COMPLEMENTI DI ANALISI DEI SEGNALI

AA 2021-22

Agostino Accardo (accardo@units.it)

Parte I - Analisi statistica di dati biomedici

Parte II – Analisi non lineare di segnali biom

Parte III – modulo Ajcevic

Parte I - Analisi statistica di dati biomedici

Testi di riferimento:

Altman – Practical statistics for medical research

Steel & Torrie – Principles and procedures of statistics

Di Orio – Statistica medica

slide in: www.units.it/accardo

CONTENUTI DELLA PARTE I

- Strumenti statistici: caratteri, modalità, rango, frequenza
- Parametri statistici: media, varianza, mediana, momenti, ecc.
- Richiami sulle principali distribuzioni di probabilità
- Test di normalità
- Inferenza statistica, stimatori corretti ed efficienti, intervallo di confidenza
- Dimensione ottima del campione
- Verifica delle ipotesi
- Test parametrici e non parametrici
- Test su una media: z-test, t-test, test binomiale, test del segno e di Wilcoxon, $\chi 2$ -test, Kolmogorov
- Test su due campioni: z-test, t-test, Fisher, Mann-Whitney, tabelle di contingenza RxC
- Test su più campioni: ANOVA, Bartlett, Kruskal-Wallis

STATISTICA MEDICA: offre metodologia per analizzare quantitativamente fenomeni inerenti la medicina

2 APPROCCI:
- **POPOLAZIONE => EPIDEMIOLOGICO**
- **DIAGNOSI/PROGNOSI/TERAPIA => CLINICO-SPERIMENTALE**

MOMENTI ESSENZIALI:
- **SCELTA DELLA/E VARIABILE/I**
- **DETERMINAZIONE DELLA POPOLAZIONE DI RIFERIMENTO E DEL CAMPIONE DA CUI INFERIRE**
- **SCELTA DELLA DIMENSIONE DEL CAMPIONE E SUA ESTRAZIONE**

STATISTICA DESCRITTIVA:

insieme dei metodi che riguardano raccolta, presentazione e sintesi di un insieme di dati per descriverne le caratteristiche essenziali

STATISTICA INFERENZIALE:

insieme dei metodi con cui si possono elaborare i dati dei campioni per dedurne omogeneità o differenze nelle caratteristiche analizzate

PROCEDURA CORRETTA:

PRIMA DI PROGETTARE UNA RICERCA ANALIZZARE QUALE METODOLOGIA STATISTICA UTILIZZARE E QUINDI RACCOGLIERE I DATI

NO PRIMA I DATI E POI SCEGLIERE IL METODO STATISTICO

-- SINGOLO, DOPPIO O TRIPLO CIECO

STUDI DI:

- COORTE (LONGITUDINALI) => NEL TEMPO SUGLI STESSI SOGGETTI (prospettico o retrospettivo)

- CROSS-SECTIONAL (TRASVERSALI) => 1 MISURA PER SOGGETTO DI UNA POPOLAZ. IN UN DET. MOMENTO

- CASI-CONTROLLO (LONGITUDINALI) => 1 FATTORE SU 2 GRUPPI

STRUMENTI STATISTICI

CARATTERE = VARIABILE
MODALITA' = VALORE

CARATTERE	QUALITATIVO (=> CONTEGGI)	QUANTITATIVO
NOMINALE	NON ORDINATO	
ORDINALE	ORDINATO O ORDINABILE*	
A INTERVALLI		NUMERABILE**

* **RANGO:** POSIZIONE OCCUPATA IN UN INSIEME ORDINATO
** DISCRETI / CONTINUI

RILEVAZIONI:

- SALTUARIE / CONTINUE
- PUBBLICHE / PRIVATE
- PARZIALI (CAMPIONI) / TOTALI (POPOLAZIONE)

RICHIEDONO:

- PERIODO DI RACCOLTA DATI
- GRADO DI PRECISIONE
- SCHEDE/QUESTIONARI X COSTRUIRE TABELLE
- IPOTESI STATISTICHE
- .....

PRODUCONO DATI STATISTICI DIVISI IN CLASSI O TABELLE DI CLASSI DI MEDESIMA O DIFFERENTE AMPIEZZA (MAX 10-20 CLASSI NON SOVRAPPOSTE)

ESPRESSIONE DEL DATO STATISTICO

FREQUENZA DI UNA MODALITA' :

RELATIVA= n_i/N ASSOLUTA=n_i DENSITA'=n_i/α_i

α_i = Ampiezza della classe

INTENSITA' = VALORE

L'INSIEME DELLE FREQUENZE O INTENSITA' =

DISTRIBUZIONE DI FREQUENZE/INTENSITA'

CLASSI DI X	CLASSI DI Y				TOTALI
	Y_1	Y_2	.....	Y_k	
X_1	n_{11}	n_{12}	...	n_{1k}	n_{10}
X_2	n_{21}	.....			
....	.....				.....
X_h	n_{h1}	n_{h2}	....	n_{hk}	n_{h0}
TOTALE	n_{01}	n_{02}	....	n_{0k}	N

n_{ij} = frequenze; n_{0j} e n_{i0} = frequenze marginali

DISTRIBUZIONE DI FREQUENZE/INTENSITA'

	y_1	y_2	…	…	…	y_k	Totale
x_1	n_{11}	n_{12}	…	…	…	n_{1k}	n_{1o}
x_2	n_{21}	n_{22}	…			n_{2k}	n_{2o}
.	…	…				…	.
.	…	…		$\boldsymbol{n_{ij}}$		…	.
.	…	…				…	.
x_n	n_{n1}	n_{n2}	…	…	…	n_{nk}	n_{no}
Totale	n_{o1}	n_{o2}	…	…	…	n_{ok}	N

$x_1, \ldots, x_n$ possibili valori di x (o classi)
$y_1, \ldots, y_k$ possibili valori di y (o classi)
n_{ij} frequenze (o intensità)
n_{oj} e n_{io} frequenze marginali (tengono conto di una sola variabile)

Le variabili continue sono caratterizzate dalla funzione densità
Le variabili discrete sono caratterizzate dalla funzione di frequenza

ESEMPIO:

Conteggio del numero di foglie (variabile discreta) nate su 45 rami di uguale lunghezza di una pianta in un dato intervallo di tempo :

5 6 3 4 7 2 3 2 3 2 6 4 3 9 3 2 0 3 3 4 6 5 4 2 3 6 7 3 4 2 5 1 3 4 3 7 0 2 1 3 1 5 0 4 5

Distribuzione di frequenze assolute e relative delle foglie:

classe (xi)	0	1	2	3	4	5	6	7	8	9
freq. assol. (ni)	3	3	7	12	7	5	4	3	0	1
freq. rel. (fi)	0,07	0,07	0,15	0,27	0,15	0,11	0,09	0,07	0,0	0,02
freq.cumulata	0,07	0,14	0,29	0,56	0,71	0,82	0,91	0,98	0,98	1

Quante classi di frequenza costruire?

- da un minimo di 4-5 ad un massimo di 15-20 in funzione del numero di osservazioni

Infatti:

- se il numero di classi è troppo basso: perdita d'informazione sulle caratteristiche della distribuzione rendendola non significativa
- se il numero di classi è troppo alto: dispersione dei valori e perdita della forma della distribuzione

Non è necessario costruire intervalli uguali; ma la loro rappresentazione grafica ed il calcolo dei parametri fondamentali esigono alcune avvertenze non sempre intuitive

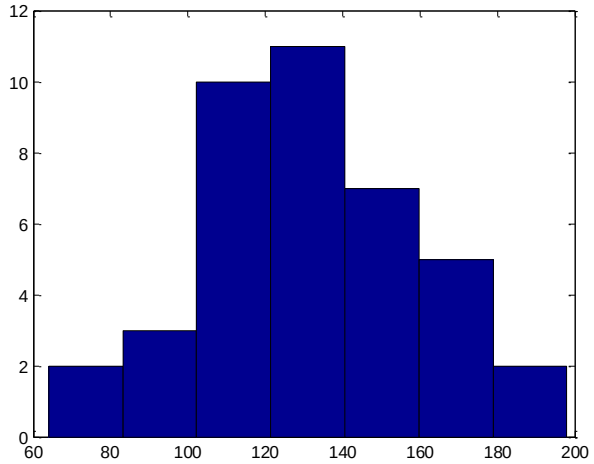
ESEMPIO 2:

Raggruppamento in classi di una variabile continua: altezza (cm) di 40 piante:
107 83 100 128 143 127 117 125 64 119 98 111 119 130 170 143 156 126 113 127
130 120 108 95 192 124 129 143 198 131 163 152 104 119 161 178 135 146 158 176

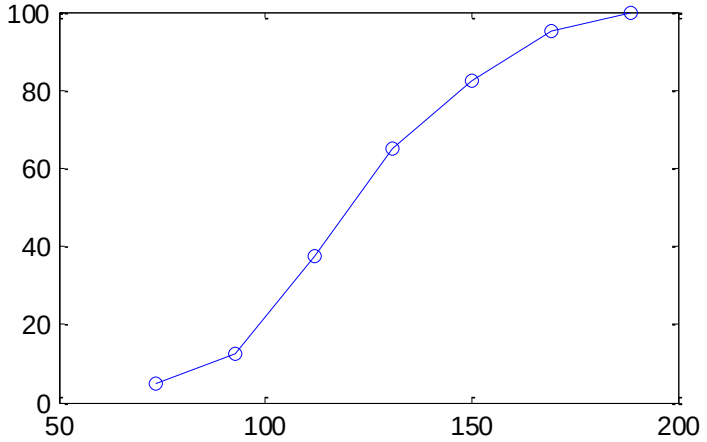
Distribuzione di frequenze assolute e relative (%) dell'altezza delle 40 piante :

classe (xi)	60-79	80-99	100-119	120-139	140-159	160-179	180-199
freq. ass. (ni)	1	3	10	12	7	5	2
freq. rel. (fi)	2,5	7,5	25	30	17,5	12,5	5
freq. cumul.	2,5	10	35	65	82,5	95	100

Nota: la classe iniziale e terminale non devono essere aperte (es.: < 80 quella iniziale; >180 quella finale), poiché si perderebbe l'informazione del loro valore minimo e massimo e quindi del valore centrale (indispensabili per calcolare la media e gli altri parametri da essa derivati)



Curva frequenza cumulata



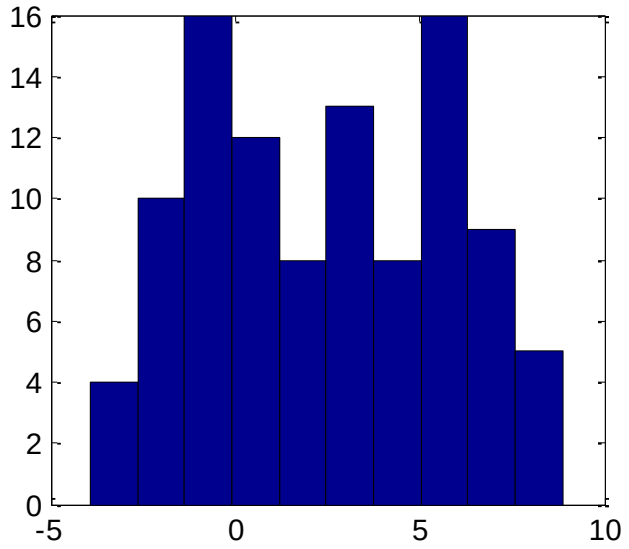
RAPPRESENTAZIONI

- ANALITICHE ESPRIMONO LEGAMI FUNZIONALI/MODELLI INTEPRETATIVI

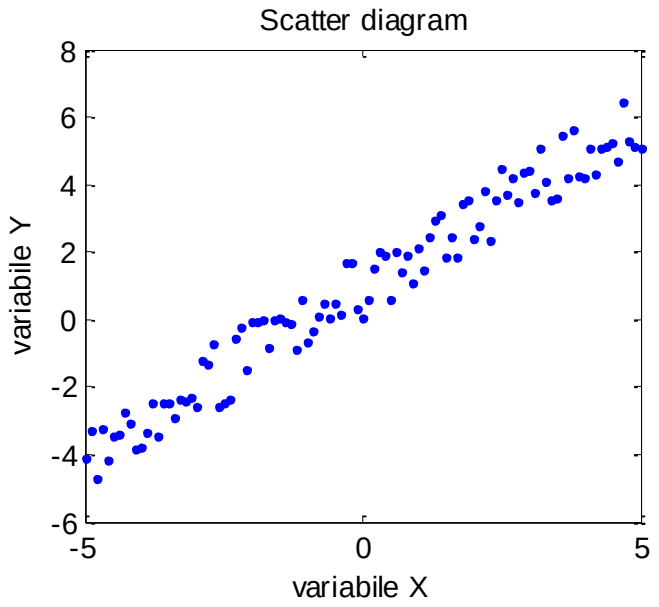
- GRAFICHE DI DATI QUANTITATIVI, FORNISCONO:

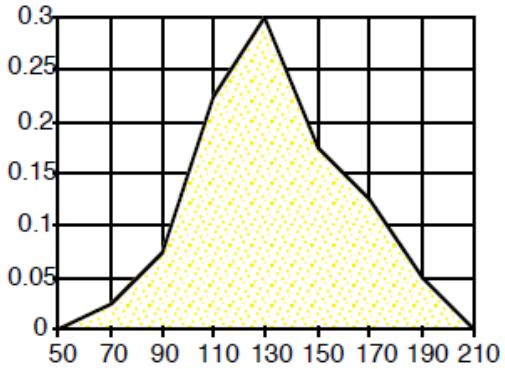
- una sintesi visiva delle caratteristiche fondamentali delle distribuzioni
- impressioni percepite con maggiore facilità
- meno particolari
- una descrizione espressa mediante una interpretazione soggettiva

ISTOGRAMMI, POLIGONI E TORTE

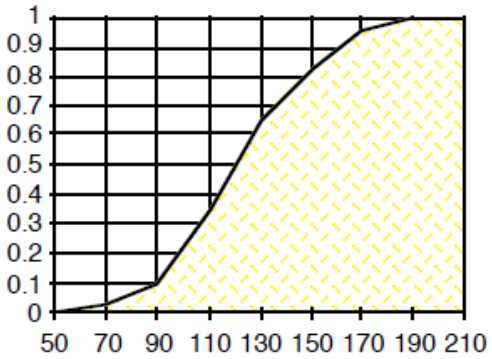


SCATTER DIAGRAM (GRAFICO A PUNTI, PER 2 VARIABILI)





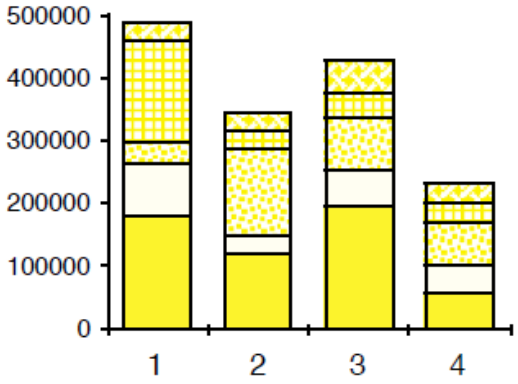
Poligono

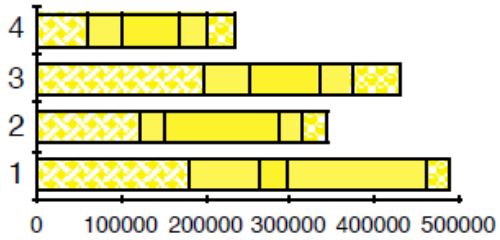


Poligono cumulato

GRAFICI A NASTRI (**ORTOGRAMMI**)

Sono simili ai rettangoli distanziati, ma con le classi di frequenza sequenziali sulla stessa barra, per una migliore lettura e comparazione





SINTESI DEI DATI

Valor Medio (Sample Mean): $$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Media Ponderata (Pesata): $$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$ dove n_i è la frequenza di ripetizione di x_i

Proprietà della Media: $\sum_{i=1}^{N}(x_i - \overline{x}) = 0$ e $\sum_{i=1}^{N}(x_i - \overline{x})^2$ è MINIMO rispetto ad ogni altro k≠$\bar{x}$

Media Geometrica: $M_g = \sqrt[N]{\prod_{i=1}^{N} x_i}$ **ponderata:** $M_{g,pond} = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}}$

OSS: $\log M_g = \frac{1}{N}\sum_{i=1}^{k} n_i \log x_i$ si applica a fenomeni che seguono leggi esponenziali.

Mediana: divide in 2 aree uguali la distribuzione:

per distribuzioni **DISCRETE**:

- Se N dispari $Me = x_{\frac{N+1}{2}}$
- Se N pari $Me = \frac{x_{\frac{N}{2}} + x_{\frac{N+1}{2}}}{2}$

Per distribuzioni **CONTINUE** (a classi): $Me = b + \frac{\frac{N}{2} - N_i}{f} \cdot c$

dove *b* è il valore della classe precedente al valore cumulativo pari a metà del totale.

ESEMPIO.

(Me-b):(N/2-Ni)=c:f

Classi	Valore Classe	N° elementi classe o frequenza	Frequenza cumulativa
1	700	12	12
2	900	21	33
3	1100	52	85
4	1300	70	155
5	1500	68	223
6	1700	36	259
7	1900	16	275
8	2100	11	286
9	2300	9	295
10	2500	5	300
Totale	-	300	-

f=155-85=70

La classe 4 contiene la MEDIANA, in quanto si supera il valore 150 (metà del totale 300). Da questo sappiamo quindi:

b= 1100

c=1300-1100=200

N=300

f=70

N_i *= 85*

In definitiva avremo:

Me=1100 + $\frac{150-85}{70}$ *200 = 1285.7*

La **mediana** serve nelle distribuzioni asimmetriche (non Normali)

<u>Proprietà della Mediana</u>:

- $\sum_{i=1}^{N} |x_i - Me|$ = min rispetto ad ogni k≠Me;
- Non è sensibile ai valori estremi;
- può essere usata anche per variabili non numeriche.

Quantili: indicano una determinata posizione nella distribuzione della variabile.
<u>Quartili</u>: divisione della distribuzione in 4 parti (aree) uguali → 2° quantile ≙ Mediana
<u>Percentili</u>: divisione in percentuale

Moda: il valore che presenta la massima frequenza (ovvero il max). Si possono avere più massimi (bimodale, trimodale, ecc. Serve per quei fenomeni che presentano tante unità con tendenza a presentarsi (≡ più massimi). Si usa per variabili qualitative.

Campo di variazione: R=$x_{max} - x_{min}$

Scarto quadratico medio: σ = deviazione standard → indice di variabilità di una popolazione, indica la dispersione dei dati intorno al valore atteso.

Varianza: $\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{N(-1)} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{N} - \bar{x}^2$

NB: La correzione (-1) serve nei casi in cui N sia molto piccolo.

Varianza in dati raggruppati in classi => correzione di Sheppard (o 'di continuità'):

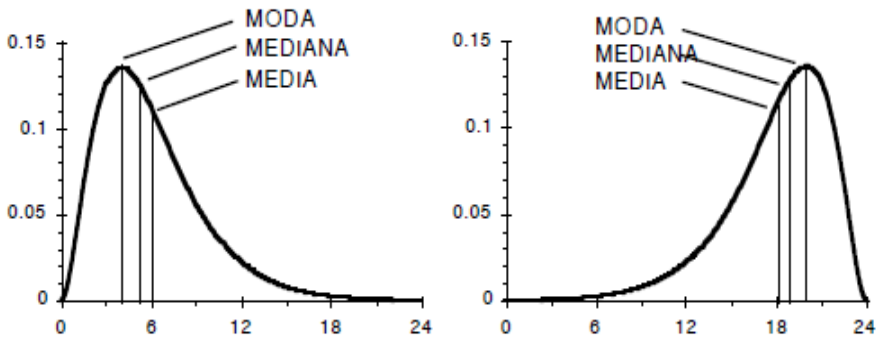
$$\sigma^2 = \sigma^2 - \frac{h^2}{12}$$ con h= ampiezza delle classi

Coefficiente di variazione: $CV = \sigma / \bar{x}$ mette in rilievo la variabilità di un fenomeno. (è un numero puro, ottimo per variabili eterogenee o con medie molto diverse)

Skewness (asimmetria): $b_1 = \dfrac{\sqrt{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}^{3/2}}$ rapporto tra momento di ordine 3 e mom. di ordine 2 alla $^3/_2$.

Kurtosis (curtosi): $b_2 = \dfrac{n \cdot \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}^{2}}$ fornisce una misura della forma, appiattita o allungata rispetto alla distribuzione Normale.

- distribuzione perfettamente normale —> 3
- dati più addensati verso il centro (lepto) —> > 3
- curva schiacciata (plati) —> < 3



Altri Indici di Asimmetria:

SKEWNESS DI PEARSON (sk)

E' la differenza (d) tra media e moda divisa per la deviazione standard (s)

sk=d/s

Proprietà :

- sk può essere nullo, positivo o negativo secondo la forma della distribuzione
- è un rapporto adimensionale: si può utilizzare per il confronto tra due o più distribuzioni

INDICE ϒ1 DI FISHER

E' il momento standardizzato di terz'ordine: $\Upsilon 1 = \frac{m^3}{\sigma^3}$

INDICE β1 DI FISHER

E' il quadrato dell'indice di Fisher: **β1 =**$\Upsilon 1^2$

MOMENTI DI ORDINE K rispetto ad un punto c

$$m_k = \frac{\sum (x_i - c)^k}{n}$$ per una serie di dati

$$m_k = \frac{\sum (x_i - c)^k \cdot f_i}{n}$$ per una distribuzione di frequenza divisa in classi

c = **origine** (c = 0) --> **momento rispetto all'origine**, oppure

c = **media** (c = media)--> **momento centrale**

CONCETTO DI PROBABILITA' MATEMATICA

A PRIORI:

Basata sul concetto che la probabilità di un evento è il rapporto tra il numero di casi favorevoli ed il numero di casi possibili, purchè tutti i casi siano egualmente probabili!

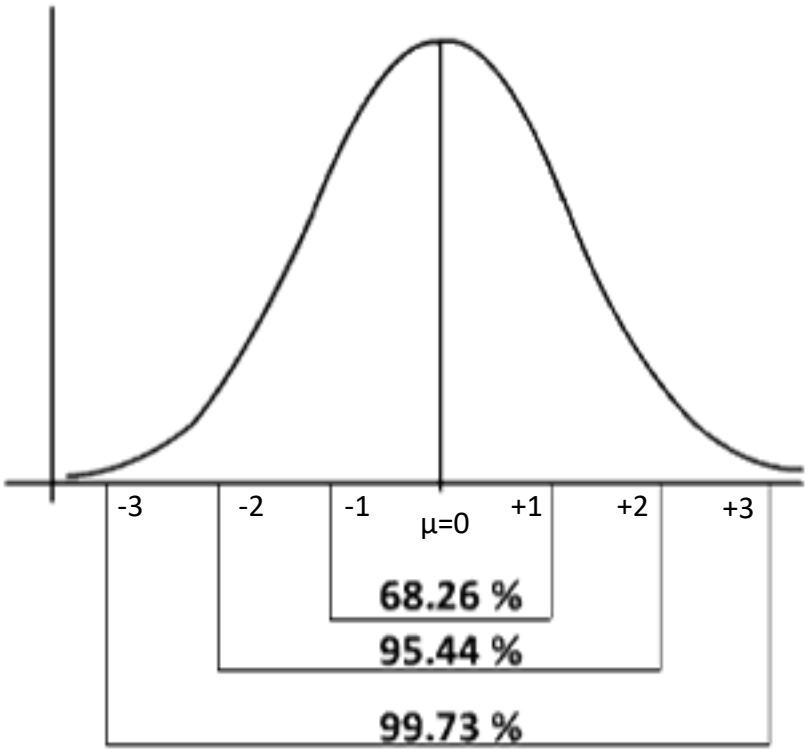
=> limitazioni per la ricerca sperimentale poichè questa è basata su un approccio non teorico ma empirico: per valutare una probabilità sarebbe necessario conoscere preventivamente le diverse probabilità dei vari eventi

A POSTERIORI (Frequentista o Statistica) :

- se in un insieme di prove la frequenza di un evento è all'incirca costante, questo valore di frequenza è assunto come probabilità
- si basa sul **principio di von Mises** (formulato nel 1920) : la probabilità di un evento, in una serie di prove condotte nelle stesse condizioni, è il limite a cui essa tende al crescere del numero delle osservazioni
- si applica in tutti quei casi in cui **non sono note a priori** le leggi dei fenomeni studiati, ma possono essere determinate a posteriori; ovvero per calcolare la probabilità attesa di trovare un numero stabilito di individui in un conteggio, deve essere nota la percentuale di presenza rilevata attraverso una precedente serie di osservazioni. Infatti, l'unico modo per rispondere ai quesiti empirici è condurre una serie di osservazioni od esperimenti, in condizioni controllate statisticamente, per rilevare la frequenza relativa del fenomeno

DISTRIBUZIONE DI PROBABILITA'

- **NORMALE**



Equazione della curva: $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{(x-\mu)^2 / 2\sigma^2}$$

Data la media $\boldsymbol{\mu}$ e la devizione standard $\boldsymbol{\sigma}$ si può conoscere la percentuale (%) dei casi compresi in un certo intervallo di valori.

Posso normalizzare questa variabile attraverso la formula:

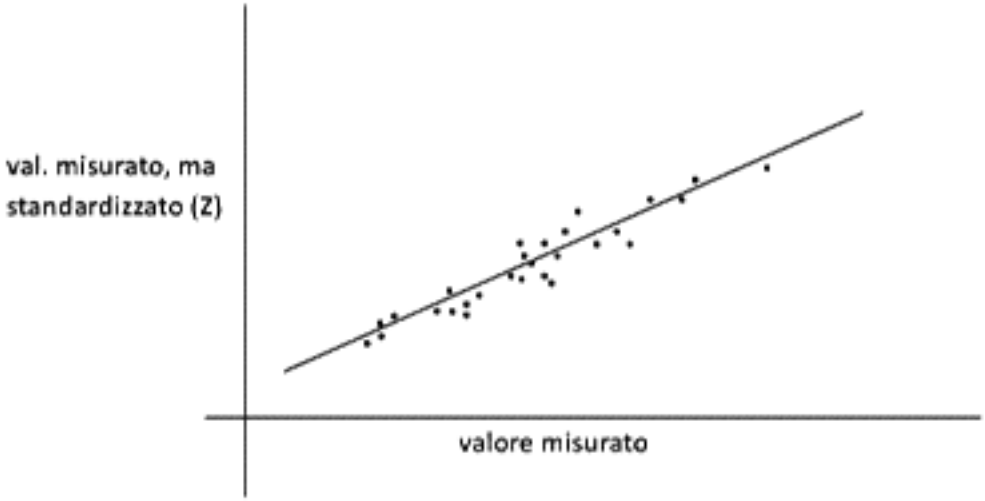
$$\boldsymbol{Z \triangleq \frac{x-\mu}{\sigma} \triangleq}$$ **variabile Normale Standardizzata**

Questa variabile ha: $\boldsymbol{\mu = 0, \sigma = 1, b_1 = 0, b_2 = 3.}$

95%=1.96σ
99%=2.58σ

TEOREMA DEL LIMITE CENTRALE

Qualunque distribuzione di variabile casuale regoli il fenomeno in esame, se il n° di osservazioni $\rightarrow +\infty$, allora essa è riconducibile ad una DISTRIBUZIONE NORMALE.



Esistono diversi modi per testare la normalità dei dati, per esempio passando attraverso il Normal Plot: si ottengono dai dati standardizzati, ordinando le osservazioni in modo crescente e "plottando" i dati contro i corrispondenti "Normal Scans" (= porzioni di σ dalla media). Se la distribuzione è Normale, dovrei ottenere una retta.

NB: Esiste il test di Shapiro-Wilk (W-test) che esegue la correlazione tra i dati e una normale, dando un valore per accettare o meno l'ipotesi di Normalità.

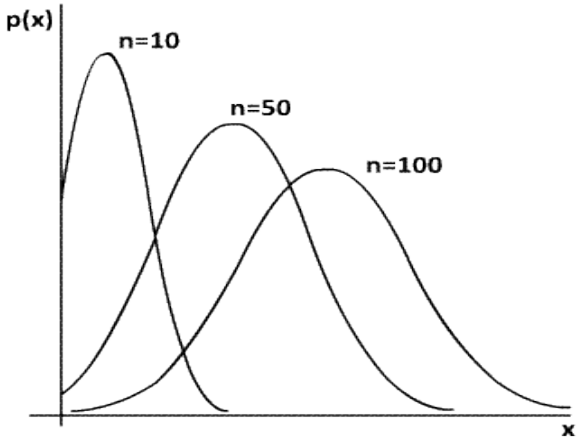
Nel caso in cui non si verifichi questa ipotesi, esistono diversi modi per trasformare la variabile in Normale ($\log x, \sqrt{x}, e^x, \dots$) sotto la condizione di MONOTONIA. Dopo la normalizzazione va sempre eseguito il W-test per verificare che l'ipotesi sia vera. Importante riuscire a Normalizzare, perché poi sarà possibile sfruttare TEST PARAMETRICI.(NB: non sempre è possibile!)

- **BINOMIALE**: solamente 2 valori possibili con probabilità p e q

p è la probabilità evento favorevole;
q=1-p è la probabilità dell'evento sfavorevole.

=> Gaussiana per n=>∞

$$p(k) = \binom{n}{k} p^k q^{n-k} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \sum_{k=0}^{n} p(k) = 1$$



k= n° successi in n prove
n-k=n° insuccessi in n prove

CONDIZIONI NECESSARIE:

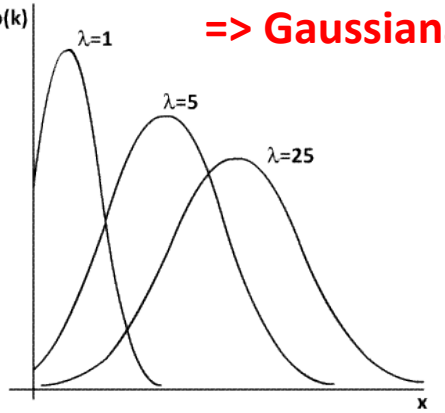
1. 2 sole risposte possibili;
2. prove indipendenti tra loro;
3. la probabilità non cambia tra le prove.

Media: $\boldsymbol{\mu = n\,p}$
Varianza: $\boldsymbol{\sigma^2 = np(1-p) = npq}$

- **POISSON**: per eventi che si verificano con bassa frequenza (rari). È il caso limite della binomiale $\begin{cases} n \to \infty \\ p \to 0 \end{cases}$

=> Gaussiana per λ>>0

$$\lim_{n\to+\infty} p(n) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \lambda = np$$



Media: $\boldsymbol{\mu = \lambda}$
Varianza: $\boldsymbol{\sigma^2 = npq \simeq np = \lambda}$

NB: Per $\lambda \to +\infty$ la distribuzione tende ad una distribuzione Gaussiana (basta $\lambda > 20$)

- **t-Student**: simile alla Gaussiana, è possibile variare il numero dei gradi di libertà. La si usa ogni qual volta si lavora su un campione piuttosto che su una popolazione, per tener conto di n e si ha solamente una stima di σ. Si definisce il parametro grado di libertà, legato alla numerosità del campione. Per gdl>100 la distribuzione è assimilabile ad una Gaussiana.

- χ^2 **(chi quadrato)**: somma di tante gaussiane al quadrato. Con gdl=1, avrò il quadrato di 1 sola Gaussiana.

  $\chi^2_k = \sum_{i=1}^{k} Z_i^2$ con $Z_i = N(0,1)$. Se $K > 25$, la distribuzione avrà forma Gaussiana.

- **F**(**di Fisher**): viene sfruttata per valutare rapporti di varianze.

STATISTICA INFERENZIALE

La conduzione dell'indagine (o ESPERIMENTO) è un percorso di ricerca scientifica articolabile in quattro fasi:

1 - **disegno sperimentale**

- osservazioni in natura e ripetizioni in laboratorio non raccolte ed attuate a caso, ma scelte e programmate in funzione della ricerca e delle ipotesi esplicative
- chiarire a priori la formulazione dell'**IPOTESI ESPLICATIVA** (alternativa all'**IPOTESI NULLA**)

Le eventuali differenze riscontrate dovranno essere imputate a

FATTORI **CAUSALI SPECIFICI** ?

oppure solamente a FATTORI **CASUALI IGNOTI** ?

attribuibili alla naturale variabilità di misure e materiale utilizzato

2 - **campionamento**

- raccogliere i dati in funzione dello scopo della ricerca
- rispettare le caratteristiche della popolazione

Numero limitato di dati –> conclusioni generali –> tutta la popolazione (UNIVERSO)

STATISTICA INFERENZIALE

3 - **descrizione dei dati raccolti** per verificare l'adeguatezza di:
- disegno sperimentale
- campionamento
- analisi condotte
- risultati conseguiti

4 - **utilizzo dei tests** (programmati nel disegno sperimentale e in funzione dei quali viene effettuato il campionamento)
processo logico-matematico che, mediante il calcolo di probabilità, porta alla conclusione di non poter respingere oppure di dover **respingere l'ipotesi nulla**

Soltanto con una corretta applicazione del campionamento e dei test di confronto statistico è possibile rispondere alla **DOMANDA INFERENZIALE** di verifica dell'**ipotesi nulla**:

LE DIFFERENZE FRA LE OSSERVAZIONI EMPIRICHE SONO DOVUTE A FATTORI PURAMENTE CASUALI ?

STATISTICA INFERENZIALE

Quale è la probabilità che, fra le alternative possibili, si presenti proprio la situazione descritta dai dati raccolti?

- probabilità alta (convenzionalmente => 5%) ——> **fattori casuali**
- probabilità bassa (convenzionalmente < 5%) ——> **fattori non casuali**
  cioé rientranti tra i criteri con cui i dati sono stati raggruppati

Analisi e conclusioni sono rese complesse fondamentalmente da tre aspetti:

- **errori nelle misurazioni** generati da strumenti e da differenti abilità degli sperimentatori
- **utilizzo di campioni:** i dati utilizzati in una ricerca non sono mai identici a quelli rilevati nelle altre
- **fattori contingenti di disturbo:** possono incidere in modo differente sul fenomeno indagato (es.: tempo, luogo, …)

INFERENZA STATISTICA

Per effettuare uno studio non utilizzo tutta la popolazione d'interesse, ma limito lo studio ad un solo sottoinsieme, **un CAMPIONE**, per poi estendere i risultati a tutta la popolazione => INFERENZA

Immaginiamo di avere a disposizione TUTTA la popolazione (di media $\boldsymbol{\mu}$ e deviazione standard $\boldsymbol{\sigma}$)

Estraendo dei campioni lo si può fare con 2 modalità: Esaustiva e Bernoulliana:

Esaustiva: estraggo un campione di n soggetti, in modo casuale, misurando media m_i e deviazione standard s_i e non considero più i soggetti scelti

Bernoulliana: estraggo ugualmente un campione di n soggetti casualmente, misuro media e dev. std, ma poi i soggetti possono essere riestratti nel successivo campione.

OSS: Nel nostro caso utilizzeremo SEMPRE l'estrazione Bernoulliana!

La distribuzione delle medie campionarie, in entrambe le estrazioni, m_i è GAUSSIANA, con

media pari a μ (stimatore **CORRETTO**)

varianza uguale a $\frac{\sigma^2}{n}$ per la Bernoulliana e $\frac{\sigma^2}{n} * \frac{N-n}{N-1}$ per l'esaustiva (stimatore **NON CORRETTO**, in quanto ha bisogno di un fattore correttivo)

con *n* = numerosità del campione. Per n << N le due varianze coincidono.

Se si considerasse la media delle mediane campionarie (ovvero dei singoli campioni), anche essa sarebbe pari a μ (anche la **mediana** **è** uno stimatore **CORRETTO**) mentre la **varianza** sarebbe maggiore per cui la **media** $\boldsymbol{m_i}$ rappresenta uno stimatore di μ più **EFFICIENTE** della **mediana**

Considereremo l'estrazione Bernoulliana

STIMA PUNTUALE DELLA MEDIA (della popolazione)

La eseguo tramite la media sul campione inserendo un errore della stima:

$\hat{\mu} = \mathbf{m_i} \pm \frac{\sigma}{\sqrt{n}}$ Stima **CONSISTENTE**, perché al crescere di *n* l'errore tende a 0

Se **σ** (dev.standard della popolazione) è ignota allora useremo la dev. st. del campione, $\mathbf{s_i}$, per stimarla:

$$\hat{\sigma} = \frac{s_i}{\sqrt{n-1}}$$

Le stime puntuali sono affette da errore per cui spesso si usa stimare un intervallo entro il quale, con un prefissato livello di probabilità, cadrà il valore del parametro => **INTERVALLO DI CONFIDENZA**

STIMA DELLA MEDIA (della popolazione) PER INTERVALLI - INTERVALLO DI CONFIDENZA

Dato α = probabilità che il valore vero cada fuori dell'intervallo individuato allora P = 1 – α = prob. che cada dentro = livello di confidenza (o **p-value**)

Poiché la variabile campionaria è distribuita come una gaussiana la si può normalizzare:

$$Z_i = \frac{m_i - \mu}{\sigma / \sqrt{n}}$$

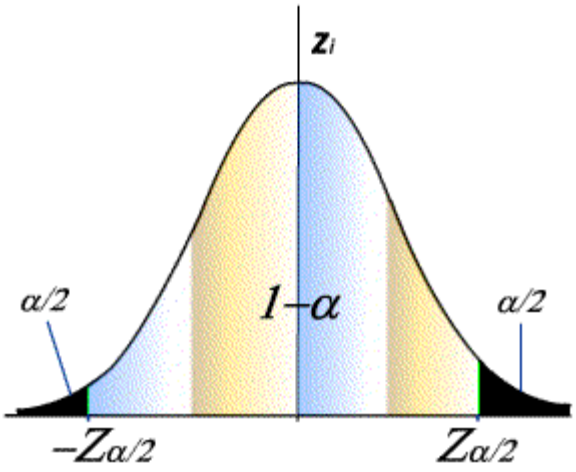
e si può scrivere:

$$\text{Prob}\left(-Z_{\alpha/2} \leq \frac{m_i - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

da cui si ricava l'**INTERVALLO DI CONFIDENZA**:

$$\hat{\mu} = m_i \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

nel quale la media μ è compresa con prob. pari a 1-α



1-α	$Z_{\alpha/2}$
0.68	1
0.8	1.28
0.9	1.64
0.95	1.96
0.9544	2
0.99	2.58
0.9973	3

$\pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ rappresenta l'incertezza della stima = l'errore! da minimizzare:

- Aumentando *n*
- Selezionando α opportuna
- Selezionando un'altra variabile ……

Se σ è ignota o *n* è 'piccolo' (<≈100) al suo posto utilizzerò S_i (dev. st. del campione) e al posto della Gaussiana utilizzerò la t-Student ($\mathbf{t_{\alpha/2,n-1}}$) per tener conto dell'ulteriore incertezza introdotta:

Un discorso analogo vale se si utilizza una coda e non due

$$\hat{\mu} = m_i \pm \mathbf{t_{\alpha/2,n-1}} \cdot \frac{S_i}{\sqrt{n}}$$

Nel caso di frequenze campionarie, l'intervallo si trasforma in:

$$\hat{p} = f_i \pm t_{\alpha/2,n-1} \cdot \frac{\sqrt{f_i \cdot (1-fi)}}{\sqrt{n}}$$

in quanto la distribuzione delle medie campionarie di una variabile binomiale ha **$\mu = p$** e **$\sigma^2 = p*q$**

Esempio: epidemia di influenza, su 100 soggetti il 70% è affetto da influenza => f=0.7.
Valutare la bontà della stima della vera % nella popolazione.
Scelgo 1-α=0.9545 => $Z_{\alpha/2}=2$
=> p=0.7±2√(0.7)(1-0.7)/100 = 0.7±0.09 p1=0.61, p2=0.79*
Con probabilità 1-α la vera p di cittadini affetti da influenza è compresa tra 61% e 79%

DIMENSIONE OTTIMA DEL CAMPIONE *n*

Esistono diversi metodi per ottenere l'***n*** che porti alla significatività dello studio statistico.
A partire dall'intervallo di confidenza si definisce l'errore come:

$$E = Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Fissato E (errore tollerato, espresso in termini percentuali, 2-3%), si avrà come sola incognita ***n***, la numerosità del campione, ovvero:

$$\boldsymbol{n = \frac{Z^2_{\alpha/2} \cdot \sigma^2}{E^2}}$$

NB: Se σ non è disponibile (come succede nella maggior parte dei casi) estraggo un piccolo campione (~ 10) e cerco una sovrastima della deviazione standard in maniera molto approssimativa mediante la formula: $\hat{\sigma} = \frac{max - min}{4}$

Se distribuzione BINOMIALE, essendo, per la distribuzione campionaria, $\sigma^2 = p \cdot q$, si avrà:

$$n = \frac{Z^2_{\alpha/2} \cdot p \cdot q}{E^2}$$

Se non noto, il valore di p si può stimarlo nel peggiore dei casi, ovvero quando $p = q = 0.5$.

ESEMPIO: Determinare la dimensione campionaria sufficiente per condurre una ricerca con questionario sulle caratteristiche dei medici di base in Italia. Fissando:

$$1 - \alpha = 95.46\% \Rightarrow Z\alpha_{/2} = 2$$

e $E = 3\%$, $p = q = 0.5$ *(non noti a priori)*

Avremo che $n = \frac{4 \cdot 0.5 \cdot 0.5}{0.03^2} = 1111$ *cioè con 1111 risposte prese a caso nella popolazione dei medici di base vi saranno 95% di prob che i risultati del campione siano validi con un margine di errore del 3%.*

E' importante notare che questo dato non è vincolato dalla popolazione totale, quindi se si fosse effettuato uno studio più limitato a livello geografico (Regione o Provincia,) il numero del campione sarebbe stato sempre di 1111 medici, alla stregua dello studio Nazionale.

VIA NOMOGRAMMA

can be, but greater power requires a larger sample, as we will see. It is common to require a power of between 80% and 90%. In effect, we try to make clinical importance and statistical significance agree, and thus reduce problems of interpretation.

The necessary sample size is usually obtained from complicated formulae or there are extensive tables available (Machin and Campbell, 1987), but it is much simpler to use a graphical method. Figure 15.2 shows a nomogram that can be used to calculate the appropriate sample size for all the situations considered in this chapter. It is simple to use and has the added

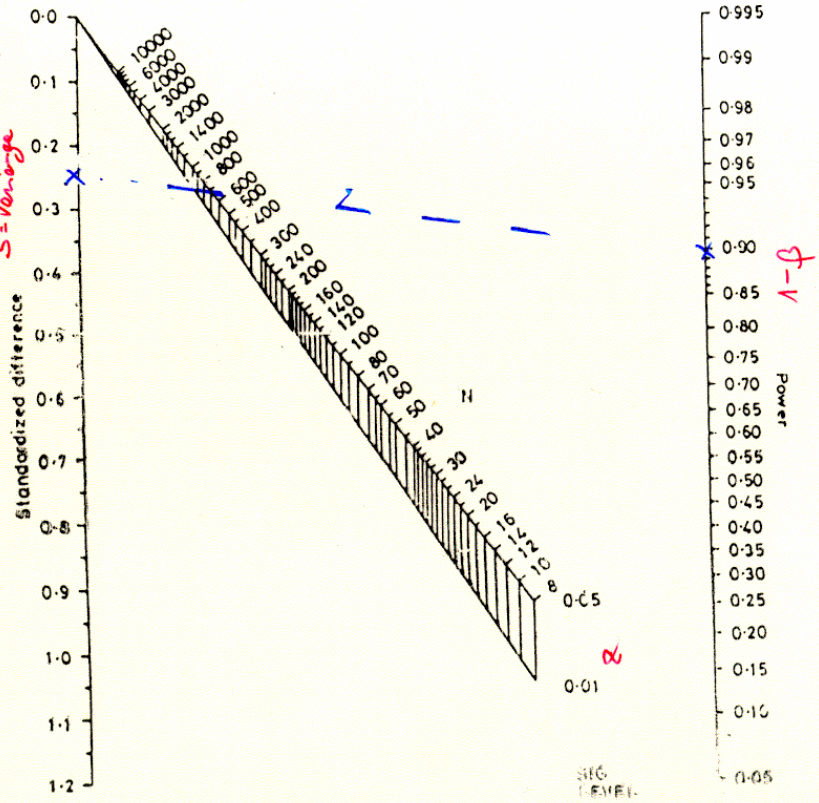


Figure 15.2 Nomogram for calculating sample size or power (reproduced from Altman, 1982b, with permission)

advantage of being equally easy to use in reverse for determining the power of a study of given sample size.

I shall first consider the case where we intend to have two groups of equal size. The nomogram can be used, however, for unequal sample sizes, as I shall show later. All of the sample size calculations are based on the quantity known as the **standardized difference**. This is calculated in a different way for continuous or categorical outcome variables, but in principle it is based in each case on the ratio of the difference of interest to the standard deviation of the observations. In other words, we express the difference of interest as a multiple of the standard deviation. As we would expect, the smaller this ratio is the larger the required size of the trial.

(a) Continuous data – two independent groups

For studies of two independent groups of patients with a continuous outcome measure we need to specify the following quantities:

1. standard deviation of the variable (in each group) (s);
2. clinically relevant difference (δ);
3. the significance level (α – two-sided);
4. the power ($1 - \beta$);

and it is assumed that the variable has a Normal distribution in the population. The total sample size is N.

The standardized difference is calculated simply as the ratio of the difference of interest to the standard deviation, that is δ/s. We can use Figure 15.2 to calculate the necessary sample size from the standardized difference for any desired power, choosing either a 5% or 1% level of significance.

For example, suppose that we are planning a milk-feeding trial in five-year-old children, to see if a daily supplement of milk for a year will lead to an increased gain in height compared with a control group. (Such a study would in fact be difficult to carry out, for practical and ethical reasons.) We know from published data that at this age children grow on average about 6 cm in a year, with a standard deviation of 2 cm. Suppose that the effect of the milk on height gain will be considered important if it is at least 0.5 cm. We want a high probability of detecting such a difference, so we set the power to be 0.9 (90%) and choose a 1% significance level. The standardized difference is $0.5/2.0 = 0.25$. We can now use Figure 15.2 to calculate the necessary sample size. We 'draw' a straight line from the value 0.25 on the scale for the standardized difference to the value 0.90 on the scale for power and read off the value for N on the line corresponding to $\alpha = 0.01$, which gives a total sample size of 900, i.e. 450 in each group.

There are several possible approaches if no estimate of the standard deviation is available. One way is to start the trial and use the data for the

Caso di due campioni (p.es. confronto casi-controllo)

- Il potere del test (1-β) che si desidera ottenere (valori di riferimento standard sono l'80% e il 90%). Probabilità di rigettare l'ipotesi nulla quando esiste una reale differenza o associazione.

- La differenza minima tra i trattamenti che sia "clinicamente rilevante". Può essere espresso in generale come il grado di beneficio che il nuovo trattamento dovrebbe fornire rispetto a quello vecchio perché valga la pena di utilizzarlo. Negli studi di *superiorità* è la minima differenza clinicamente rilevante; negli studi di *non-inferiorità* massimo svantaggio clinicamente tollerabile

- Il livello di significatività (α), cioè la probabilità di ottenere una differenza "statisticamente significativa" quando di fatto differenza non c'è.

$$n = \frac{2\sigma^2}{(\mu_1 - \mu_2)^2}(Z_\beta + Z_{\alpha/2})^2$$

Per i valori più comuni di α e 1-β

	1-β		
α	0,80	0,90	0,95
0,10	6,2	8,6	10,8
0,05	7,9	10,5	13,0
0,01	11,7	13,0	17,8

$(Z_\beta + Z_{\alpha/2})^2$

$$n = \frac{p_1(100 - p_1) + p_2(100 - p_2)}{(p_1 - p_2)^2}(Z_\beta + Z_{\alpha/2})^2$$

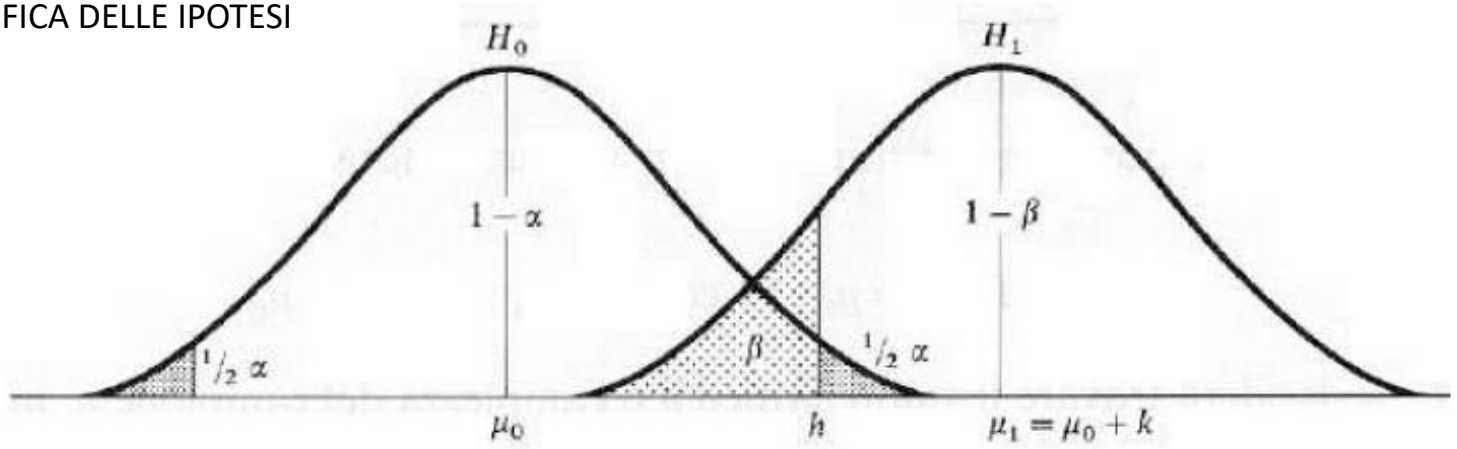
VERIFICA DELLE IPOTESI

L'ipotesi da verificare riguarda in genere la generalizzazione di un risultato ottenuto su un campione, a tutta la popolazione.
La verifica avviene tramite varie fasi che si concludono con un test.

FASI:

1. Formulazione dell'**ipotesi NULLA (H0)**, nulla perché viene formulata allo scopo di rifiutarla. H1 rappresenta l'ipotesi alternativa (generalmente contraria ad H0). H0 di solito pone l'assenza di relazioni significative tra variabili o tra campioni (p.es. delle differenze presenti tra campioni), per cui le eventuali differenze sono dovute al caso
2. Individuazione della distribuzione campionaria che, a seconda del test usato, dovrà soddisfare ad opportuni requisiti: la casualità dell'estrazione è sempre richiesta. Possono inoltre essere richieste indipendenza dei campioni, distribuzione nota, ecc. **SEMPRE RICHIESTA: CASUALITA' 'ESTRAZIONE'**
3. Scelta del livello α di significatività (in genere tra 0.01 e 0.05) che si può ricavare anche a posteriori (p-value = valore di probabilità di ottenere la differenza osservata, se H0 è vera)
4. Selezione del test e sua applicazione con relativa decisione di accettare o rifiutare l'H0 con un livello di probabilità α di sbagliare

VERIFICA DELLE IPOTESI



Distribuzioni legate alle due ipotesi H0 e H1, mutuamente esclusive

Il test statistico confronta la stima campionaria con le distribuzioni H0 e H1 e la associa ad una delle due

- A priori non sono mai sicuro che il valore misurato appartenga di diritto più all'una che all'altra distribuzione
- Esiste un'area di sovrapposizione che dipende dalla distanza dei valori medi e dalle variabilità
- Scelto α ed il numero di 'code' viene determinato h (=soglia di significatività) e quindi β

VERIFICA DELLE IPOTESI

Ogni test è associato a 4 probabilità interdipendenti che misurano il rischio che si corre, ovvero della sicurezza che si ha, nel formulare una conclusione:

- **Errore di I tipo** (rischio **α** o livello di significatività a cui corrisponde il p-value): probabilità che esprime il rischio di RIFIUTARE H0 quando è VERA (falso negativo)
- **Errore di II tipo** (rischio **β**): probabilità del rischio di ACCETTARE H0 quando è FALSA (falso positivo)
- **Protezione del test** (1- α): probabilità di ACCETTARE H0 quando è VERA
- **Potenza del test** (1- β): probabilità di RIFIUTARE H0 quando è FALSA

	REALTA'	
CONCLUSIONE DEL TEST	**H_0 vera**	**H_0 falsa**
accetto H_0 statisticamente non significativo	**Esatto** **p = 1-α** PROTEZIONE	**Errore β** di II^ tipo **p = β**
rifiuto H_0 statisticamente significativo	**Errore α** di I^ tipo **p = α**	**Esatto** **p = 1-β** POTENZA

VERIFICA DELLE IPOTESI

Poiché è impossibile diminuire un errore senza aumentare l'altro:

- scelgo tra i test a disposizione quello più 'potente' (che minimizza β)
- aumento la numerosità dei dati nel campione => cala la varianza delle distribuzioni campionarie (si 'stringono') e di conseguenza il β

Note:

- Di solito la distribuzione H1 non è nota e ci si basa solo sui dati per l'Hp. H0, di conseguenza non si valutano bene gli errori β
- Si può determinare la dimensione ottima N del campione imponendo un certo β, utilizzando il Nomogramma precedente (che parte da 1-β, da α, da 's' = dev.std nella popolazione e dalla differenza significativa accettata, δ, da cui si ricava δ/s = differenza standardizzata), ovvero, dato N, dal medesimo Nomogramma posso stimare il β

VIA NOMOGRAMMA

can be, but greater power requires a larger sample, as we will see. It is common to require a power of between 80% and 90%. In effect, we try to make clinical importance and statistical significance agree, and thus reduce problems of interpretation.

The necessary sample size is usually obtained from complicated formulae or there are extensive tables available (Machin and Campbell, 1987), but it is much simpler to use a graphical method. Figure 15.2 shows a nomogram that can be used to calculate the appropriate sample size for all the situations considered in this chapter. It is simple to use and has the added

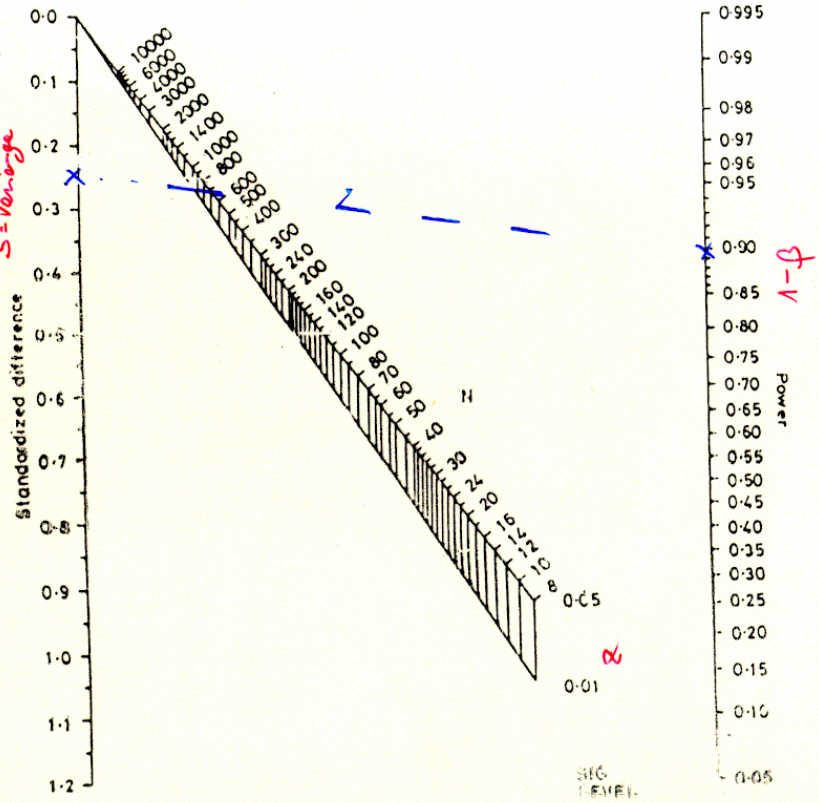


Figure 15.2 Nomogram for calculating sample size or power (reproduced from Altman, 1982b, with permission)

advantage of being equally easy to use in reverse for determining the power of a study of given sample size.

I shall first consider the case where we intend to have two groups of equal size. The nomogram can be used, however, for unequal sample sizes, as I shall show later. All of the sample size calculations are based on the quantity known as the **standardized difference**. This is calculated in a different way for continuous or categorical outcome variables, but in principle it is based in each case on the ratio of the difference of interest to the standard deviation of the observations. In other words, we express the difference of interest as a multiple of the standard deviation. As we would expect, the smaller this ratio is the larger the required size of the trial.

(a) Continuous data – two independent groups

For studies of two independent groups of patients with a continuous outcome measure we need to specify the following quantities:

1. standard deviation of the variable (in each group) (s);
2. clinically relevant difference (δ);
3. the significance level (α – two-sided);
4. the power ($1 - \beta$);

and it is assumed that the variable has a Normal distribution in the population. The total sample size is N.

The standardized difference is calculated simply as the ratio of the difference of interest to the standard deviation, that is δ/s. We can use Figure 15.2 to calculate the necessary sample size from the standardized difference for any desired power, choosing either a 5% or 1% level of significance.

For example, suppose that we are planning a milk-feeding trial in five-year-old children, to see if a daily supplement of milk for a year will lead to an increased gain in height compared with a control group. (Such a study would in fact be difficult to carry out, for practical and ethical reasons.) We know from published data that at this age children grow on average about 6 cm in a year, with a standard deviation of 2 cm. Suppose that the effect of the milk on height gain will be considered important if it is at least 0.5 cm. We want a high probability of detecting such a difference, so we set the power to be 0.9 (90%) and choose a 1% significance level. The standardized difference is $0.5/2.0 = 0.25$. We can now use Figure 15.2 to calculate the necessary sample size. We 'draw' a straight line from the value 0.25 on the scale for the standardized difference to the value 0.90 on the scale for power and read off the value for N on the line corresponding to $\alpha = 0.01$, which gives a total sample size of 900, i.e. 450 in each group.

There are several possible approaches if no estimate of the standard deviation is available. One way is to start the trial and use the data for the

TEST DI SIGNIFICATIVITA'

Parametrici: richiedono assunzioni sul tipo di distribuzione della popolazione (Normale, Binomiale, t-Student, Fisher, ….)

se: campioni poco numerosi e/o
forma della distribuzione sulla popolazione non certa

=> EVITARLI!

Sono i più POTENTI e consentono, fissato α, di minimizzare gli errori di II tipo

Non Parametrici ('distribution free'): richiedono solo l'ordinabilità della variabile (quindi vanno bene anche con variabili non numeriche purchè ordinabili).

Sono meno POTENTI/EFFICIENTI (ma non troppo); si basano sui RANGHI

Ok per basse numerosità, dati fortemente asimmetrici (sui quali si può eventualmente agire mediante trasformazione con funzioni –p.es.log), ….

- I TEST SI APPLICANO PER INFERIRE SULLA MEDIA, SU DIFFERENZE DI MEDIE, SU PROPORZIONI, SU UN CAMPIONE, SU PIU' CAMPIONI.
- DIETRO TUTTI I TEST C'E' L'IDEA DI UN MODELLO STATISTICO, OVVERO DI UN LEGAME/RELAZIONE ESISTENTE TRA 2 O PIU' VARIABILI, CHE GIUSTIFICHI I RISULTATI OSSERVATI SUL CAMPIONE

VERIFICA DELLE IPOTESI

1 CAMPIONE vs 1 POPOLAZIONE DI RIFERIMENTO => H0: il campione appartiene alla popolazione

2 CAMPIONI (non serve conoscere la popolazione) => H0: i due campioni appartengono alla stessa popolazione

PIU' DI 2 CAMPIONI: considero coppie di campioni oppure => H0: tutti i campioni appartengono alla medesima popolazione; se H0 è falsa allora almeno 1 campione è 'estraneo' e posso individuarlo/i ripetendo il test su sottoinsiemi di campioni

VERIFICA - TEST SU 1 CAMPIONE

Test sulla media: confronto il valore medio di un campione con quello **NOTO** della popolazione.

H_0: il campione appartiene alla popolazione, ovvero l'eventuale differenza tra i valori medi $\bar{x}$ e μ è dovuta al caso e non è *statisticamente significativa*

H_1: $\mu_0 > \mu$ (e/o $\mu_0 < \mu$, dipende dal caso in questione, 1 o 2 code)

Z - Test (parametrico):

- **Popolazione con distribuzione gaussiana di note media μ e σ**
- **Campione di numerosità *n* di media $\bar{x}$**

Si calcola $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ e si confronta con i valori di Z_α di una Normale

Se $\boldsymbol{Z > Z_\alpha}$ rifiuterò l'ipotesi nulla (test ad 1 coda) con un livello di significatività pari ad α

ovvero (test a 2 code)

se $\boldsymbol{Z > Z_{\alpha/2}}$ oppure se $\boldsymbol{Z < -Z_{\alpha/2}}$, allora rifiuterò l'ipotesi nulla con un livello di significatività pari ad α

ovvero concluderò con una probabilità pari a $1-\alpha$ che il campione non proviene dalla stessa popolazione che aveva media μ e deviazione standard σ

Se il valore di σ della popolazione **non è noto o la numerosità del campione è bassa (<100)** allora posso stimare σ attraverso la deviazione standard del campione *s* e, per tener conto dell'incertezza sulla stima di σ, si utilizza il

T-Test o t-Student (parametrico):

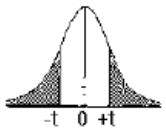
- **Popolazione con distribuzione normale di nota media μ**
- **Campione di numerosità *n* di media $\bar{x}$ e deviazione standard *s***

Si calcola $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$ con $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ (stima della varianza del campione)

Si confronterà il valore di t con le tabelle della t-Student (a 1 o 2 code) con livello di significatività α e $n-1$ gradi di libertà => $t_{\alpha,n-1}$ ovvero $t_{\alpha/2,n-1}$

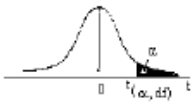
TABELLE t-Student

Valori critici della distribuzione t di Student per un test bilaterale



Gradi di libertà	α								
	0,500	0,400	0,200	0,100	0,050	0,025	0,010	0,005	0,001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	.741	.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	.727	.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	718	.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	.711	.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	.706	.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	.703	.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	.700	.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	.697	.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	.695	.873	1.356	1.782	2.179	2.560	3.055	3.428	4.318
13	.694	.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	.692	.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	.691	.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	.690	.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	.689	.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	.688	.862	.330	1.734	2.101	2.445	2.878	3.197	3.922
19	.688	.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	.687	.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	.686	.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	.686	.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	.685	.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	.685	.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	.684	.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	.684	.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	.684	.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	.683	.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	.683	.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	.683	.854	1.310	1.697	2.042	2.360	2.750.	5.030	3.646
35	.682	.852	1.306	1.690	2.030	2.342	2.724	2.996	3.591
40	.681	.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
45	.680	.850	1.301	1.680	2.014	2.319	2.690	2.952	3.520
50	.680	.849	1.299	1.676	2.008	2.310	2.678	2.937	3.496
55	.679	.849	1.297	1.673	2.004	2.304	2.669	2.925	3.476
60	.679	.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
70	.678	.847	1.294	1.667	1.994	2.290	2.648	2.899	3.435
80	.678	.847	1.293	1.665	1.989	2.284	2.638	2.887	3.416
90	.678	.846	1.291	1.662	1.986	2.279	2.631	2.878	3.402
100	.677	.846	1.290	1.661	1.982	2.276	2.625	2.871	3.390
120	.677	.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
∞	.6745	.841	1.28161	1.6448	1.9600	2.2414	2.5758	2.8070	3.2905

Valori critici della distribuzione t di Student per un test unilaterale (prima parte)



Gradi Di Libertà	Aree della coda superiore					
	0.25	0.10	0.05	0.25	0.01	0.005
1	1.0000	3.07	6.313	12.706	31.8207	63.6574
2	0.8165	1.88	2.920	4.302	6.9646	9.9248
3	0.7649	1.63	2.353	3.182	4.5407	5.8409
4	0.7407	1.53	2.131	2.776	3.7469	4.6041
5	0.7267	1.47	2.015	2.570	3.3649	4.0322
6	0.7176	1.43	1.943	2.446	3.1427	3.7074
7	0.7111	1.41	1.894	2.364	2.9980	3.4995
8	0.7064	1.39	1.859	2.306	2.8965	3.3554
9	0.7027	1.38	1.833	2.262	2.8214	3.2498
10	0.6998	1.37	1.812	2.228	2.7638	3.1693
11	0.6974	1.36	1.795	2.201	2.7181	3.1058
12	0.6955	1.35	1.782	2.178	2.6810	3.0545
13	0.6938	1.35	1.770	2.160	2.6503	3.0123
14	0.6924	1.34	1.761	2.144	2.6245	2.9768
15	0.6912	1.34	1.753	2.131	2.6025	2.9467
16	0.6901	1.33	1.745	2.119	2.5835	2.9208
17	0.6892	1.33	1.739	2.109	2.5669	2.8982
18	0.6884	1.33	1.734	2.100	2.5524	2.8784
19	0.6876	1.32	1.729	2.093	2.5395	2.8609
20	0.6870	1.32	1.724	2.086	2.5280	2.8453
21	0.6864	1.32	1.720	2.079	2.5177	2.8314
22	0.6858	1.32	1.717	2.073	2.5083	2.8188
23	0.6853	1.31	1.713	2.068	2.4999	2.8073
24	0.6848	1.31	1.710	2.063	2.4922	2.7969
25	0.6844	1.31	1.708	2.059	2.4851	2.7874
26	0.6840	1.31	1.705	2.055	2.4786	2.7787
27	0.6837	1.31	1.703	2.051	2.4727	2.7707
28	0.6834	1.31	1.701	2.048	2.4671	2.7633
29	0.6830	1.31	1.699	2.045	2.4620	2.7564
30	0.6828	1.31	1.697	2.042	2.4573	2.7500
31	0.6825	1.30	1.695	2.039	2.4528	2.7440
32	0.6822	1.30	1.693	2.036	2.4487	2.7385
33	0.6820	1.30	1.692	2.034	2.4448	2.7333
34	0.6818	1.30	1.690	2.032	2.4411	2.7284
35	0.6816	1.30	1.689	2.030	2.4377	2.7238
36	0.6814	1.30	1.688	2.028	2.4345	2.7195
37	0.6812	1.30	1.687	2.026	2.4314	2.7154
38	0.6810	1.30	1.686	2.024	2.4286	2.7116
39	0.6808	1.30	1.684	2.022	2.4258	2.7079
40	0.6807	1.30	1.683	2.021	2.4233	2.7045
41	0.6805	1.30	1.682	2.019	2.4208	2.7012
42	0.6804	1.30	1.682	2.018	2.4185	2.6981
43	0.6802	1.30	1.681	2.016	2.4163	2.6951
44	0.6801	1.30	1.680	2.015	2.4141	2.6923
45	0.6800	1.30	1.679	2.014	2.4121	2.6896
46	0.6799	1.30	1.678	2.012	2.4102	2.6870
47	0.6797	1.29	1.677	2.011	2.4083	2.6846
48	0.6796	1.29	1.677	2.010	2.4066	2.6822
49	0.6795	1.29	1.676	2.009	2.4049	2.6800
50	0.6794	1.29	1.675	2.008	2.4033	2.6778
51	0.6793	1.29	1.675	2.007	2.4017	2.6757
52	0.6792	1.29	1.674	2.006	2.4002	2.6737
53	0.6791	1.29	1.674	2.005	2.3988	2.6718
54	0.6791	1.29	1.673	2.004	2.3974	2.6700
55	0.6790	1.29	1.673	2.004	2.3961	2.6682
56	0.6789	1.29	1.672	2.003	2.3948	2.6665
57	0.6788	1.29	1.672	2.002	2.3936	2.6649
58	0.6787	1.29	1.671	2.002	2.3924	2.6633
59	0.6787	1.29	1.671	2.001	2.3912	2.6618
60	0.6786	1.29	1.670	2.000	2.3901	2.6603

***Esempio**: verificare se il tasso di glicemia, riscontrato nel sangue su un campione casuale di 26 individui, sia significativamente diverso da quello medio in soggetti normali (μ=90mg/100ml) con distribuzione normale.*
Nel campione $\bar{x}$ = 140mg/100ml e s = 52.5
H0: μ_0 (della popolazione da cui proviene il campione) = μ
H1: $\mu_0 > \mu$ (basta 1 coda in quanto $\bar{x} > \mu$)

Siccome σ è ignota si applicherà la t-Student. Si sceglie α (p.es.=0.01), si calcola $t_{\alpha,n-1} = t_{0.01,25}$= 2.48 e poi
$t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$ *= 50/(52.5/5) = 4.76 > valore tabulato*

=> con probabilità pari all'1% di commettere errore, possiamo respingere H0
Ovvero il campione proviene da una popolazione con media 90mg/100ml e concludere che PRESUMIBILMENTE i soggetti del campione appartengono ad una popolazione con ridotta tolleranza glucidica.

NOTE:
- Se la distribuz. delle misure presenta qualche outlier, si può usare la mediana al posto della media
- Per n => ∞ (ma basta >100) la t-Student tende alla Z, altrimenti ci vuole la correzione che aumenta il coefficiente tenendo conto dell'ulteriore incertezza

Se i dati sono qualitativi/numerabili classificabili in 2 gruppi e la numerosità *n* è bassa, si utilizza un test binomiale

Test Binomiale (parametrico):

Si confronta $p_{n,k}$ (probabilità dell'evento) con α, dove:

$$p_{n,k} = \binom{n}{k} p^k \cdot q^{n-k}$$

k = nr elementi con un certo carattere

$\binom{n}{k}$= n!/(k! (n-k)!)

n = numerosità del campione

oppure si utilizza l'approssimazione Normale:

$$Z = \frac{(|k - np| - {}^1/_2)}{\sqrt{np(1-p)}}$$

(k=valore osservato - np=valore ipotizzato – 'continuity correction')/errore standard

e la si confronta con una Normale ($Z_{\alpha/2}$)

***Esempio**: dalla letteratura è noto che un certo trattamento di una data malattia ottiene risultati positivi nel 50% dei casi. Si vuole valutare se l'introduzione di un farmaco in quel trattamento comporta aumento dei casi positivi. Supponiamo di avere n=10 e che su 8 si è avuto miglioramento. Vogliamo valutare se il risultato è significativamente diverso da quello noto.*

H0: terapia nuova non ha effetto sulla malattia e quindi la differenza riscontrata è dovuta al caso.
Poiché la variabile è dicotomica (risponde/non risponde alla terapia) e il campione è piccolo (n=10) si può usare la binomiale.
p=q=50%=0.5 (i soggetti erano già in una terapia che dava risultati nel 50% dei casi)
Per verificare l'hp calcoliamo la probabilità $p_{n,k} = \binom{n}{k} p^k \cdot q^{n-k}$ *dove k può valere 8, 9 o 10 perché la prob di ottenere in un gruppo di 10 soggetti 8 risposte positive alla terapia solo per caso equivale a dire 'almeno' 8. La prob che il campione provenga da una popolazione in cui il 50% risponde positivamente vale la somma delle* $p_{n,k}$ *per k=8,9,10 che, con n=10, vale:* 0.055 *=prob di ottenere 8 risposte alla terapia soltanto per caso. Scegliendo un livello* α *pari al 5% allora bisognerà accettare l'H0! Sostenendo che la terapia supplementare non migliora significativamente il quadro clinico (nella popolazione da cui proviene il campione)*

SIGN Test (non parametrico): si ordinano le differenze tra le osservazioni ed il riferimento; N1 conta le differenze negative e N2 quelle positive. H0= Differenze non significative.
Si calcola $\chi^2 = (N1-N2)^2 / (N1+N2)$ e si confronta con $\chi^2_{N1+N2-1,\ \alpha}$

meglio:

WILCOXON-Signed Rank Sum Test (non parametrico): si usa quando oltre a considerare se ogni osservazione è sotto o sopra un certo valore di interesse, si vuole tener conto anche dell'ampiezza delle osservazioni.
1. Calcolo la differenza tra ciascuna osservazione e il valore d'interesse
2. Si ordinano ignorando il segno e si assegna il Rango a partire dal più piccolo. NOTA: se 2 o più Ranghi sono uguali, si prende quello intermedio
3. Si calcola la somma dei ranghi delle osservazioni che sono negative rispetto al valore ipotetico quella delle osservazioni positive;
4. Si confronta l'una o l'altra somma con l'opportuna tabella di WILCOXON (per $n \leq 25$),altrimenti si procede con la normale, assumendo $\mu = \frac{n\cdot(n+1)}{4}$ e $\sigma^2 = \frac{n\cdot(n+1)\cdot(2n+1)}{24}$ e confrontando con la $Z_{\alpha/2}$.

TEST SULLA FREQUENZA

Si utilizzano quando le frequenze osservate di una modalità (valore) di un carattere (variabile), su un campione, differiscono (in modo signif. statistico) da quelle teoriche (attese o note a priori). Si parla anche di conteggi e tavole di contingenza.

Test χ^2 (chi quadrato) (test non parametrico, in genere la numerosità è elevata)

ESEMPIO: Si ha un campione di 800 famiglie con 2 figli, aventi distribuzione rispetto al sesso:

MM	*MF (o FM)*	*FF*
160	*440(n_i^2)*	*200*

Poiché la probabilità che nasca un Maschio o una Femmina è 1/2 anche dopo il primo figlio (eventi indipendenti), la distribuzione teorica dei sessi dovrebbe essere:

MM	*MF (o FM)*	*FF*
200	*400 ($n_i^{2^*}$)*	*200*

H_0: i dati del campione seguono la distribuzione teorica e le diversità sono imputabili solo al caso.

Il test χ^2 è adatto per confrontare frequenze empiriche (n_i^2) con frequenze teoriche $n_i^{2^*}$

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i^2 - n_i^{2^*})^2}{n_i^{2^*}}$$

Il risultato si confronterà con il $\chi^2_{n-1,\alpha}$ essendo α il livello di significatività.

TEST SULLA FREQUENZA

Nell'esempio il test sarà bilaterale perché non c'è motivo di supporre che le diversità rispetto ai valori teorici abbiano una propensione verso un segno piuttosto che l'altro.

$\chi^2 = \frac{(160-200)^2}{200} + \frac{(440-400)^2}{400} + \frac{(200-200)^2}{200} = 12$, *mentre* $\chi^2_{2,0.01} = 7.38$

quindi l'ipotesi nulla va rifiutata.

Test di Kolmogorov (non parametrico):

indicato per *n* piccoli e se i caratteri sono continui.
Si propone di verificare l'ipotesi che un campione provenga da un dato universo continuo con distribuzione nota (verifica la forma della distribuzione).
Si calcola: $D_{calc} = max_i|F_i^* - F_i|$ dove F_i^* è la distribuzione teorica delle frequenze cumulate e F_i quella empirica delle freq cumulate => si valuta quanto le due distribuzioni siano 'simili' mediante una misura di distanza (il max su tutti i valori considerati)
Dato α ed *n*, dalle apposite tabelle si ricava il $D_{teorico}$ e lo si confronta con il precedente.
Se $D_{calc} > D_{teorico}$ si rifiuta l'ipotesi nulla.

TEST SU DUE CAMPIONI

Test sulla differenza tra le medie

Servono per stabilire con quale probabilità i 2 campioni (con medie diverse) provengono da popolazioni con la stessa media.
Si estendono tutti i test precedenti e anziché parlare di una sola media, si tratteranno DIFFERENZE DI MEDIE.

Z-test a 2 campioni (parametrico)**:**

H_0= i 2 campioni appartengono alla stessa popolazione.

La variabile differenza di due distribuzioni campionarie ha:

$$\textbf{media= } \boldsymbol{\mu_1 - \mu_2} \textbf{ e varianza= } \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Si calcola $$Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Se le due popolazioni sono le stesse, si avrà $\mu_1 = \mu_2$ e $\sigma_1 = \sigma_2$ per cui la formula applicata sarà:

$$Z = \frac{(\overline{x_1} - \overline{x_2})}{\sigma \cdot \sqrt{1/n_1 + 1/n_2}}$$

La Z si confronterà con la $Z_{\alpha/2}$

t-Test (t-Student) a 2 campioni (test parametrico):

Se i valori di σ non sono noti allora: se si ipotizza che tra i 2 campioni le varianze siano diverse, si dovranno utilizzare test complessi (test di Behrens-Fisher); altrimenti se si ipotizzano uguali, (attraverso il test di Fisher si verificherà quanto questa supposizione sia valida) si valuta:

$$t = \frac{(\overline{x_1} - \overline{x_2})}{s \cdot \sqrt{1/n_1 + 1/n_2}}$$

con

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\sum(x_{i_1} - \overline{x_1})^2 + \sum(x_{i_2} - \overline{x_2})^2}{n_1 + n_2 - 2}$$

'pooled variance', varianza ponderata

Con s_1 e s_2 calcolate sui due campioni. Si confronta con la t_{α,n_1+n_2-2}.
Se t è < t_{α,n_1+n_2-2}, prima di accettare H0 devo verificare l'omogeneità delle varianze mediante il test di Fisher.

Esempio:

Si confrontano le medie di 2 campioni casuali, indipendenti estratti da due popolazioni di forma Normale, relativamente ai giorni di degenza degli operati per un motivo xx in due diverse strutture ospedaliere (non posseggo σ della popolazione)

Ospedale A: n_1=16, $\overline{x_1}$=13, s_1=4
Ospedale B: n_2=16, $\overline{x_2}$=9, s_2=6
Prima di concludere che $\overline{x_2}$ è significativamente < $\overline{x_1}$ (e la diff non sia dovuta la caso) eseguo il test
H0: le due medie provengono dalla medesima popolazione e quindi anche le due dev standard devono essere comparabili
Stimo σ a partire dalla s ponderata e ricavo t => 2.22 che confronto con $t_{0.01,\,30}$= 2.46 e ……. concludo che H0 è vera!
Con prob. 1% di commettere un errore di tipo I potremo accettare l'hp che non esiste differenza significativa nella degenza media relativa all'intervento xx nei 2 ospedali.

A questo punto eseguo il test di Fisher per verificare l'omogeneità delle varianze

Test di Fisher

Utile per verificare se (H0) due varianze possano considerarsi provenire da una medesima popolazione, con un certo grado α di significatività

Dati n_1, n_2 (dei 2 campioni) e s_1^2, s_2^2 (con $s_1^2 > s_2^2$) si confronta se:

$$F = \frac{s_1^2}{s_2^2} \quad < \quad F_{\alpha, n_1-1, n_2-1}$$

Nell'eventualità che questo si verifichi, si accetterà l'ipotesi H0 e si confermerà il risultato della t-Student che i due campioni provengono dalla medesima popolazione

In caso contrario **NULLA** si può concludere e bisogna utilizzare altri test, p.es. quello di Welch (che stima diversamente la varianza complessiva ed il nr di gradi di libertà)

TABELLA di Fisher per α=0.01

Valori critici della distribuzione F di Fisher-Snedecor

I gradi di libertà del numeratore (o varianza maggiore) sono riportati in orizzontale (prima riga)

I gradi di libertà del denominatore (o varianza minore) sono riportati in verticale (prima colonna)

$\alpha = 0.01$

NUMERATORE

DEN.	1	2	3	4	5	6	7	8	12	24	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6106	6235	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,41	99,46	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,05	26,60	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,37	13,93	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	9,89	9,47	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,72	7,31	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,47	6,07	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,67	5,28	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,11	4,73	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,71	4,33	3,91
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,16	3,78	3,36
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	3,80	3,43	3,00
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,55	3,18	2,75
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,37	3,00	2,57
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,23	2,86	2,42
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	2,84	2,47	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,66	2,29	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,50	2,12	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,34	1,95	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,18	1,79	1,00

Se i campioni **NON sono indipendenti**, p.es. misure ripetute sugli stessi soggetti in due tempi diversi, si può utilizzare la

t-Student per campioni 'accoppiati' (paired)

che parte dalle differenze tra le misure nei medesimi soggetti nei due momenti = D_i

Si calcolano la media $\bar{D}$ e la **σ** delle differenze D_i e quindi si stima la

$$t = \frac{\bar{D}}{\sqrt{\sigma^2/n}}$$

che confronterò con $t_{\alpha,n-1}$

In pratica valuto quanto $\bar{D}$ si discosta dallo zero….

Test di Wilcoxon-Mann-Whitney o U-test **(non parametrico)**

(Wilcoxon Rank Sum test)

H_0= i 2 campioni (indipendenti) appartengono alla stessa popolazione.

Si ordinano le osservazioni dei 2 gruppi come se fossero di uno unico e si associa il rango *(se ci sono troppe osservazioni uguali ci sono correzioni da apportare)*
Si calcola il rango complessivo per ciascun gruppo (R_1 e R_2) e si valutano:

$$U = n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1) - R_1$$

$$U' = n_1 n_2 + \frac{1}{2} n_2 (n_2 + 1) - R_2$$

Mediante apposita tabella si entra con *min(U,U')* e si confronta il valore relativo con un α prefissato: se il valore è maggiore di α, si accetterà H0.
Per $n_1, n_2 > 10$, il valore di R_1 (o R_2, il minimo dei due) =**T** tende ad una Normale con media e SD:

$$\mu_T = \frac{n_s(n_s+n_l+1)}{2} \quad \text{e} \quad \sigma_T = \sqrt{\frac{n_l * \mu_T}{6}}$$

dove n_s è il campione meno numeroso e n_l il più numeroso tra i 2.

A questo punto si utilizza la statistica Z = $\frac{T-\mu_T}{\sigma_T}$ e si confronta con la $Z_{\alpha/2}$

Se i due campioni sono appaiati si utilizza il (equivalente al t-Student appaiato)

Test di Wilcoxon per dati appaiati (non parametrico)

(Wilcoxon Signed Rank test)

H_0= i 2 campioni (accoppiati) appartengono alla stessa popolazione.

Si ordinano le **differenze** (in valore assoluto) dei 2 gruppi e si associa il rango
Si calcola il rango (p.es.) delle differenze positive (**T**) e si valutano (per *n>20*):

$$\mu_T = \frac{n_1(n_1+1)}{4} \quad \text{e} \quad \sigma_T = \sqrt{\frac{(2n_1+1) * \mu_T}{6}}$$

dove n_1 è la numerosità del singolo campione.

A questo punto si utilizza la statistica Z = $\frac{T-\mu_T}{\sigma_T}$ e si confronta con la $Z_{\alpha/2}$.

Se *n<20* si utilizza apposita tabella entrando col parametro W che tiene conto della somma dei quadrati dei ranghi:

$$W=\frac{T}{\sqrt{\sum R^2}}$$

Test sulla differenza tra proporzioni o frequenze

Differenze tra proporzioni (caso binomiale)

IPOTESI: 2 gruppi di osservazioni INDIPENDENTI
Un carattere con 2 sole possibilità (valori), p.es. 'miglioramento'/'non miglioramento', con probabilità ***p*** e ***q=1-p***.
(avremo $p_1, q_1 \; e \; n_1$ e $p_2, q_2 \; e \; n_2$ nei due gruppi)

Per valori di n_1, n_2 grandi (>80-100), una buona stima di p_1 e p_2 è data dalle frequenze relative $f_1 \; e \; f_2$, inoltre la distribuzione **binomiale** si può approssimare con una Normale.
Poiché i gruppi sono indipendenti, allora la differenza tra i gruppi avrà media pari alla differenza delle medie e Varianza pari alla somma delle varianze, quindi

H_0: i 2 campioni provengono dalla stessa popolazione
devono quindi avere lo stesso valore di ***p***, stimabile con

$$\hat{\boldsymbol{p}} = \frac{\boldsymbol{n_1 f_1 + n_2 f_2}}{\boldsymbol{n_1 + n_2}}$$

e $\hat{\boldsymbol{q}} = 1 - \hat{\boldsymbol{p}}$

Utilizzeremo:

$$\boldsymbol{Z = \frac{f_1 - f_2}{\sqrt{p \cdot q \cdot (1/n_1 + 1/n_2)}}}$$

Che si potrà confrontare con $Z_\alpha \; o \; Z_{\alpha/2}$.

Per *n* piccoli al posto di $(\boldsymbol{f_1 - f_2})$ si userà la correzione (continuity correction): $|\boldsymbol{f_1 - f_2}| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

altrimenti il test sarebbe troppo ottimista.

Esempio:

25 pazienti:

12=n1, ricevettero un trattamento e 9 dissero di aver ricevuto benefici

13=n2, ricevettero un placebo e 4 dissero di aver ricevuto benefici

Frequenze osservate:

f1= 9/12=0.75, f2= 4/13=0.3077 => f1 - f2=0.4423

Stimiamo $\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ = (9+4)/(12+13) = 0.52 e lo standard error

$\sqrt{p \cdot q \cdot (1/n_1 + 1/n_2)}$ = 0.20

Z= 0.4423/0.20 = 2.21 corrispondente ad un p-value di 0.027 che è < 5% ($Z_{0.05}$ = 1.64)

QUINDI esiste evidenza di una differenza significativa tra i 2 trattamenti

Differenze tra proporzioni (caso analogo al binomiale ma su osservazioni NON indipendenti - paired)

p.es. due osservazioni sullo **stesso** individuo (p.es effetto di due farmaci)

In questo caso l'errore standard della differenza non è più basato sulla sola varianza di ciascuna proporzione, ma deve tener conto dei risultati 'correlati'. Si dividono le osservazioni (Ti) in 4 gruppi a seconda che la caratteristica sia presente o meno in ciascun membro della coppia (es*: presenza di un sintomo prima di un trattamento (T_1) e dopo (T_2)):*

T_1	T_2	n° coppie
Si	Si	a
Si	No	b
No	Si	c
No	No	d
		n

Nota: Considero '**b**' e '**c**' perché sono le sole situazioni che cambiano!

In questo caso si valuta:

$$Z = \frac{b - c}{\sqrt{b + c}}$$

e si confronta con Z_α.

Per *n* piccoli (correzione):

$$Z = \frac{|b - c| - 1}{\sqrt{b + c}}$$

Differenze tra frequenze (Tabelle di frequenza o di contingenza)

CASO GENERALE: TABELLE r x c

	CONSUMO CAFFEINA (mg/day)				
STATO CIVILE	0	1-150	151-300	>300	TOTALE
Coniugato	652	1537	598	242	3029
Separato/Divorziato/Vedovo	36	46	38	21	141
Single	218	327	106	67	718
TOTALE	906	1910	742	330	3888

H_0: le 2 variabili (stato civile/ consumo caffeina) sono scorrelate nella popolazione da cui è stato estratto il campione.

Dalla tabella delle frequenze osservate, si ricava la tabelle di quelle attese (teoriche), basandosi sul mantenimento delle distribuzioni marginali, le quali sono esenti da interdipendenza tra le variabili.
Le frequenze attese sono così ricavabili:

									Totale
	$T_1 \cdot \frac{\sum A}{T}$	$T_2 \cdot \frac{\sum A}{T}$	$T_3 \cdot \frac{\sum A}{T}$	$T_4 \cdot \frac{\sum A}{T}$	…	…	…	…	$\sum A$
	$T_1 \cdot \frac{\sum B}{T}$	…	…	…	…	…	…	…	$\sum B$
	…	$T_2 \cdot \frac{\sum C}{T}$	…	…	…	…	…	…	$\sum C$
	…	…	…	$T_4 \cdot \frac{\sum D}{T}$	…	…	…	…	$\sum D$
Totale	T_1	T_2	T_3	T_4	…	…	…	…	T

Infine si calcola:

$$\boldsymbol{\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$$

dove le $\boldsymbol{O_{ij}}$ sono le frequenze osservate e le $\boldsymbol{E_{ij}}$ sono le frequenze attese.
Tanto maggiore sarà questo valore, tanto più i valori osservati sono diversi da quelli attesi.
Si confronta quindi la $\boldsymbol{\chi^2}$ con $\chi^2_{\alpha,\nu}$ con $\nu = (c-1)(r-1)$ (gradi di libertà)

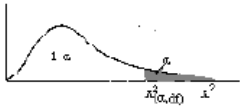
Nell'esempio $\boldsymbol{\chi^2 = 51.61}$ *e* $\boldsymbol{\chi^2_{0.01,6} = 22.46}$ $\Rightarrow$ *si rifiuta* H_0*, portando alla conclusione che ESISTE un legame significativo tra le due variabili!*

NOTE:

- è importante ricordare che se si trova un legame tra le variabili (H_0 falsa) questo NON INDICA necessariamente che esiste una RELAZIONE CAUSALE tra esse!
- In generale ci sono altri fattori che influenzano entrambe le variabili e provocano l'associazione trovata.
- L'ampiezza di $\boldsymbol{\chi^2}$ non indica la forza del legame tra le variabili, ma piuttosto la forza dell'evidenza che l'ipotesi nulla è falsa.
- Il $\boldsymbol{\chi^2}$ si può applicare solo se l'80% delle celle nella tabella delle frequenze attese è >5 e se ciascuna frequenza attesa è >1, altrimenti altri metodi (per tabelle piccole)

TABELLA χ^2

VALORI CRITICI DELLA DISTRIBUZIONE χ^2 (con gdl da 1 a 30)



Gradi di libertà	*Area della coda superiore*											
	.995	*.99*	*.975*	*.95*	*.90*	*.75*	*.25*	*.10*	*.05*	*.025*	*.01*	*.005*
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	17.240	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	29.339	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	20.843	30.435	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	21.749	31.528	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	22.657	32.620	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	23.567	33.711	39.087	42.557	45.722	49.5B8	52.336
30	13.787	14.954	16.791	18.493	20.599	24.478	34.800	40.256	43.773	46.979	50.892	53.672

CASO PARTICOLARE: TABELLE 2X2

Tabella delle Osservazioni

N= a+b+c+d

	C_1	C_2	totale
R_1	a	b	a+b
R_2	c	d	c+d
totale	a+c	b+d	a+b+c+d

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

da confrontare con $\chi^2_{\alpha,1}$

Per piccoli campioni, si usa la correzione di Yates:

$$\chi^2 = \frac{N(|ad-bc| - \frac{N}{2})^2}{(a+b)(a+c)(b+d)(c+d)}$$

Se più di un elemento della tabella dei valori attesi è <5 si userà il **test esatto di Fisher**.

Test esatto di Fisher

Questo test è adatto anche nel caso in cui si hanno a disposizione dati NON NORMALI. È basato sul calcolo diretto della probabilità che venga 'estratta' proprio quella tabella.
Si calcola:

$$P = \frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{N!\,a!\,b!\,c!\,d!}$$

per ciascuna possibile differente tabella che produca gli stessi totali:

Esempio:

	V_1	V_2	totale
Z_1	1	5	6
Z_2	8	2	10
totale	9	7	16

0 6	1 5	2 4	3 3	4 2	5 1	6 0
9 1	8 2	7 3	6 4	5 5	4 6	3 7
P_1	P_2	P_3	P_4	P_5	P_6	P_7

P1=0.00087, P2=0.0236, P3=0.157, P4=0.327, P5=0.33, P6=0.11, P7=0.01
Si sommano quindi tutti i P_i delle distribuzioni che cadono fino a quella osservata (nell'esempio sino alla SECONDA situazione $P_1 + P_2$), si raddoppia il valore (per avere 2 code) e si confronta direttamente con la probabilità di rifiutare H_0, ovvero con l'α prefissato
Siccome $(P_1+P_2)*2 = 0.049$ è $< \alpha = 0.05$, allora rifiuteremo l'ipotesi nulla e diremo che esiste una relazione tra V e Z.

Tabella 2 x 2, osservazioni dipendenti

Non si usa il χ^2 ma si confronta $Z=\frac{|b-c|-1}{\sqrt{b+c}}$ con Z_α o $Z_{\alpha/2}$

Oppure si fa il quadrato, Z^2, e si confronta con $\chi^2_{\alpha,1}$ = **test di Mc Nemar**

TEST A PIU' CAMPIONI

ANOVA (Analisi della Varianza) (parametrico)

H_0: i campioni provengono dalla medesima popolazione (stesse MEDIA e VARIANZA)

Anziché esaminare la differenza tra le medie si analizza la differenza tra le varianze (altrimenti si utilizza t-Student per i confronti a coppie i campioni, applicando opportune procedure che tengano conto che la probabilità dell'errore di I tipo cresce col numero di confronti => test di Bonferroni, di Tukey, di Scheffè, di Dunnet….).

Requisiti: i Campioni, tra loro indipendenti, provengono da popolazioni Normali (per testare Normalità→Normal Plot) con varianze omogenee (per testarlo => test Bartlett)

La variabilità dei dati è dovuta
- sia dal fatto che i soggetti appartengono a gruppi (o trattamenti) diversi, VARIANZA TRA GRUPPI, var(t)
- sia ad una variabilità individuale tra i soggetti anche di uno stesso gruppo/trattamento (che è la parte dovuta a errori di misura, diversità individuali, fattori non controllabili, ecc.), VARIANZA ENTRO I GRUPPI, var(E)

La varianza totale sarà quindi $var_{tot}=var(t)+var(E)$

Se i campioni provengono tutti dalla medesima popolazione (o da popolazioni 'indistinguibili') allora

$var(t) \sim 0$ e $var(E) \sim$ la varianza del fenomeno

altrimenti $var(t) \gg var(E)$ (tanto maggiore, quanto maggiori saranno le differenze tra i gruppi)

Si userà allora il **test di Fisher** per vedere quanto le varianze siano diverse tra loro:

$$F = \frac{var(t)}{var(E)}$$

Si confronta F con F_{α,n_t,n_E} con n_t e n_E = gradi di libertà del numeratore e denominatore.
Se F risulta essere minore, allora l'ipotesi nulla è vera, ovvero tutti i campioni provengono dalla medesima popolazione.

OSS: Se $var(t) < var(E)$, allora evidentemente H_0 è VERA!

La **varianza entro i gruppi** vale (la devianza media complessiva):

$$var(E) = \frac{\sum_{k=1}^{j} \sum_{i=1}^{n_k} (x_{i_k} - \overline{x_k})^2}{\sum_{k=1}^{j} (n_k - 1)}$$

j è il numero di gruppi; n_k è il numero di elementi del gruppo k-esimo;

$\sum_{k=1}^{j} (n_k - 1)$ è il numero di gradi di libertà ENTRO i gruppi;

$\sum_{i=1}^{n_k} (x_{i_k} - \overline{x_k})^2$ è la devianza nel k-esimo gruppo.

Questa espressione rappresenta una **stima della varianza della popolazione** (di cui i gruppi sono i campioni estratti), MIGLIORE di ciascuna varianza ottenibile separatamente in ciascun gruppo (in quanto tiene conto di un numero maggiore di osservazioni rispetto ciascun gruppo).

La **varianza tra i gruppi** (che si avvicina a zero quanto più le medie sono simili tra loro) sarà:

$$var(t) = \frac{\sum_{k=1}^{j} (\overline{x_k} - \bar{x})^2 \cdot n_k}{j - 1}$$

$\bar{x}$ è la media totale su tutte le osservazioni;

$\sum_{k=1}^{j} (\overline{x_k} - \bar{x})^2$ è la devianza di ciascun gruppo dalla media totale.

NOTA: solitamente è più comodo calcolare var_{tot} e var(t) per poi ricavare var(E) per semplice differenza, con

$$var_{tot} = \sum_{k=1}^{j} \sum_{i=1}^{n_k} {x_{i_k}}^2 \; - \; \frac{(\sum_{k=1}^{j} n_k * \overline{x_k})^2}{\sum_{k=1}^{j} n_k}$$

Si applicherà quindi il test di Fisher:

$$F = \frac{var(t)}{var(E)} < F_{\alpha, n_t, n_E}$$

con $n_t = j - 1$; $n_E = \sum_{k=1}^{j}(n_k - 1) = \sum_{k=1}^{j} n_k - j$

Esempio: Tasso colesterolo in 3 gruppi H0: no diff significative tra le medie

Professionisti (A) n_A=12; $\overline{x_A}$= 285; ${\sigma_A}^2$=3140

Impiegati (B) n_B=14; $\overline{x_B}$= 224; ${\sigma_B}^2$=1380

Agricoltori (C) n_C=10; $\overline{x_C}$= 195; ${\sigma_C}^2$=666

$var(E) = 1772$ $var(t) = 23768$ F=13.4 $n_t = 2$ n_E = 33 $F_{0.05, n_t, n_E}$ =3.31

$\Rightarrow$ Rifiuto H0: Almeno uno dei 3 gruppi è significativamente distinto dagli altri due

Nota: con 2 gruppi il test di Fisher dà gli stessi risultati del t-Student

Test di Bartlett: necessario per valutare se le varianze sono tra loro omogenee e quindi nel caso in cui l'ANOVA dia valida l'H0.

H_0: le varianze sono stime indipendenti di varianza di una popolazione e le differenze sono dovute al caso.

Si valuta:

$$\frac{A}{B} = \frac{2.3026 \cdot \left[(n-k)\log_{10}\bar{s}^2 - \sum_{i=1}^{k}(n_i - 1)\log_{10} s_i^2\right]}{1 + \frac{1}{3(k-1)} \cdot \left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right]}$$

k è il numero di gruppi; n_i è il numero di campioni nell'i-simo gruppo; $n = \sum_{i=1}^{k} n_i$; s_i^2 è la varianza nell'i-simo gruppo; $\bar{s}^2$ è la varianza complessiva.

Il rapporto si distribuisce come una χ^2 con k-1 gradi di libertà.

Se $\chi^2_{A/B} < \chi^2{}_{\alpha,k-1}$, allora l'ipotesi nulla va accettata e le varianze sono OMOGENEE

Test di Kruskal-Wallis (non Parametrico)

se le ipotesi richieste per l'analisi della varianza non sono soddisfatte
ma almeno si hanno: a) Indipendenza tra i campioni; b) Numerabilità;
allora si può utilizzare questo test che rappresenta il caso generale del test di Mann-Whitney

H_0: i gruppi appartengono alla stessa popolazione (le differenze tra le sommatorie dei ranghi sono attribuibili solo al caso)

Per valutare il test, si prendono le N osservazioni tutte insieme (tutti i gruppi), si valutano i ranghi e quindi la statistica:

$$H = \frac{12 \cdot \sum_{i=1}^{k} n_i (\bar{R}_i - \bar{R})^2}{N(N+1)}$$

Con:

n_i numero di osservazioni nell'i-esimo gruppo;
R_i sommatoria dei Ranghi dell'i-esimo gruppo;
$\bar{R}_i$ rango medio dell'i-esimo gruppo;
$\bar{R}$ media dei Ranghi;
k numero di gruppi.

H cresce col crescere della variazioni fra i gruppi e si distribuisce come una $\chi^2_{\alpha,k-1}$ (ad una sola coda perché H può solo crescere)

Se $H > \chi^2_{\alpha,k-1}$, si respingerà l'ipotesi nulla => esiste una differenza significativa tra i gruppi

VERIFICA DELLE IPOTESI (sintesi)

1 CAMPIONE:

Test sulla media: Popolaz con distrib Normale e nota σ => Z – test

Popolaz con distrib Normale ma ignota σ => t-Student

Ignota distribuzione => Sign / Wilcoxon signed rank sum test

Test sulla frequenza: Tabelle con sufficiente numerosità => χ^2

Tabelle con bassa numerosità => Kolmogorov

2 CAMPIONI:

Test su differenza di medie: distrib Normale e nota σ => Z – test

distrib Normale ma ignota σ => t-Student + Fisher (se H0 è vera)

Ignota distribuzione => Wilcoxon-Mann-Whitney

Test sulle diff di frequenza (tabelle 2 x 2): test Z, Z modificato, χ^2 + Yates, test esatto di Fisher, test di McNemar

3 O PIU' CAMPIONI:

Test su differenze di varianze: Popolaz con distr Normale => ANOVA + Bartlett

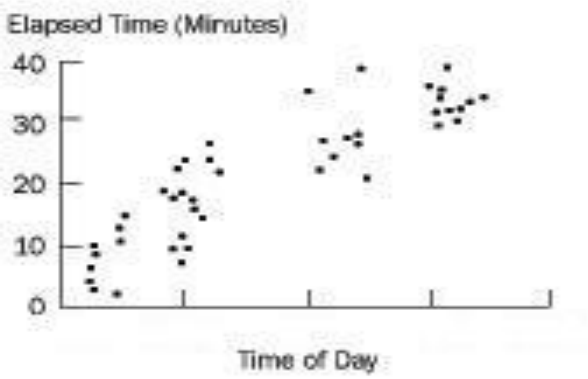
Ignota distribuzione => Kruskal – Wallis

Tabelle con 2 variabili (tabelle r x c) => χ^2

RELAZIONI TRA FENOMENI (VARIABILI)

RELAZIONE TRA 2 VARIABILI

Retta di Regressione (relazione lineare): si può utilizzare per predire Y, data una qualsiasi X



Si parte da uno SCATTER DIAGRAM

Assunti:

- i valori di Y (variabile DIPENDENTE) devono essere distribuiti come una Normale per ciascun valore di x
- la varianza di Y deve essere identica per ciascuna x, ovvero deve essere verificata l'OMOSCHEDASTICITA'
- la relazione deve essere lineare.

NON è necessario che entrambe le variabili siano aleatorie (casuali), né che X sia Normale!

Per verificare i 3 assunti, si calcola la relazione: $Y = a + bX$

a e *b* opportuni per minimizzare le distanze verticali (RESIDUI): $\sum_{i=1}^{n}(y_i - Y(x_i))^2$, dove y_i sono i valori osservati e $Y(x_i)$ sono quelli teorici; si ricava:

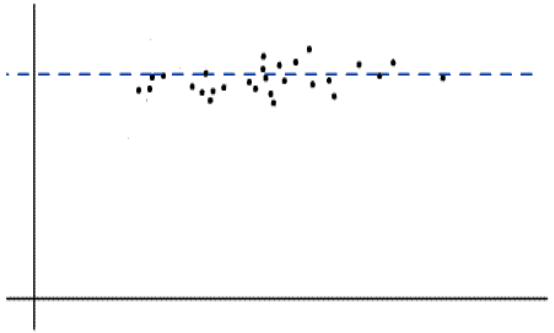
$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{cov(x,y)}{var(x)} \qquad a = \bar{y} - b\bar{x}$$

Un discorso simile si può effettuare scambiando le 2 variabili considerando la X come variabile dipendente e la Y indipendente
Si dovrà quindi calcolare:

$$X = a' + b'Y$$

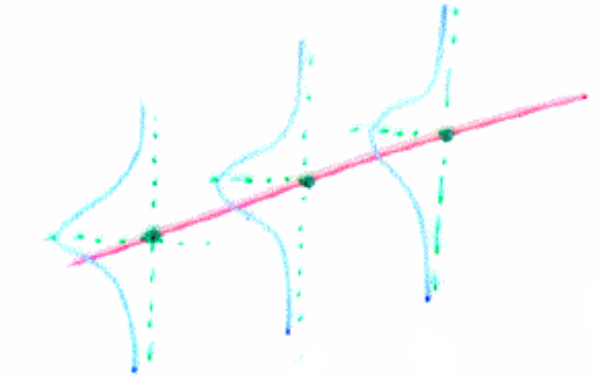
Distanze calcolate in orizzontale anziché in verticale
Se gli assunti sono veri, i residui devono essere distribuiti Normalmente e questo si può testare con il Normal Plot

Residui =>



Nota: *b* e *b'* possono essere utilizzati come indici di concordanza, indicando quanto cresce in media una variabile al crescere unitario dell'altra. Esse rappresentano anche asimmetria nel rapporto tra variabili (=> coeff. correlazione)

Nell'ipotesi che la distribuzione di Y in corrispondenza ad ogni x_i sia normale:



nell'ipotesi che la varianza sia uguale per tutti i punti (OMOSCHEDASTICITA') si possono valutare le deviazioni standard della pendenza (*b*) e dell'intercetta (*a*):

$$\sigma_a = \sigma \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}} \qquad \sigma_b = \frac{\sigma}{\sqrt{\sum(x-\bar{x})^2}}$$

Per stimare σ ($\hat{\sigma} = errore\ standard\ della\ stima$) utilizzo S_{yx}:

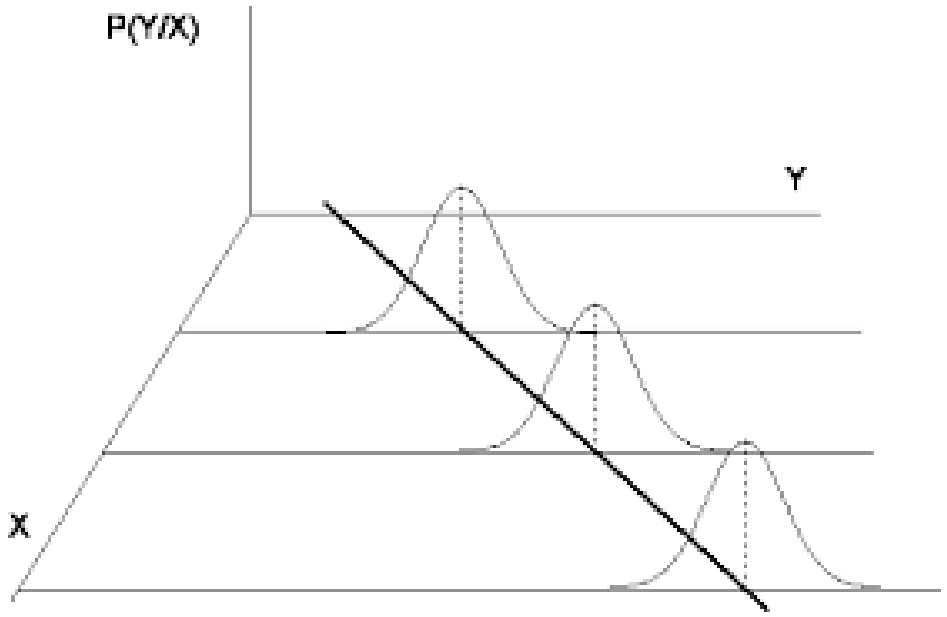
$$\hat{\sigma} = S_{yx} = \sqrt{\frac{\sum(y_i - Y(x_i))^2}{n-2}}$$

Per testare la bontà dei valori della pendenza e dell'intercetta trovati si userà il t-test
Nel caso della **pendenza** (b) si avrà:

$$t = \frac{b - \beta_0}{\sigma_b}$$

da confrontare con $t_{\alpha,n-2}$. β_0 è il valore da testare *(Es. H_0: non esiste relazione tra qual è la probabilità che un campione con particolari y in x prefissate, dia una pendenza $b \geq \beta_0$)*.
Se $\beta_0 = 0$, si testa l'ipotesi che non vi sia alcun legame tra x e y.
L'intervallo di confidenza della pendenza sarà quindi:

$$b \pm t_{\alpha,n-2} \cdot \sigma_b$$


P(Y/X)
Y
X

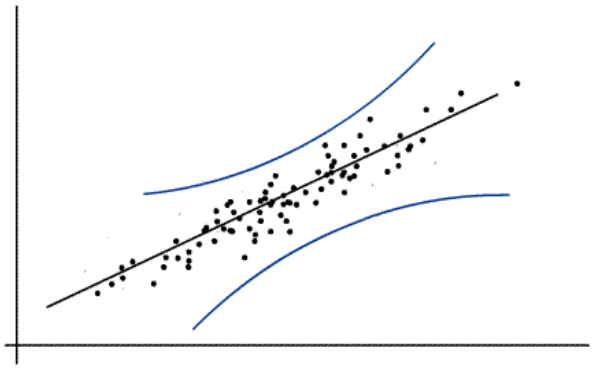
Per l’**intercetta** (a) si usa:

$$t = \frac{a - \alpha_0}{\sigma_a}$$

con $t_{\alpha,n-2}$ e α_0 un valore da testare, come per la pendenza.
L’intervallo di confidenza dell’intercetta sarà:

$$a \pm t_{\alpha,n-2} \cdot \sigma_a$$

Limiti di confidenza rispetto la retta si ottengono combinando i possibili valori (a intervalli) di a e b



INTERPRETAZIONI E LIMITI
affinchè i risultati siano significativi
- le osservazioni devono essere indipendenti (Es: 1 sola misura per ogni individuo)
- non si deve usare la relazione oltre il campo delle x da cui si è partiti (no estrapolazioni)
- data x, si può predire Y, ma non viceversa
- gli intervalli di confidenza per b indicano l’incertezza nella forza della relazione tra y e x
- la retta di regressione indica quanto della variabilità di y può essere spiegata (in modo lineare) da x e quanta variabilità resta non spiegata (quota parte dovuta a rumore)

Coefficiente di correlazione (lineare)

Per uniformare le informazioni delle 2 rette di regressione, ovvero non considerare più una variabile dipendente e una indipendente, ma entrambe aleatorie, si utilizza il coefficiente di correlazione lineare di Bravais-Pearson:

$$r = \pm\sqrt{b' \cdot b} = \frac{cod(x,y)}{dev(x)dev(y)} = \frac{cov(x,y)}{var(x)var(y)} =$$

$$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)\sigma_x\sigma_y} =$$

$$= \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sqrt{\left[\sum x_i^2 - (\sum x_i)^2/n\right]\left[\sum y_i^2 - (\sum y_i)^2/n\right]}}$$

$$-1 \leq r \leq 1 \quad \text{(adimensionale)}$$

È una misura simmetrica che dà informazione sull'interdipendenza tra le variabili, ovvero una misura della dispersione dei dati rispetto ad un andamento lineare.
Se r=0, non c'è correlazione, più $r \to 1\ (-1)$ maggiore è la correlazione.

Per vedere se è significativamente distante da 0, si valuta:

$$t = \frac{r}{\sqrt{(1-r^2)/n-2}}$$

e si confronta con $t_{\alpha,n-2}$.

Assunti:
- per calcolare l'intervallo di confidenza è necessario che sia la x che la y provengano da distribuzioni normali (W-test per valutarlo)
- le osservazioni devono essere indipendenti (1 sola osservazione per ciascun individuo, NON ripetute!), altrimenti **l'ANALISI NON E' VALIDA**!

Una volta trovato che r è significativamente vicino a 1 (o -1) **non si può** direttamente **dire** se x dipende da y o viceversa o addirittura che x e y dipendano da un terzo fattore

Esiste anche il **COEFFICIENTE DI DETERMINAZIONE**: $r^2 = b \cdot b'$ che esprime la variabilità di Y, attraverso la variabilità di X.

$1 - r^2$ esprime la porzione di varianza di Y che dipende da fattori diversi da X.

Esistono anche correlazioni basate sui Ranghi (NON parametriche), con ipotesi iniziale di sola indipendenza, non di Normalità.
- Coefficiente di Spearman=r_s, si calcola come r ma sui ranghi delle x e delle y
- Coefficiente di Kendall=r_t

APPROCCIO MATRICIALE

regressione lineare $\boldsymbol{Y = XB + \varepsilon}$

$$Y = \begin{vmatrix} y_1 \\ \vdots \\ y_n \end{vmatrix}, X = \begin{vmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{vmatrix}, \mathrm{B} = \begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix}, \varepsilon = \begin{vmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{vmatrix}$$

←con valor medio nullo e varianza = σ

↑ la colonna con "1" indica che l'intercetta è inclusa in questa matrice

Ai minimi quadrati avremo (noti che siano X e Y):

$$\widehat{\mathrm{B}} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

$$E(Y) = X\mathrm{B}$$

$$\varepsilon\varepsilon^T = \begin{vmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \dots \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \dots \\ \vdots & \vdots & \varepsilon_n^2 \end{vmatrix} \Longrightarrow E(\varepsilon\varepsilon^T) = \begin{vmatrix} \varepsilon_1^2 & \dots & 0 \\ 0 & \varepsilon_2^2 & 0 \\ 0 & \dots & \varepsilon_n^2 \end{vmatrix}$$

perché le ε sono indipendenti => $E(\varepsilon_i\varepsilon_j) = 0$ per $i \neq j$

correlazione parziale e multipla

$$R = \begin{vmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \cdots & \cdots \\ r_{k1} & \cdots & \cdots & 1 \end{vmatrix}$$

regressione multipla lineare: $\boldsymbol{Y = XB + \varepsilon}$

$$X = \begin{bmatrix} x_0 & \cdots & x_{k1} \\ \vdots & \ddots & \vdots \\ x_0 & \cdots & x_{kn} \end{bmatrix}, \mathrm{B} = \begin{vmatrix} \beta_0 \\ \vdots \\ \beta_n \end{vmatrix}$$

$$X \cdot X^T = \begin{vmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & & \vdots \ \vdots \ \vdots & \\ \vdots & & & \vdots \ \vdots \ \vdots & \\ \sum x_{ki} & \vdots & \cdots & \cdots & \sum x_{kn}^2 \end{vmatrix} \Longrightarrow \hat{\beta} = (X^T \cdot X)^{-1} X^T Y$$

Come scegliere il test piu' adatto per la verifica di ipotesi

Scala di misura	*Due gruppi con sogg. diversi*	*>2 gruppi con sogg. diversi*	*Prima e dopo con gli stessi sogg.*	*Piu' tempi negli stessi sogg.*	*Associazione fra 2 variabili*
Intervallare ("normale")	*T-test di Student*	***ANOVA***	*Paired t test*	***ANOVA per misure ripetute***	*Correlazione di Pearson e regressione lineare*
Nominale	*Chi quadro*	*Chi quadro*	*Test di McNemar*	*Test Q di Cochrane*	*Coefficiente di contingenza (Test K di Kendall)*
Ordinale	*Test per la somma dei ranghi di Mann-Whitney*	*Test di Kruskal-Wallis*	*Test di Wilcoxon*	*Test di Friedman*	*Correlazione dei ranghi (test di Spearman)*

I primi 23 numeri fattoriali

1! =	1
2! =	2
3! =	6
4! =	24
5! =	120
6! =	720
7! =	5.040
8! =	40.320
9! =	362.880
10! =	3.628.800
11! =	39.916.800
12! =	479.001.600
13! =	6.227.020.800
14! =	87.178.291.200
15! =	1.307.674.368.000
16! =	20.922.789.888.000
17! =	355.687.428.096.000
18! =	6.402.373.705.728.000
19! =	121.645.100.408.832.000
20! =	2.432.902.008.176.640.000
21! =	51.090.942.171.709.440.000
22! =	1.124.000.727.777.607.680.000
23! =	25.852.016.738.884.976.640.000