa ma noi non la rifiutiamo (errore di secondo tipo)

Qual è la probabilità di commettere un errore di secondo tipo?
La regola di rifiuto permette di determinare i valori della media campionaria che permettono di non rifiutare l'ipotesi nulla, ovvero la regione di accettazione: essa è infatti la complementare della regione di rifiuto
La probabilità dell'errore di secondo tipo $\boldsymbol{\beta}$ è pari alla probabilità che la media campionaria appartenga alla regione di accettazione
Il suo complementare $\mathbf{1-\beta}$ si chiama potenza
Fissato un livello di significatività pari ad α e fornito il vero valore del parametro in test dobbiamo:

Definire la regola per cui $\beta = P(non\ rifiutare\ H_0 | H_1 vera)$
Applicare la regola di non rifiuto: cerco il valore critico x_c che separa la regione di rifiuto da quella di accettazione.
Calcolare la probabilità normalizzato: si ottiene β
Alla fine, calcola la potenza come $1-\beta$

</td>
</tr>
</table>

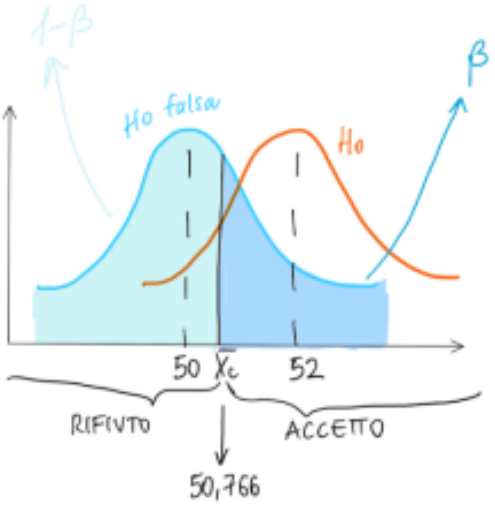
Per **aumentare la potenza del test** occorre:

- Aumentare l'**ampiezza campionaria**
- Aumentare α, ossia la probabilità di rifiutare H_0

<u>Esempio numerico con grafico</u>
Consideriamo il seguente test unilaterale sulla media della popolazione:

$$H_0: \mu \geq 52$$
$$H_1: \mu < 52$$

Assumiamo che la vera media sia $\mu^* = 50$ (ovvero, H_1 vera); assumiamo che la varianza sia nota, e pari a 6, e che la popolosità del campione sia $n = 64$. Consideriamo $\alpha = 0.05$



Come primo step dobbiamo trovare β, e per farlo occorre calcoalr eil valore critico di rifiuto x_c che separa la regione di rifiuto da quella di accettazione. Essendo la vera media (50) inferiore a quella fornita dall'ipotesi nulla (52), allora la regione di rifiuto di H_0 risulta essere quella minore di x_c
Perciò si ha:

$$\beta = P(X \geq x_c | \mu^* = 50)$$

Per trovare x_c, nel seguente caso, applico:

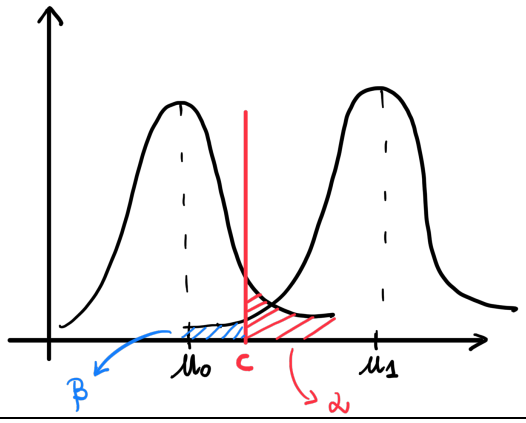
$$x_c = \mu_0 - z_\alpha * \frac{\sigma}{\sqrt{n}} = 50{,}766$$

Da cui:

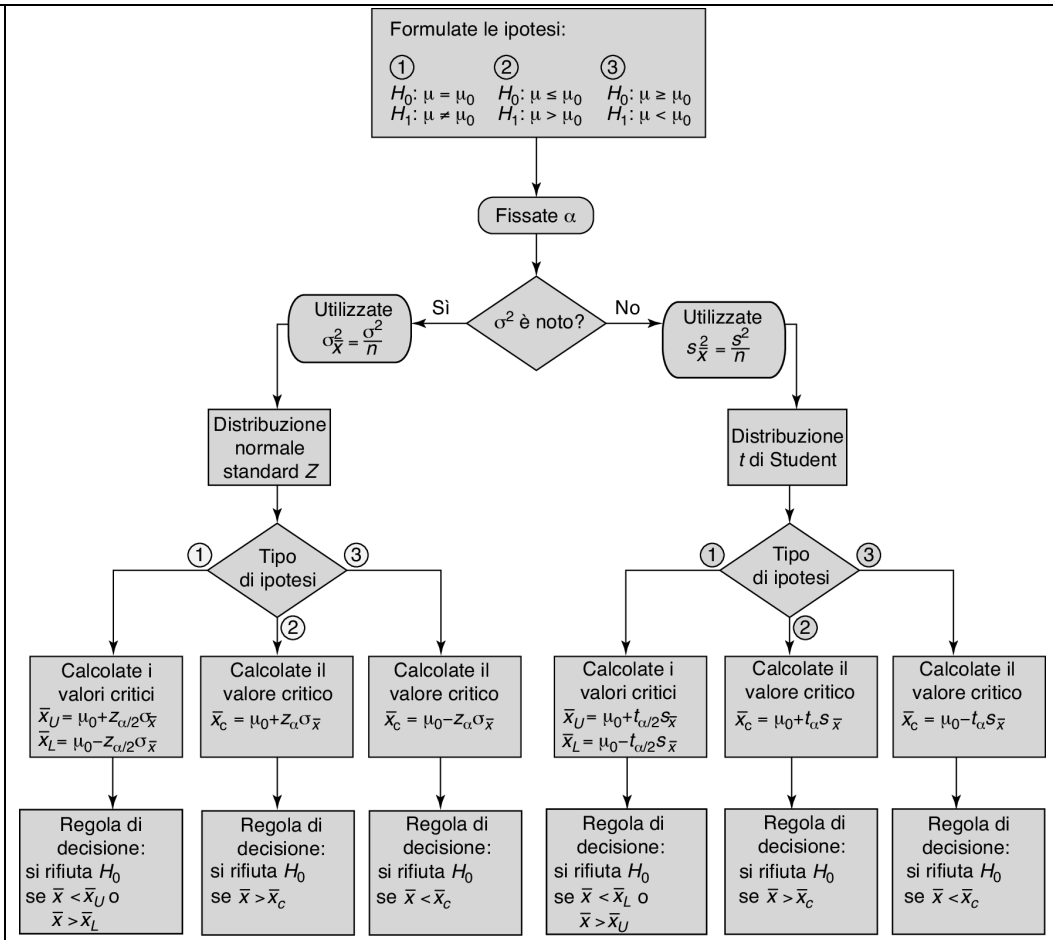
$$\beta = P(X \geq 50{,}766) = P\left(X \geq \frac{50{,}766 - 50}{6/\sqrt{64}}\right) = P(Z \geq 1{,}02) = 0{,}1539$$

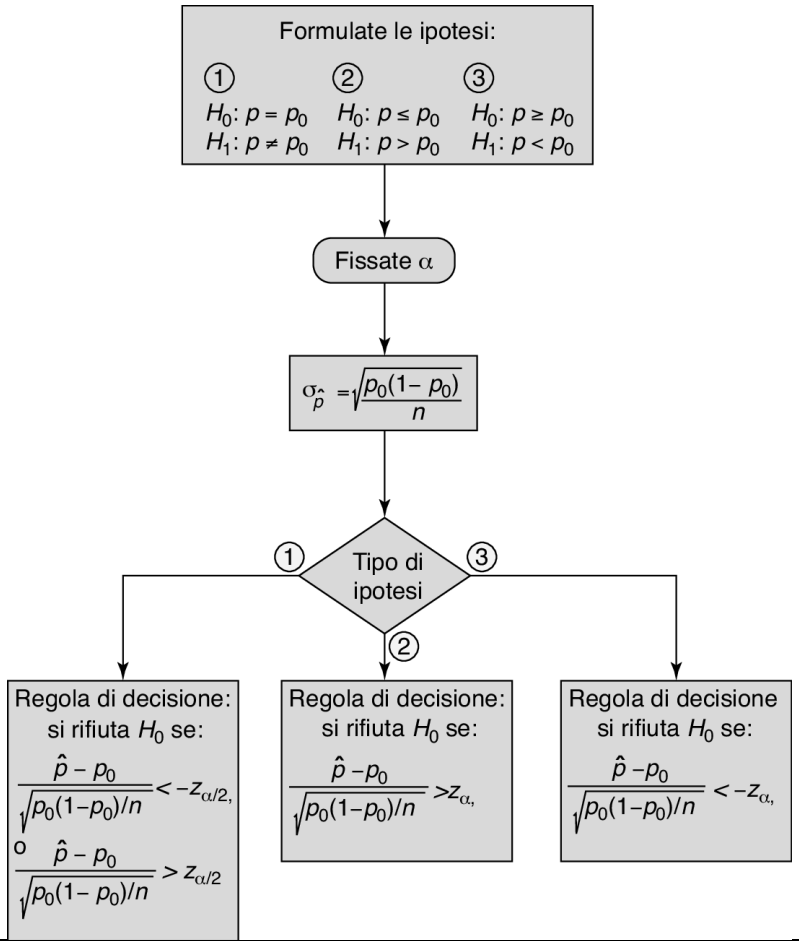
Segue

$$1 - \beta = 0{,}8461 \rightarrow \text{Potenza}$$

Relazione grafica tra α e β	Rappresentiamo graficamente e confrontiamo l'ipotesi nulla e l'ipotesi alternativa. A titolo esemplificativo, consideriamo il seguente test con la relativa regione di rifiuto: $$H_0: \mu < \mu_0$$ $$H_1: \mu > \mu_0 = \mu_1$$ $$R = \{\bar{x} > c\}$$ Consideriamo: • α = probabilità di rifiutare H_0 quando H_0 è vera (area a destra di c sotto la curva di H_0) • β = probabilità di non rifiutare H_0 quando H_1 è vera vera (area a sinistra di c sotto la curva di H_1) Può essere chiesto di descrivere la relazione tra le aree per determinati cambiamenti di c, μ_0, μ_1, n (N.B.: aumentando n, la varianza diminuisce e la curva si appiattisce, concentrandosi sul centro) 

Riassunto


Formulate le ipotesi:
① $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$
② $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$
③ $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$
Fissate α
σ^2 è noto?
Sì
No
Utilizzate $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$
Utilizzate $s^2_{\bar{x}} = \frac{s^2}{n}$
Distribuzione normale standard Z
Distribuzione t di Student
Tipo di ipotesi
Calcolate i valori critici $\bar{x}_U = \mu_0 + z_{\alpha/2}\sigma_{\bar{x}}$ $\bar{x}_L = \mu_0 - z_{\alpha/2}\sigma_{\bar{x}}$
Calcolate il valore critico $\bar{x}_c = \mu_0 + z_\alpha \sigma_{\bar{x}}$
Calcolate il valore critico $\bar{x}_c = \mu_0 - z_\alpha \sigma_{\bar{x}}$
Calcolate i valori critici $\bar{x}_U = \mu_0 + t_{\alpha/2} s_{\bar{x}}$ $\bar{x}_L = \mu_0 - t_{\alpha/2} s_{\bar{x}}$
Calcolate il valore critico $\bar{x}_c = \mu_0 + t_\alpha s_{\bar{x}}$
Calcolate il valore critico $\bar{x}_c = \mu_0 - t_\alpha s_{\bar{x}}$
Regola di decisione: si rifiuta H_0 se $\bar{x} < \bar{x}_U$ o $\bar{x} > \bar{x}_L$
Regola di decisione: si rifiuta H_0 se $\bar{x} > \bar{x}_c$
Regola di decisione: si rifiuta H_0 se $\bar{x} < \bar{x}_c$
Regola di decisione: si rifiuta H_0 se $\bar{x} < \bar{x}_L$ o $\bar{x} > \bar{x}_U$
Regola di decisione: si rifiuta H_0 se $\bar{x} > \bar{x}_c$
Regola di decisione: si rifiuta H_0 se $\bar{x} < \bar{x}_c$


Formulate le ipotesi:
① $H_0: p = p_0$ $H_1: p \neq p_0$
② $H_0: p \leq p_0$ $H_1: p > p_0$
③ $H_0: p \geq p_0$ $H_1: p < p_0$
Fissate α
$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$
Tipo di ipotesi
Regola di decisione: si rifiuta H_0 se: $\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < -z_{\alpha/2}$, o $\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_{\alpha/2}$
Regola di decisione: si rifiuta H_0 se: $\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha$,
Regola di decisione si rifiuta H_0 se: $\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < -z_\alpha$,

VERIFICA DI IPOTESI SULLA **MEDIA DI DUE POPOLAZIONI**

Problemi tipo

CASO I	CASO II	CASO III
$H_0: \mu_x - \mu_y = d_0$ $H_1: \mu_x - \mu_y \neq d_0$	$H_0: \mu_x - \mu_y \leq d_0$ $H_1: \mu_x - \mu_y > d_0$	$H_0: \mu_x - \mu_y \geq d_0$ $H_1: \mu_x - \mu_y < d_0$

Occorre distinguere due casi fondamentali:

- **Campioni dipendenti** → campioni uguali, provenienti dalla medesima popolazione, vengono analizzati prima e dopo un trattamento
- **Campioni indipendenti** → campioni diversi vengono analizzati prima e dopo un trattamento

Popolazioni normali - campioni **dipendenti** **popolazione normale**

Sia $D_i = X_i - Y_i$ con $D_1, \dots, D_n \sim^{iid} N(\mu_D, \sigma_D^2)$ ed in particolare:

- $\bar{D} = \bar{x} - \bar{y}$
- $S_D^2 = S_X^2 + S_Y^2 - 2r_{XY}$

Statistica test	$T = \dfrac{\bar{D} - d_0}{S_D/\sqrt{n}} \approx^{\mu_D = d_0} T_{(n-1)}$
CASO I $H_0: \mu_x - \mu_y = d_0$ $H_1: \mu_x - \mu_y \neq d_0$	
Regola di rifiuto	• $\bar{D} < d_0 - t_{n-1,\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$ oppure $\bar{D} > d_0 + t_{n-1,\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$ • $\lvert T \rvert = \left\lvert \frac{\bar{D} - d_0}{\frac{S_D}{\sqrt{n}}} \right\rvert > t_{n-1,\frac{\alpha}{2}}$ • $p - value = 2P(T > \lvert t_{oss} \rvert) < \alpha$
CASO II $H_0: \mu_x - \mu_y \leq d_0$ $H_1: \mu_x - \mu_y > d_0$	
Regola di rifiuto	• $\bar{D} > d + t_{n-1,\alpha} \frac{S}{\sqrt{n}}$ • $T = \frac{\bar{D} - d_0}{\frac{S_D}{\sqrt{n}}} > t_{n-1,\alpha}$ • $p - value = P(Z > t_{oss}) < \alpha$
CASO III $H_0: \mu_x - \mu_y \geq d_0$ $H_1: \mu_x - \mu_y < d_0$	
Regola di rifiuto	• $\bar{D} < d + t_{n-1,\alpha} \frac{S}{\sqrt{n}}$ • $T = \frac{\bar{D} - d_0}{\frac{S_D}{\sqrt{n}}} < -t_{n-1,\alpha}$ • $p - value = P(Z < t_{oss}) < \alpha$

Varianza pooled

La principale differenza del caso in cui le varianze siano incognite ma assunte uguali con campioni indipendenti è l'utilizzo della **varianza pooled S_P**

$$S_P^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_x + n_y - 2}$$

Popolazioni normali - campioni **indipendenti** con **varianze incognite ma**

Statistica test	$T = \dfrac{(\bar{X} - \bar{Y}) - d_0}{S_P \sqrt{\dfrac{1}{n_x} + \dfrac{1}{n_y}}} \sim^{\mu_x - \mu_y = d_0} T_{(n_x + n_y - 2)}$
CASO I $H_0: \mu_x - \mu_y = d_0$ $H_1: \mu_x - \mu_y \neq d_0$	

<table>
<tr><td rowspan="5">assunte uguali</td><td rowspan="5"></td><td>Regola di rifiuto</td><td>• $\bar{X}-\bar{Y} < d_0 - t_{n_x+n_y-2,\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$ oppure $\bar{X}-\bar{Y} > d_0 + t_{n_x+n_y-2,\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$
• $|T| = \left|\frac{\bar{X}-\bar{Y}-d_0}{S_P\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}\right| > t_{n_x+n_y-2,\frac{\alpha}{2}}$
• $p-value = 2P(T > |t_{oss}|) < \alpha$</td></tr>
<tr><td colspan="2">CASO II
$H_0: \mu_x - \mu_y \leq d_0$
$H_1: \mu_x - \mu_y > d_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>• $\bar{X}-\bar{Y} > d_0 + t_{n_x+n_y-2,\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$
• $T = \frac{\bar{X}-\bar{Y}-d_0}{S_P\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} > t_{n_x+n_y-2,\alpha}$
• $p-value = P(Z > t_{oss}) < \alpha$</td></tr>
<tr><td colspan="2">CASO III
$H_0: \mu_x - \mu_y \geq d_0$
$H_1: \mu_x - \mu_y < d_0$</td></tr>
<tr><td>Regola di rifiuto</td><td>• $\bar{X}-\bar{Y} < d_0 - t_{n_x+n_y-2,\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$
• $T = \frac{\bar{X}-\bar{Y}-d_0}{S_P\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} < -t_{n_x+n_y-2,\alpha}$
• $p-value = P(Z < t_{oss}) < \alpha$</td></tr>
<tr><td>Popolazioni arbitrarie con grandi campioni (n≥30)</td><td colspan="3">Applicazione del teorema:
• Teorema centrale del limite → distribuzione normale

a) Campioni dipendenti con varianze non note e uguali

Statistica test: $Z = \frac{\bar{D}-d_0}{\frac{S_D}{\sqrt{n}}} \approx^{\mu_d=d_0} N(0,1)$
Le regole di rifiuto sono uguali a quelle viste per la singola popolazione normale varianza incognita, cambiando però i quantili con quelli della normale standard ($t_{n-1,\alpha} \to z_\alpha$)

b) Campioni indipendenti con varianze non note e differenti

Statistica test: $Z = \frac{(\bar{X}-\bar{Y})-d_0}{\sqrt{\frac{S_P^2}{n_x}+\frac{S_P^2}{n_y}}} \approx^{\mu_X-\mu_Y=d_0} N(0,1)$
Le regole di rifiuto sono uguali a quelle viste per la singola popolazione normale varianza incognita, cambiando però i quantili con quelli della normale standard ($t_{v,\alpha} \to z_\alpha$)</td></tr>
<tr><td>Popolazioni arbitrare</td><td colspan="3">Assunzioni:
• Indipendenza delle popolazioni
• Uguaglianza delle varianze (non note) delle popolazioni
• Assunzione di normalità (ampiezza campionaria non sufficiente per applicare il teorema centrale del limite)

Ci possiamo ricondurre al caso di popolazioni indipendenti con varianze assunte uguali → calcolo la varianza pooled e distribuzione t di student</td></tr>
</table>

Test sulla bontà di adattamento

Consideriamo un campione casuale di *n* osservazioni che possano essere classificate in K categorie o classi di misura. Indichiamo con $O_1, O_2, \dots, O_K$ il numero di osservazioni effettive.
Formuliamo un'ipotesi H_0 che ci permetta di descrivere i dati: in particolare supponiamo che l'ipotesi nulla specifichi $p_i \; con \; i = 1, 2, \dots, K \; t.c. \sum_{i=1}^{k} p_i = 1$ come la probabilità che un'osservazione appartenga all'i-esima categoria. Sia infine:

$$E_i = np_i$$

La frequenza attesa della i-esima classe sotto l'ipotesi H_0. Verificare che i dati effettivamente riscontrati si adattano ai valori attesi, sotto l'ipotesi H_0, vuol dire condurre un **test sulla bontà di adattamento**

$$H_0: p_i = p_i \; per \; ogni \; i$$
$$H_1: p_i \neq p_i \; per \; almeno \; un \; i$$

Abbiamo la tabella:

Categoria	1	2	…	K	Totale
Frequenze osservate	O_1	O_2	…	O_K	n
Probabilità (sotto H_0)	p_1	p_2	…	p_k	1
Frequenze attese (sotto H_0)	$E_1 = np_1$	$E_2 = np_2$	…	$E_K = np_K$	n

Per n elevato (grandi campioni) e $E_i > 5 \; \forall i$ avremo la seguente statistica test

$$\chi^2_{oss} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \approx^{H_0} \chi^2_{(k-1)}$$

E presenta la seguente regola di rifiuto:

- $\chi^2_{oss} > \chi^2_{k-1,\alpha}$
- $p - value = P(\chi^2 > \chi^2_{oss}) < \alpha$

Test chi-quadrato sulla indipendenza (o sulla associazione di due variabili)

Si considerino le variabili **categoriche** X ed Y. Vogliamo testare le seguenti ipotesi:

$$H_0: X \; e \; Y \; indipendenti$$
$$H_1: X \; e \; Y \; dipendenti$$

Sulla base dei dati campionari raccolti su n unità e sintetizzati in una **tabella di contingenza**

$X \setminus Y$	y_1	…	y_j	…	y_c	Tot.
x_1	O_{11}	…	O_{1j}	…	O_{1c}	R_1
…	…	…	…	…	…	…
x_i	O_{i1}	…	O_{ij}	…	O_{ic}	R_i
…	…	…	…	…	…	…
x_r	O_{r1}	…	O_{rj}	…	O_{rc}	R_r
Tot.	C_1	…	C_j	…	C_c	n

Dove O_{ij} è la frequenza congiunta assoluta osservata di unità nel campione.
Consideriamo la **frequenza attesa della cella** definibile come:

<table>
<tr><td></td><td>

$$E_{ij} = \frac{R_i C_j}{n}$$

Sotto l'ipotesi nulla di assenza di dipendenza, ci aspettiamo, infatti, di osservare una distribuzione *proporzionale* al totale delle osservazioni totali per riga e per colonna. Se n elevato e $E_{ij} < 5$ al massimo nel 20% delle celle, possiamo costruire la seguente **statistica test**:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx^{H_0} \chi^2_{(r-1)(c-1)}$$

Avremo la seguente **regola di rifiuto**

- $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$
- $p - value = P(\chi^2 > \chi^2_{oss}) < \alpha$

</td></tr>
<tr><td>Test sulla correlazione lineare (r)</td><td>

Si considerino due variabili **<u>numeriche</u>** X ed Y estratte da un campione casuale semplice di ampiezza n la cui distribuzione sia normale bidimensionale; si vuole indagare circa l'eventuale esistenza e forza, nella popolazione, di una relazione lineare tra le variabili considerate: l'obiettivo dell'analisi è il **coefficiente di correlazione lineare** nella popolazione e valutare **se esista una correlazione lineare tra le due variabili**

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var[X] * Var[Y]}}$$

Poiché disponiamo unicamente di un campione dobbiamo riferirci al **coefficiente di correlazione lineare campionario**

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 * \sum_i (Y_i - \bar{Y})^2}}$$

Si può dimostrare che è possibile ottenere la seguente **statistica test**

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim^{\rho=0} T_{(n-2)}$$

E avremo le seguenti regole di rifiuto:

CASO	REGOLA DI RIFIUTO
$H_0: \rho = 0$ $H_1: \rho \neq 0$	$\lvert t_{oss}\rvert > t_{n-2,\frac{\alpha}{2}}$ $p - value = 2P(T > \lvert t_{oss}\rvert) < \alpha$
$H_0: \rho = 0$ $H_1: \rho > 0$	$t_{oss} > t_{n-2,\alpha}$ $p - value = 2P(T > t_{oss}) < \alpha$
$H_0: \rho = 0$ $H_1: \rho < 0$	$t_{oss} < t_{n-2,\alpha}$ $p - value = 2P(T < t_{oss}) < \alpha$

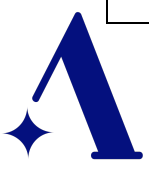
</td></tr>
</table>

<table>
<tr><th colspan="2">MODELLO DI REGRESSIONE LINEARE SEMPLICE</th></tr>
<tr><td>Terminologie e modello statistico</td><td>

Il modello di regressione lineare fornisce il valore atteso della variabile aleatoria Y quando X assume un particolare valore, aggiungendovi un margine di errore ε:

$$\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon_i}$$

</td></tr>
</table>



<table>
<tr>
<td></td>
<td>
Essa necessita alcune assunzioni di base:

La x è costante o è una realizzazione di una variabile aleatoria X, indipendente dalle componenti aleatorie di errore ε_i
Il valore atteso della variabile aleatoria Y è una funzione lineare della variabile aleatoria esplicativa X
I termini di errore sono variabili aleatorie con media 0 e varianza costante σ^2. La seconda condizione è detta omoschedasticità
$$E[\varepsilon_i] = 0 \quad Var[\varepsilon_i] = \sigma^2$$
Gli errori aleatori non sono correlati tra loro e quindi $E[\varepsilon_i \varepsilon_j] = 0$

Da queste assunzioni, seguono alcune importanti osservazioni:

L'ipotesi di media nulla implica $\boldsymbol{Cov(X, \varepsilon) = 0 \rightarrow}$ X e ε non sono correlati
L'ipotesi di varianza costante significa che la variabilità di ε non dipende dal valore di X

Determinazione dei parametri del modello

$E[Y|X = x] = E[b_0 + b_1 x + \varepsilon|X = x] = E[b_0 + b_1 x|X = x] + E[\varepsilon|X = x] = b_0 + b_1 x$
$b_0 = E[Y|X = 0]$ = medio di Y nella sottopopolazione con X = 0 → è l'intercetta all'origine
b_1 = variazione nella media di Y se X aumenta di una unità; rappresenta il coefficiente angolare → $E[Y|X = x + 1] - E[Y|X = x] = b_0 + b_1(x - 1) - (b_0 + b_1 x) = b_1$

Ulteriore conseguenza:

$Var[Y|X = x] = \sigma^2$ → la varianza di Y è costante in ogni sottopopolazione

</td>
</tr>
<tr>
<td>Metodo dei minimi quadrati</td>
<td>
Per determinare, da n coppie di osservazioni, gli stimatori dei coefficienti incogniti del modello, β_0 e β_1, possiamo ricorrere al modello dei minimi quadrati.
Dato un campione di ampiezza n:
$$(x_1, y_1) \dots (x_n, y_n)$$
Stimiamo β_0 e β_1 in modo tale che minimizzino:
$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$
Tali stime hanno espressione esplicita:
$$b_1 = \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{XY}\frac{s_Y}{s_X}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$
Interpretazione dei coefficienti della regressione (<u>se esplicativi</u>)

b_1 indica quanto <u>in media</u> aumenta y all'aumentare di una unità di x

Il segno positivo indica una relazione diretta (retta positivamente inclinata)
Il segno negativo indica una relazione inversa (retta negativamente inclinata)

b_0 indica quanto vale <u>in media</u> y, per $x = 0$; occorre valutare il suo senso osservando i dati del problema

</td>
</tr>
</table>

Capacità esplicativa della retta di regressione

Vogliamo sviluppare delle misure per indicare **quanto efficacemente la variabile X spiega il comportamento di Y** nel modello lineare.

Partiamo dalla retta di regressione stimata:

$$\hat{y} = b_0 + b_1 x$$

Questa consente di determinare:

- I **valori previsti**

$$\hat{y}_i = b_0 + b_1 x_i$$

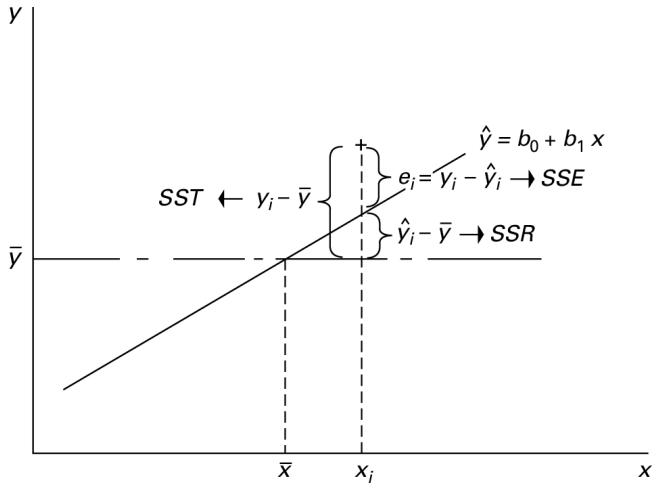
- I **valori residui**

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

Come è facile osservare i valori osservati y_i si discostano dai valori previsti $\hat{y}_i$ per un residuo e_i:

$$y_i = \hat{y}_i + e_i$$

Vi è quindi una parte che il modello non riesce a spiegare, ovvero e_i



Quello che stiamo facendo si chiama **scomposizione della devianza**. Questa può essere condotta anche considerando:

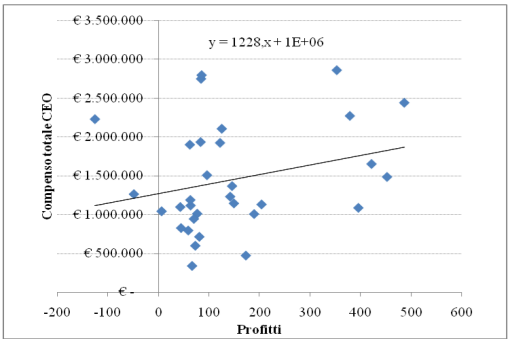
- $y_i - \bar{y}$ scarto della singola osservazione dalla media
- $\hat{y}_i - \bar{y}$ scarto del singolo valore previsto dalla media
- $y_i - \hat{y}_i$ scarto della singola osservazione dal corrispondente valore previsto

La loro analisi ci permette di scomporre lo scarto di una singola osservazione di Y dalla media nella somma algebrica degli altri due:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Elevano al quadrato ciascun membro dell'uguaglianza e sommando gli n addendi si ottiene:

	$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$ $$SST=SSR+SSE$$ Note le **varianze della variabile risposta Y e della variabile esplicativa X** possiamo giungere alle seguenti formule: $$SSR=b_1^2(n-1)s_X^2=b_1^2\sum(x_i-\bar{x})^2$$ $$SST=(n-1)s_Y^2=\sum(y_i-\bar{y})^2$$ $$SSE=SST-SSR=(n-1)s_Y^2-b_1^2(n-1)s_X^2$$
Coefficiente di determinazione	Si chiama **coefficiente di determinazione:** $$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$$ Rapp**resenta la percentuale di variabilità di Y spiegata dal modello di regressione**. R^2 aumenta in modo direttamente proporzionale alla dispersione della variabile indipendente X. **Varia tra 0 ed 1: valori più elevati indicano una migliore bontà del modello, in quanto meglio capace di spiegare la variabilità di Y** N.B.: **la bontà del modello mediante R^2 è sempre relativa**, nel senso che è corretto valutare due R^2, ed il maggiore indicherà il modello migliore; non è corretto valutare R^2 in senso assoluto Proprietà: • $0 \le R^2 \le 1$ • Maggiore è il suo valore, migliore è l'adattamento del modello ai dati Esiste una relazione tra R^2 ed il coefficiente di correlazione lineare $$R^2=r^2=\left(\frac{Cov(x,y)}{\sqrt{Var(x)}*\sqrt{Var(y)}}\right)^2$$
Stima della varianza del modello	Chiamiamo **stimatore dei residui** o **varianza dell'errore** o **varianza dei residui**: $$s_e^2=\frac{\sum_i e_i^2}{n-2}=\frac{\sum_i(y_i-\hat{y}_i)^2}{n-2}=\frac{SSE}{n-2}$$ Si chiama **errore standard** del modello: $$s_e=\sqrt{s_e^2}$$

Interpretazione grafica del modello (esempio)	Considerando la distribuzione dei valori osservati per la variabile dipendente e quella esplicativa riportata nel diagramma a dispersione, è possibile notare come **i punti non sembrino suggerire l'esistenza di una relazione lineare tra le due variabili.** Essendo, inoltre, la distanza tra i punti e la retta piuttosto ampio (accentuata dalla presenza di alcuni valori anomali) ci si aspetta di commettere un **errore di stima piuttosto elevato** e di ottenere, di conseguenza, un coefficiente di determinazione **R quadrato non molto elevato.**	
Proprietà degli stimatori dei minimi quadrati	**Teorema di GAUSS-MARKOV** Sotto le assunzioni di base del modello di regressione lineare, gli **stimatori dei minimi quadrati** sono i **migliori stimatori lineari e non distorti per il modello di regressione lineare** (stimatori BLUE) Valgono perciò le seguenti proprietà (σ^2 è non noto) **Valore atteso** $E[b_0] = \beta_0$ $E[b_1] = \beta_1$ **Varianza** $Var[b_0] = S_{b_0}^2 = s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)$ $Var[b_1] = S_{b_1}^2 = \frac{s_e^2}{\sum_i (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_X^2}$ Inoltre, fissa l'**assunzione di Normalità degli errori**, ovvero $\epsilon_1, \dots \epsilon_n \sim^{iid} N(0, \sigma^2)$, gli stimatori dei minimi quadrati hanno **distribuzione normale**: $b_0 \sim N(\beta_0, \sigma_{b_0}^2) \quad b_1 \sim N(\beta_1, \sigma_{b_1}^2)$ **Efficienza degli stimatori** **Uno stimatore è tanto più efficiente tanto più la sua varianza è piccola**. La varianza dello stimatore del coefficiente angolare è pari a $Var[b_1] = \frac{s_e^2}{\sum_i (x_i - \bar{x})^2}$, per cui, affinché $Var[b_1]$ sia piccolo è necessario che $\sum_i (x_i - \bar{x})^2$ sia elevato e cioè che la variabile indipendente sia dispersa intorno alla media.	
Inferenza sui parametri	**Intervalli di confidenza (assunzione di normalità degli errori)** Dato che σ^2 è incognito, stimiamo la varianza con s^2, la varianza campionaria; per b_1 e b_0 si ha che: $s_{B_1} = \sqrt{\frac{s_e^2}{\sum_i (x_i - \bar{x})^2}} = \sqrt{\frac{\boldsymbol{s_e^2}}{\boldsymbol{(n-1)Var(x)}}}$ $S_{B_0} = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)} = \sqrt{\boldsymbol{s^2}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}} + \frac{\boldsymbol{\bar{x}^2}}{\boldsymbol{(n-1)Var(x)}}\right)}$	

Ricorda che $\sum_i (x_i - \bar{x})^2 = (n-1)Var(x)$

Da cui determiniamo i pivot:

$$\frac{B_1 - \beta_1}{S_{B_1}} \sim T_{(n-2)} \quad \frac{B_0 - \beta_0}{S_{B_0}} \sim T_{(n-2)}$$

Da cui otteniamo gli intervalli di confidenza, a livello $1-\alpha$

$$IC_{1-\alpha}(\beta_1) = \left(b_1 \pm t_{n-2,\frac{\alpha}{2}} S_{B1}\right)$$

$$IC_{1-\alpha}(\beta_0) = \left(b_0 \pm t_{n-2,\frac{\alpha}{2}} S_{B0}\right)$$

Test d'ipotesi (assunzione di normalità degli errori)

	REGOLA DI RIFIUTO
CASO I $H_0: \beta_1 = b_1^*$ $H_1: \beta_1 \neq b_1^*$	$b_1 < b_1^* - t_{n-2,\frac{\alpha}{2}} S_{B_1} oppure\ b_1 > b_1^* + t_{n-2,\frac{\alpha}{2}} S_{B_1}$ $\lvert T \rvert = \left\lvert \frac{b_1 - b_1^*}{S_{B_1}} \right\rvert > t_{n-2,\frac{\alpha}{2}}$ $p - value = 2P(T > \lvert t_{oss} \rvert) < \alpha$
CASO II $H_0: \beta_1 \leq b_1^*$ $H_1: \beta_1 > b_1^*$	$b_1 > b_1^* + t_{n-2,\alpha} S_{B_1}$ $T = \frac{b_1 - b_1^*}{S_{B_1}} > t_{n-2,\alpha}$ $p - value = P(T > t_{oss}) < \alpha$
CASO III $H_0: \beta_1 \geq b_1^*$ $H_1: \beta_1 < b_1^*$	$b_1 < b_1^* - t_{n-2,\alpha} S_{B_1}$ $T = \frac{b_1 - b_1^*}{S_{B_1}} < -t_{n-2,\alpha}$ $p - value = P(T < t_{oss}) < \alpha$

Test di significatività

Un particolare test d'ipotesi da condurre è il ***test di significatività***, atto a verificare se **X sia una variabile significativa o meno**; se rifiuto l'ipotesi nulla, X è da considerarsi una variabile significativa. Tale test prevede le seguenti ipotesi:

$$\boldsymbol{H_0: \beta_1 = 0\ (non\ significativa)}$$
$$\boldsymbol{H_1: \beta_1 \neq 0\ (significativa)}$$

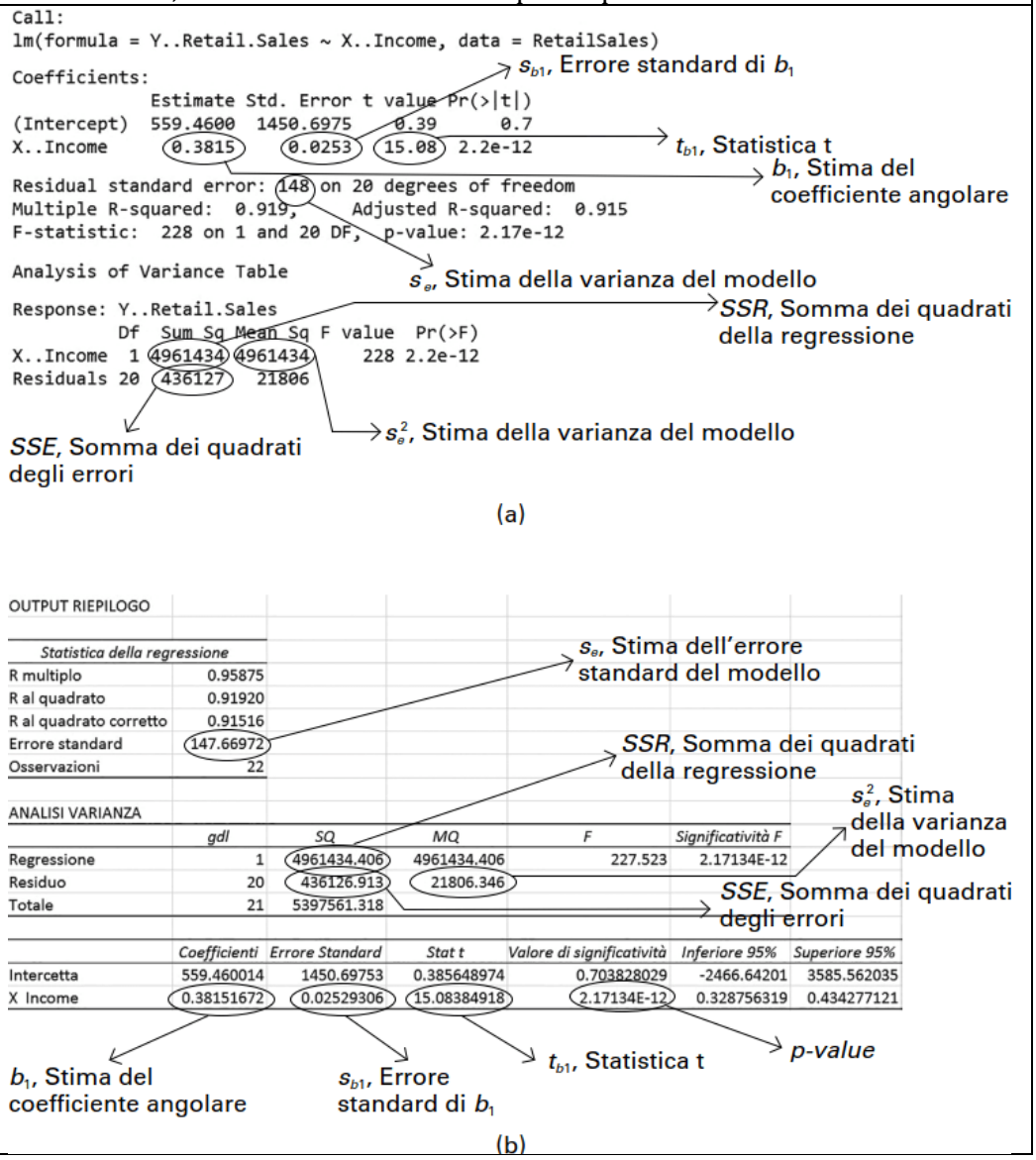
Per cui vale la seguente statistica test:

$$T = t - value = \frac{b_1}{S_{b_1}} \sim^{b_1=0} T_{(n-2)}$$

Valgono le seguenti regole di rifiuto:

$$\lvert T \rvert = \left\lvert \frac{b_1 - 0}{S_{b_1}} \right\rvert = \lvert t - value \rvert > t_{n-2,\frac{\alpha}{2}}$$
$$p - value = 2P(T_{n-2} > \lvert t_{oss} \rvert) < \alpha$$

	In **caso di rifiuto** si dirà che: • b_1 è significativamente diverso da 0 • ***X* è una variabile esplicativa statisticamente significativa** **<u>Relazione tra test di significatività e test di correlazione lineare (ρ)</u>** $$\rho = 0 \Leftrightarrow \beta_1 = 0$$
Previsione	Sono possibili due tipi di previsione: • Previsione puntuale del valore y_{n+1} risultate da una singola osservazione, x_{n+1} • Previsione del valore atteso condizionato $E[Y_{n+1}\|X = x_{n+1}]$, cioè della media della variabile risposta quando la variabile esplicativa assume il valore prefissato x_{n+1} **a) Stima puntuale** La **stima** è uguale per entrambi e, in particolare, avremo che: $$y_{n+1} = b_0 + b_1 x_{n+1}$$ **b) Stima per intervallo** **<u>Intervallo di confidenza per il valore atteso della stima</u>** $$IC_{1-\alpha}(E[Y_{n+1}\|X = x_{n+1}]) = \left(y_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right)$$ • Viene fornito x_{n+1} • Calcolo il centro dell'intervallo $y_{n+1} = b_0 + b_1 x_{n+1}$ • Calcolo $s_e = \sqrt{\frac{SSE}{n-2}}$ • Considero il termine sotto radice: ○ Il denominatore della seconda frazione può essere riscritto come $\sum_i (x_i - \bar{x})^2 = (n-1)Var(X)$ **<u>Intervallo di previsione per il valor singolo stimato</u>** $$IP_{1-\alpha}(Y_{n+1}) = \left(y_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right)$$ Si osserva che **l'intervallo di confidenza è più piccolo dell'intervallo di previsione ed entrambi sono centrati sul valore previsto *y***: con l'intervallo di confidenza prevediamo un valore medio; con l'intervallo di previsione una variabile aleatoria, che non ci permette di avere un intervallo più stretto a causa della sua aleatorietà SI possono ottenere ulteriori informazioni studiando la forma generali degli intervalli e tenendo presente il fatto che tanto maggiore è l'ampiezza intervallare, tanto minore è la previsione puntuale: 1. Più l'ampiezza n del campione è elevata, tanto minore è l'ampiezza dell'intervallo 2. Tanto maggiore è s_e^2, tanto più ampio è l'intervallo di confidenza: infatti tale misura è una stima della varianza degli errori

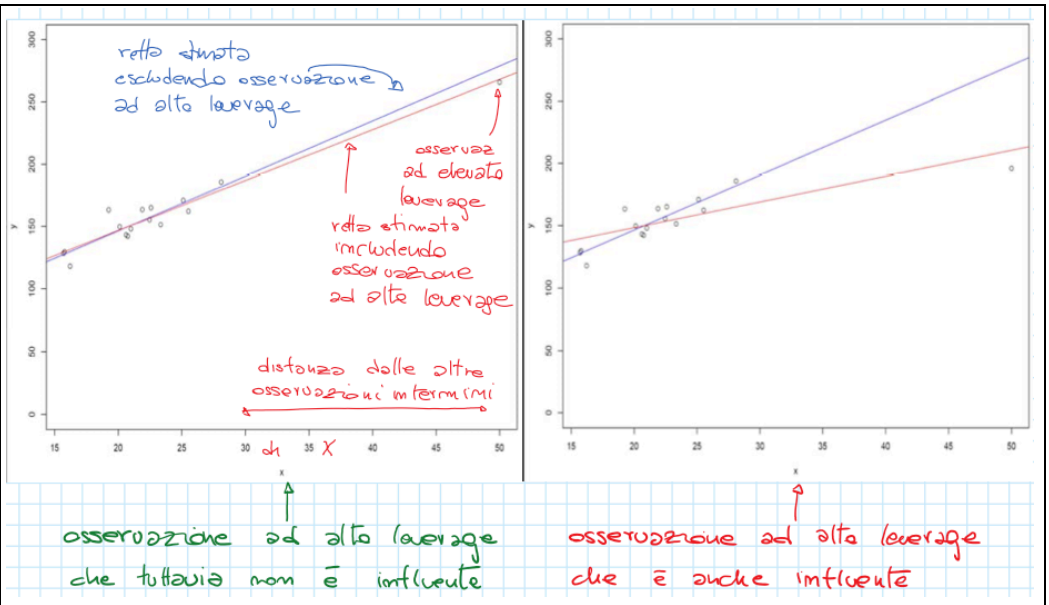
	3. Tanto maggiore è $\sum_i (x_i - \bar{x})^2$ e tanto minore risulta l'ampiezza dell'intervallo: infatti, si avranno più informazioni circa i valori più lontani dalla media 4. Se il valore per il quale vogliamo fare la previsione x_{n+1} si allontana dalla media, allora otterremo intervalli più ampi
Output software	

IL MODELLO DI REGRESSIONE MULTIPLA	
Introduzione e assunzioni di base	Il modello teorico della regressione multipla definisce la relazione tra una **variabile risposta Y** e un **insieme di variabili esplicative X_K** con k = 1,…; n rappresenta il numero totale delle osservazioni, mentre k rappresenta il numero totale delle variabili esplicative Avremo infatti che: $$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon_i}$$ Ci aspettiamo che la varianza di ε del modello di regressione semplice sia maggiore dell'ε del modello di regressione multipla **Assunzioni di base valide per ogni *i*** 

	1. Le x_{ij} sono **costanti** o **realizzazioni** di una **variabile aleatoria X**, indipendenti dalle componenti aleatorie di errore ε_i 2. Il valore atteso della variabile aleatorie Y è una **funzione lineare** delle variabili indipendenti $X_i \rightarrow E[Y\|X = x_k, k = 1,2, \dots]$ 3. I termini di errore sono variabili aleatorie con media 0 e varianza costante σ^2. La seconda condizione è detta **omoschedasticità** $$E[\varepsilon_i] = 0 \quad Var[\varepsilon_i] = \sigma^2$$ 4. Gli errori aleatori non sono correlati tra loro e quindi $\boldsymbol{E[\varepsilon_i \varepsilon_j] = 0}$ 5. Non è possibile trovare un insieme di coefficienti, tutti non nulli, $c_1, c_2, \dots, c_n$ tali che: $$c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_n x_n = 0$$ Ovvero **non esiste una relazione lineare tra le X_i**
Stima dei minimi quadrati	Anche nel caso della regressione multipla, le stime dei parametri vengono determinati mediante gli **stimatori dei minimi quadrati**; tali stimatori sono BLUE, ovvero sono i migliori stimatori lineari non distorti
Interpretazione della stima dei coefficienti	Il coefficiente di regressione stimato di una variabile, se significativo, rappresenta l'effetto sulla media della variabile Y dovuto ad una variazione unitaria di quella variabile, a parità dei valori delle altre variabili esplicative. È importante specificare che la variabile è "**significativa**" e l'effetto è presente "**in media**" e "**a parità delle altre variabili**", altrimenti la conclusione non è corretta
Scomposizione della varianza e indice R^2	Vale la scomposizione $$SST = SSR + SSE$$ Ovvero: $$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$ E l'indice R^2 (o coefficiente di determinazione) è definito come: $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SSR} \; con \; 0 \leq R^2 \leq 1$$ **R^2 misura la variabilità spiegata dal modello rispetto alla variabilità totale e fornisce un'informazione circa la bontà del modello stesso: tanto maggiore è R^2 tanto migliore sarà la spiegazione fornita dal modello** Vale la relazione per cui: $$R^2 = corr(y, \hat{y})^2$$ Ovvero il coefficiente di determinazione è uguale al quadrato della correlazione campionaria, al quadrato, tra i valori previsti ed osservati per y
Stima della varianza del modello	Una stima non distorta della varianza degli errori aleatori ε_i è chiamata **varianza dell'errore** o **varianza dei residui** è: $$s_e^2 = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$ Dove K rappresenta il numero di variabili indipendenti del modello.

	Si chiama **errore standard della stima** $$s_e = \sqrt{s_e^2}$$
Coefficiente di determinazione corretto	R^2 non può diminuire se nuove variabili esplicative vengono incluse nel modello: il confronto tra modelli basati sul valore di **R^2 porterà sempre a scegliere il modello che include tutte le possibili variabili esplicative disponibili**: $$R^2_{multipla} \geq R^2_{semplice}$$ Per **confrontare due modelli di regressione multipla (o un modello di regressione multipla con un modello di regressione semplice) con un diverso numero di variabili esplicative** ($\boldsymbol{K \neq K^*}$) è possibile usare il coefficiente di determinazione corretto, che tiene conto della significatività statistica delle variabili inserite nel modello $$\overline{R^2} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$ Con una formula alternativa risulta che: $$\overline{R^2} = 1 - (1-R^2) * \frac{n-1}{n-k-1}$$ Si utilizza questo indice per **bilanciare la riduzione, anche se minima, della somma dei quadrati degli errori che si determina con l'aggiunta di variabili esplicative non rilevanti**. L'indice R^2 fornisce quindi un migliore strumento per confrontare modelli di regressione multipla con un numero diverso di variabili indipendenti. Vale sempre che $\overline{R^2} < R^2$
Coefficiente di correlazione multipla	$$R = r(\hat{Y}, Y) = \sqrt{R^2}$$ Ci permette di misurare la relazione lineare tra la variabile dipendente e le variabili indipendenti
Interpretazione dei parametri e inferenza sui **parametri singolarmente considerati**	Sia dato il modello teorico di regressione multipla e siano $b_0, b_1, \dots, b_k$ le stime dei minimi quadrati dei paramenti del modello teorico e siano s_{b_0}, s_{b_1} … le stime degli errori standard degli stimatori dei minimi quadrati. Allora possiamo costruire la seguente statistica test: $$T_{b_i} = \frac{b_i - \beta_i^*}{s_{b_i}} \sim T_{(n-k-1)}$$ Avremo che l'intervallo di confidenza: $$IC_{1-\alpha}(\beta_i) = \left(b_i \pm t_{n-k-1,\frac{\alpha}{2}} s_{b_i}\right)$$ Sarà anche possibile condurre il test d'ipotesi, supponendo la ***normalità degli errori***, sui parametri singolarmente considerati (**T-test**) $$H_0: \beta_i = \beta_i^* \quad H_1: \beta_i \neq \beta_i^*$$ Valgono le seguenti regole di rifiuto:

	$$\|T\| = \left\|\frac{b_1}{S_{B_1}}\right\| = \|t - value\| > t_{n-2,\frac{\alpha}{2}}$$ $$p - value = 2P\left(T^{(n-k-1)} > |t_{oss}|\right) < \alpha$$ I casi II e III sono analogamente risolvibili N.B.: t-value viene spesso indicato con **Stat t** Un **test di significatività del parametro singolarmente considerato** si conduce semplicemente imponendo: $$H_0: \beta_i = 0\ (non\ significativo)$$ $$H_1: \beta_i \neq 0\ (significativo)$$ Le regole di rifiuto sono le medesime del modello generale
Verifica della significatività di **tutti i parametri congiuntamente considerati** (cenni)	Conduciamo un **test di significatività generale del modello** (**F-test**): $$H_0: \beta_1, \beta_2, \ldots = 0\ (il\ modello\ non\ \text{è}\ significativo)$$ $$H_1 = almeno\ un\ \beta_i \neq 0\ (almeno\ una\ variabile\ esplicativa\ \text{è}\ rilevante \rightarrow il\ modello\ \text{è}, nel\ complesso, significativo)$$ La statistica test (**statistica F**) è: $$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{s_e^2} = \frac{MSR}{MSE} \sim^{H_0} F_{k,n-k-1}$$ E la regola di rifiuto è: $$F > F_{k,n-k-1,\alpha}$$ $$p - value < \alpha$$
Situazioni problematiche nell'analisi di regressione: l'approccio dell'**analisi grafica dei residui**	Dopo aver costruito il modello di regressione lineare è buona norma condurre l'**analisi grafica dei residui** che consente di determinare la **bontà del modello e per verificare se sono soddisfatte le ipotesi standard della regressione**. Condurre tale analisi significa costruire: • Un **istogramma** che permetta di verificare se la **distribuzione dei residui sia simmetrica e uniforme** • Costruire un **diagramma a dispersione** che permette di osservare eventuali **osservazioni ad alto leverage** o **outliers** **Osservazioni ad alto leverage**

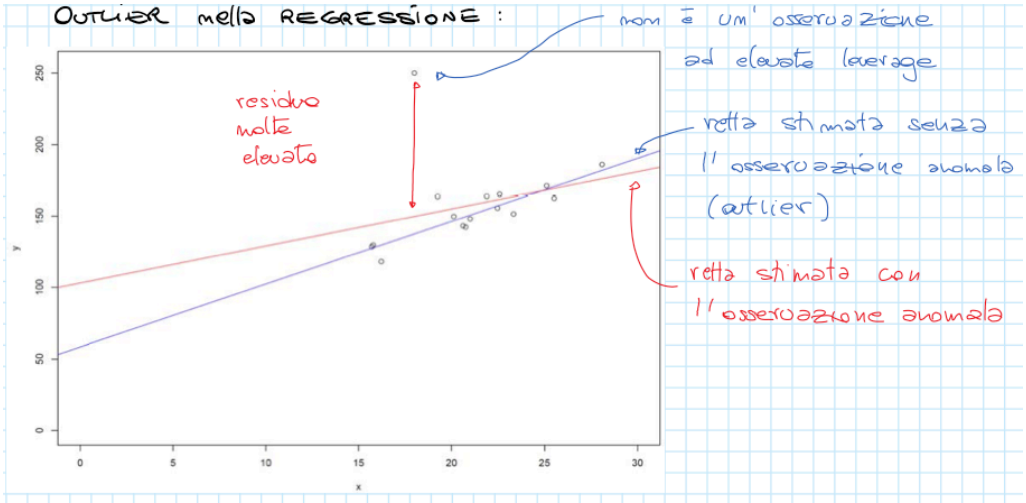


Un'osservazione ad alto leverage è quindi differente/distante dalle altre in termini di valori delle variabili esplicative: **presentano un'ordinata che si discosta notevolmente dalle altre.**

Per verificare che sia anche influente è possibile stimare i modelli con o senza l'osservazione

<u>Outlier nella regressione</u>

È una osservazione con valore elevato del corrispondente residuo, ovvero non ben descritta dal modello; in particolare, essi sono **punti la cui ascissa si discosta notevolmente**



La sua influenza può essere valutata come nel caso delle osservazioni ad alto leverage

Multicollinearità	**È l'esistenza di forti (ma non perfette) relazioni lineari tra le variabili esplicative nel dataset osservato.** Come ricordiamo (è la condizione aggiuntiva) non esistono dei coefficienti tali che: $$c_0 + c_1 x_1 + \cdots + c_k x_k = 0$$

	Escludendo l'esistenza di relazioni lineari esatte tra le variabili esplicative. Tuttavia, relazioni di forte intensità, seppur non esatte, generano problemi: $$c_0 + c_1 x_1 + \cdots + c_k x_k \approx 0$$ Alcuni **<u>indizi</u>** di multicollinearità sono: • I **coefficienti di regressione sono molto diversi**, anche dal punto di vista del segno, da quelli che ci si potrebbe attendere secondo le **teorie economiche o l'esperienza** (logica) • I coefficienti delle variabili che si ritengono rilevanti hanno una **statistica T molto bassa**, trattandosi quindi di valori nulli • Le **statistiche T** di tutti i coefficienti sono **basse** (i **coefficienti singolarmente considerati** sono ritenuti **non significativi**), ma il valore della **statistica F** del modello indica la complessiva **significatività statistica** del **modello** stesso • Vi sono forti correlazioni tra le coppie di variabili indipendenti: in un modello di **regressione semplice** una **variabile era ritenuta significativa**; inserendolo nel modello di **regressione multipla** la stessa variabile **perde la sua significatività** **<u>Effetti</u>** Gli stimatori dei minimi quadrati sono ancora non distorti ma la **varianza e lo standard error sono inflazionati**, ovvero di valore manifestamente elevato. L'aggiunta di variabili collineari porta a creare un modello più complesso ma non maggiormente esplicativo (ridondante) **<u>Come eliminarla</u>** Un *suggerimento generale* può essere quello di **eliminare semplicemente una delle variabili ritenute tra loro collineari**

R/RADIANT	
Attribuzione di nome ad un oggetto	Oggetto semplice x <- log(3) Oggetto complesso z <- c(2, 4, 3, 1)
Funzioni	Media mean(z) Matrice matrix(z, nrow=2, ncol=2, byrow=TRUE)
Caricare un dataset	Selezionare la cartella di lavoro File → Cambia Directory Caricare il file load(file="nome_file.RData")
Funzioni da file	Funzione str: permette di indagare la natura dell'oggetto str(nome_file) Funzione media mean(nome_file$nome_variabile) Funzione summary: ci permette di visualizzare gli indicatori di sintesi che il software ritiene siano appropriati per la specifica variabile) summary(nome_file$nome_variabile)

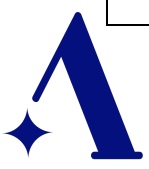
Rappresentazioni grafiche	**Funzione table**: costurisce una distribuzione di frequenza assoluta table(nome_file$nome_variabile) Grafico a torta: **funzione pie(table)** pie(table(nome_file$nome_variabile)) pie(table(nome_file$nome_variabile)/length(nome_file$nome_variabile)) Boxplot per una singola variabile: **funzione boxplot** boxplot(nome_file$nome_variabile)

Radiant: operazioni di base	Radiant è caratterizzato da un'interfaccia di menu a tendina. Mediante il pulsante *load* è possibile caricare i dataset che intendiamo analizzare
	Possiamo usare:
	• ***View*** per analizzare globalmente il dataset
	• ***Visualize*** per creare le rappresentazioni grafiche; dal menu che viene visualizzato possiamo settare tutte le caratteristiche della rappresentazione
	• ***Pivot*** per generare le distribuzioni di frequenza assolute; cliccando si *normalize by:*
	  ○ *Total* possiamo generare le frequenze congiunte relative
	  ○ *Row* possiamo generare le frequenze condizionate per riga
	  ○ *Column* possiamo generare le frequenze condizionate per colonna
	• ***Explore*** per generare gli indici di sintesi delle diverse variabili, dal sottomenu *apply function*

Analisi univariata e istogrammi	**Radiant** permette di costruire istogrammi con classi di uguale ampiezza: Data → Visualize: 1. Plot type → distribution 2. X-variabile → selezionare la variabile d'interesse 3. Create plot Lo slider Number of bins permette di scegliere il numero di classi da usare Radiant permette di costruire anche istogrammi per distribuzioni condizionate utilizzando il comando *Facet column* e selezionando la variabile condizionante Con **R/RStudio** è possibile costruire istogrammi con classi di diversa ampiezza: occorre usare la funzione ***hist()*** che si compone dei seguenti argomenti: • nome_file$variabile • main = "titolo istogramma" • xlab = "nome asse x" • ylab = "nome asse y" • breaks = c(0, estremo1, estremo2, estremo3, …, estremoN) **R** ci permette di creare rapidamente anche i boxplot relativi ad una sola variabile con la **funzione boxplot** boxplot(nome_file$nome_variabile)
Analisi bivariata	Variabili categoriche Radiant ci permette di creare tabelle a doppia entrata mediante il tab *Pivot*, selezionando le due variabili da analizzare nel box *Categorical variables* e cliccando sul pulsante *Create pivot table*: abbiamo ottenuto le distribuzioni assolute • Con il comando ***Normalize by: total*** possiamo ottenere la distribuzione relativa; aggiungendo una spunta accanto a *Percentage* possiamo visualizzare la distribuzione percentuale • Con il comando ***Normalize by: column/row*** otteniamo la distribuzione condizionata secondo la variabile disposta in colonna/riga

	• Cliccando su ***Show plot*** è possibile visualizzare il diagramma a barre accostate; con *plot type* → *fill* otteniamo il grafico a barre sovrapposte <u>Variabili numeriche</u> Dal menu Basics → Correlation possiamo selezionare le variabili da analizzare e cliccare sul checkbox *Show covariance* matrix. Nella tab *Polt* sono visualizzabili gli scatterplot relativi
Probabilità	Le probabilità possono essere valutate mediante il menu *Basics* In ***Probability calculator*** possiamo calcolare le probabilità da varie distribuzioni selezionabili in *Distribution* Possiamo lavorare con *Input type*: • ***Values*** che definisce i valori estremi dell'intervallo; da qui possiamo calcolare l'area sottesa alla curva tra gli estremi dell'intervallo, ovvero la probabilità. In pratica, ci permette di trovare la probabilità p per cui: $$P(\alpha \leq X \leq \beta) = p$$ • ***Probability*** ci permette di definire i percentili a cui è associata l'area sottesa sulla coda sinistra. In pratica, ci permette di trovare la x per cui: $$P(X \leq x) = 0{,}95$$
Intervalli di confidenza	È possibile trovare gli intervalli di confidenza utilizzando, dal menu *Basics*, la sezione relativa al parametro d'interesse (***Single mean***, ***Compare means***, ***proportions***…) Occorre eseguire un *test* ***Two Sided*** ed impostare il livello di confidenza con lo slider. Il risultato si può osservare sotto le percentuali nel tabulato, che indicano gli estremi dell'intervallo
Inferenza sulla media di una popolazione normale	R e Radiant mettono a disposizione comandi solo nel caso di **varianza non nota**; per farlo andiamo nel menu Basics → Single mean A questo punto: • Selezioniamo la variabile da analizzare nel box *Variable* • Scegliere il tipo di ipotesi alternativa da testare nel box *Alternative hypothesis* con valori possibili *Two sided, Less than, Greather than* • Scegliamo il livello di confidenza in *Confidence level* • Indichiamo il valore da testare sotto ipotesi nulla in *Comparison value* Ci viene restituito sia l'intervallo di confidenza sia il p-value; valori molto piccoli di p-value vengono indicati con la dicitura < 0.001
Inferenza sulla proporzione di successi da una popolazione bernoulliana	Basics → Single proportion • Selezioniamo la variabile da analizzare nel box *Variable* • Scegliere il tipo di ipotesi alternativa da testare nel box *Alternative hypothesis* con valori possibili *Two sided, Less than, Greather than* • Scegliamo il livello di confidenza in *Confidence level* • Indichiamo il valore da testare sotto ipotesi nulla in *Comparison value* Occorre sempre selezionale il *test type: Z-test*
Inferenza sul confronto tra medie di popolazioni normali	Ci concentreremo sul caso di **varianze non note e diverse** <u>Campioni indipendenti</u> Basics → Compare means • *Select a factor o numeric variable* inseriamo la variabile che identifica i gruppi di cui vogliamo confrontare le medie (tipo factor)

	• *Nemeric variable* selezioniamo la variabile numerica di cui vogliamo confrontare le medie • *Alternative hypothesis* indichiamo che tipo di ipotesi vogliamo testare (Two sided, less/greater than) • Inseriamo il *Confidence level* • Spuntiamo il box *Show additional statistics* • *Sample type → independent* • *Test type → t-test* <u>Campioni dipendenti</u> La procedura è uguale ma dobbiamo selezionare il box *Sample type → paired*
Indice di correlazione lineare	L'indice di correlazione lineare costituisce lo strumento principale per valutare l'intensità dell'associazione lineare tra due variabili numeriche. Basics → **Correlation** Ci vengono riportati l'indice di correlazione lineare e il p-value: bassi valori del p-value implicano un'associazione altamente significativa
Bontà di adattamento	Consente di valutare se una distribuzione di probabilità osservata per una variabile aleatoria discreta (con supporto finito) possa essere considerata "compatibile" con una distribuzione di probabilità teorica definita a priori Basics → **Goodness of fit** • La variabile nel data set che contiene i dati con cui effettuare l'analisi (box *Select a categorical variable*) • Le probabilità che definiscono la distribuzione teorica con cui confrontare quella osservata (box Probabilities); queste probabilità devono essere inserite come numeri compresi tra 0 e 1, la cui somma deve essere pari a 1 e il cui numero deve essere pari alle categorie osservate per la variabile indicata nel box precedente • Il tipo di output da mostrare (box *Observed* per le frequenze osservate, *Expected* per quelle attese sotto l'ipotesi nulla, *Chi-squared* per il contributo al calcolo dell'indice χ^2 di ogni categoria)
Test d'indipendenza	Basics → **Cross-tabs** • Il nome delle variabili che contengono i dati su cui vogliamo effettuare il test di indipendenza (due box successivi, entrambi denominati *Select a categorical variable*; consigliamo di selezionare la variabile di riga nel primo box e quella di colonna nel secondo) • Il tipo di output da mostrare, ovvero le frequenze osservate (*Observed*), quelle attese sotto l'ipotesi nulla di indipendenza (*Expected*), i contributi al calcolo dell'indice χ^2 di ogni cella della tabella (*Chi-squared*), le radici quadrate dei medesimi contributi (*Deviation std.*), le frequenze condizionate date le righe, quelle condizionate date le colonne o quelle congiunte relative (rispettivamente *Row percentages*, *Column percentages e Table percentages*).
Modello di regressione lineare semplice	La funzione principale per stimare il modello di regressione lineare in R è ***lm()*** che presenta la seguente sintassi: $$reg1 <- lm(y \sim x, data = nome\_dataset)$$ Possiamo assegnare un nome a questa funzione e successivamente impiegare la funzione



	$\boldsymbol{summary}(reg1)$
	
	Questa funzione ci restituisce:
	• Nella sezione Residuals una **sintesi dei residui del modello**
	• Nella sezione Coefficients fornisce le **stime dei coefficienti** del modello insieme ai relativi **p-value**
	• Nelle ultime tre righe vengono forniti:
	  ○ La stima di σ (***Residual standard error***)
	  ○ L'indice R^2 (***Multiple R-squared***)
	  ○ Il **test *F*** (ultima riga)
	
	Per ottenere una visualizzazione grafica possiamo usare il seguente codice:
	
	$\boldsymbol{plot}(y{\sim}x, data = nome\_dataset)$
	$\boldsymbol{abline}(reg1, lwd = 2, col = \text{"blue"})$
	
	R ci fornisce solamente i p-value per i test sui due coefficienti ma non gli **intervalli di confidenza per entrambi i coefficienti**; questi possiamo ottenerli:
	
	$\boldsymbol{confint}(reg1, level = \alpha)$
	
	Per ottenere la **scomposizione della varianza** usiamo
	
	$\boldsymbol{anova}(reg1predict)$

Modello di regressione lineare multipla	Ancora una volta usiamo la funzione *lm( )* separando le variabili con un + $reg2 <- lm(y{\sim}x_1 + x_2 + \cdots + x_n, data = comp)$ $summary(reg2)$ Come nel modello semplice possiamo usare le funzioni plot( ) e anova( )
Inferenza sui coefficienti del modello	Usiamo la funzione ***confint()*** per ottenre l'intervallo di confidenza (livello standard = 0.95) per i parametri β_0 e β_1 lm.fiori <- lm(height ~ distance, data=fiori) confint(lm.fiori, level = 0.9)
Previsioni per un modello di regressione linare	Uno dei motivi della popolarità del modello di regressione lineare consiste nella possibilità di calcolare previsioni in relazione a scenari che non sono stati necessariamente osservati nel campione. La funzione da usare in R è ***predict( )*** con i seguenti argomenti: • **object**, ovvero l'oggetto restituito dalla funzione lm() contenente i risultati della stima • **newdata**, che indica un nuovo data frame contenente i valori delle variabili indipendenti da utilizzare per il calcolo delle previsioni • **interval**, che può assumere solo tre possibili valori, ovvero ***confidence*** se si desidera calcolare gli **intervalli di confidenza** per la previsione del valore medio, ***prediction*** se si desidera calcolare gli in**tervalli di previsione per i valori individuali**, *none* se non si vuole calcolare nessun intervallo • **level**, il quale consente di specificare il livello di confidenza per il calcolo degli intervalli di confidenza o previsione Previsione di Y dato X

	$predict(object, newdata = data.frame(variabile = numero))$ N.B.: per valutare su una previsione è affidabile dobbiamo valutare la variabile esplicativa richiesta è all'interno del range di valori della variabile: per farlo, controlliamo gli estremi con la funzione *summary(variabile)* <u>Previsione intervallare per singolo valore</u> $predict(object, newdata = data.frame(variabile = numero), interval = \text{"prediction"}, level = 0.90)$ <u>Confidenza intervallare per media dei valori</u> $predict(object, newdata = data.frame(variabile = numero), interval = \text{"confidence"}, level = 0.90)$