

Metodi statistici per la Neuropsicologia Forense

A.A. 2023/2024

APPENDICE - concetti base di statistica

Giorgio Arcara

IRCCS San Camillo, Venezia
Università degli Studi di Padova





I livelli di misura

- Nominale
- Ordinale
- Intervallo
- Rapporti



I livelli di misura

- Nominale

Es. neuropsicologico. Patologia del paziente: Parkinson vs Alzheimer

I livelli di misura

- Ordinale

Es. Neuropsicologico. Scolarità dei partecipanti espressa come titolo di studio (scuola elementare, superiore, Università). Esiste un'ordinalità ma non possiamo effettuare operazioni aritmetiche sui valori della scala.

I livelli di misura

- Intervalli

Es. classico Temperatura in Celsius.

Possiamo effettuare operazioni come sottrazione e addizione (es. 45° sono 15° in più di 30° . 25° sono 15° in più di 10°). Non può essere usata moltiplicazione: es. non ha senso dire che 30° sono il doppio di 15° .

In queste scale lo 0 è arbitrario.

(es. Neuropsicologico: QI di una persona oppure la scolarità espressa in anni.)

I livelli di misura

- Rapporti

Es. Temperatura in kelvin. Esiste uno 0 assoluto. In questo caso ha senso dire che 20 gradi Kelvin sono il doppio di 10 gradi Kelvin.

(Esempio neuropsicologico: Frequenza di un certo comportamento. Potrebbe essere 0)

Note sui livelli di misura

In neuropsicologia clinica, spesso i test che si utilizzano risponderebbero solo ai criteri di *scala ordinale*. Sarebbe in questo caso importante usare metodi statistici adatti per scale ordinali (spesso non parametrici).

Nonostante ciò è molto comune utilizzare invece (anche su scale ordinali) metodi statistici *parametrici* e, sostanzialmente considerare i punteggi come se venissero da una scala ad intervalli. Ad esempio si usano *regressioni* o *correlazioni di Pearson*, etc. (che si potrebbero usare solo con scala ad intervalli o rapporti).

Note sui livelli di misura

Considerare un punteggio su scala ad intervalli ha altre conseguenze (es. si considera che avere questi I soggetti s1 ed s2, testati due volte nel tempo a t1 e t2 in un test che ha un punteggio che va da 0 a 30

	t0	t1	differenza
s1	28	30	+2
s2	8	10	+2

Se la scala fosse ad intervalli, potremmo dire che s1 ed s2 sono migliorati esattamente nello stesso modo, definito dalla differenza (+2 in entrambi I casi). In una scala ordinale questo non avrebbe senso: Entrambi sono migliorati, ma non possiamo dire se nello stesso modo.

* NOTA: pensando in termini di Item Response Theory il valore +2 potrebbe avere un diverso significato, perché la probabilità di ottenere un miglioramento da 8 a 10, potrebbe essere ben diversa dalla probabilità di ottenere un miglioramento da 28 a 30. Ad esempio potrebbe essere molto facile passare da 8 a 10, mentre molto difficile ottenere il punteggio massimo e quindi da 28 a 30

Notazione statistica

In generale nelle formule statistiche lettere latine indicano parametri (es. Media, deviazione standard) relativi al campione, mentre lettere greche indicano parametri della popolazione

s = varianza del campione

σ = (sigma) varianza della popolazione

Campione vs Popolazione

Popolazione: l'insieme di entità su cui si vogliono effettuare delle inferenze.

Campione: un subset di una popolazione.

Esempi:

Popolazione: gli studenti dell'Università di Padova

Campione: 50 studenti selezionati a random.

Popolazione: tutti gli abitanti dell'Italia


Campione: il campione miei dei dati normativi per la taratura del test

Statistica descrittiva vs Statistica inferenziale

La statistica descrittiva è quella parte della statistica che ci fornisce strumenti per descrivere i dati del nostro campione (es. Media, deviazione standard, etc), quindi sui *dati osservati*.

La statistica inferenziale è quella parte della statistica che ci fornisce strumenti per trarre delle inferenze a partire da dati da campioni (cioè dai dati osservati) sulle popolazioni da cui questi dati sono stati estratti, con certe probabilità di errore note (fatte alcune assunzioni).

Nella statistica inferenziale si tiene conto di errore e variabilità del campionamento per poter fare delle inferenze.



Popolazione: l'insieme di entità su cui si vogliono effettuare delle inferenze.

Campione: un subset di una popolazione.

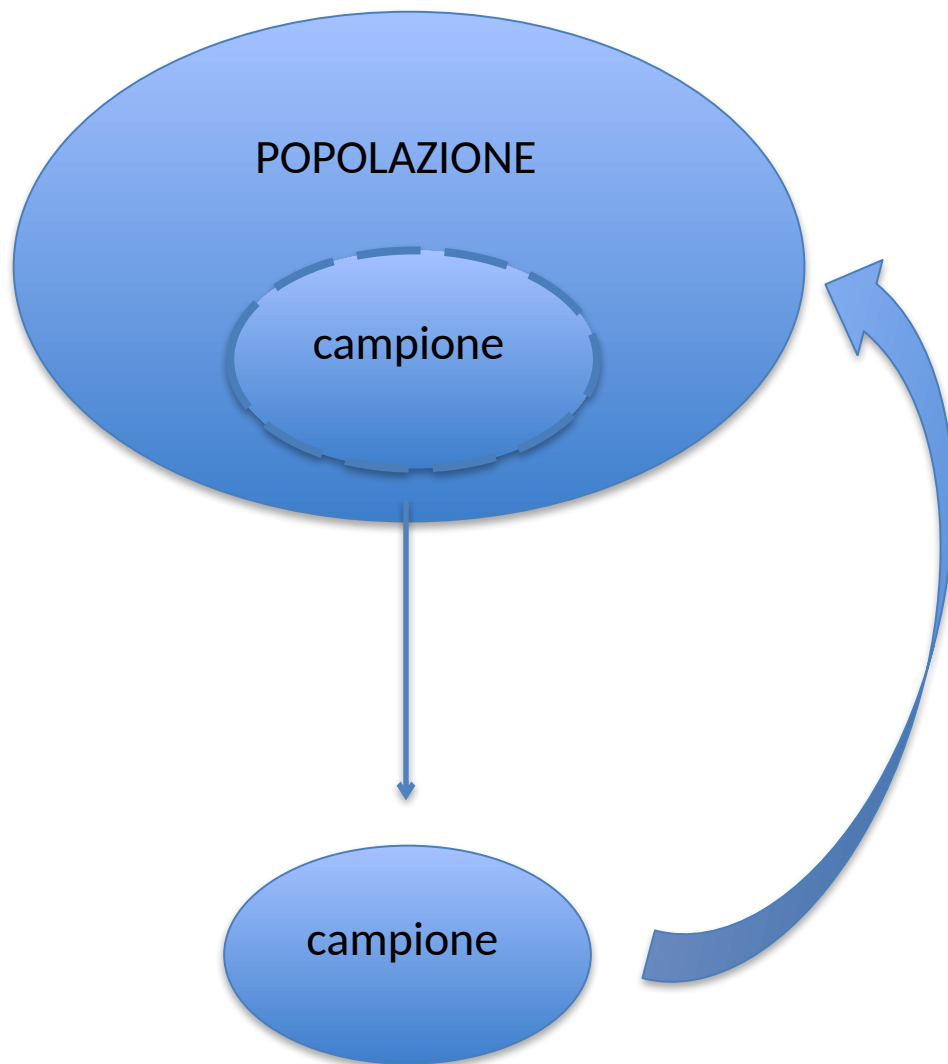
Esempio:

Popolazione: gli studenti dell'Università di Padova

Campione: 50 studenti selezionati a random.

Popolazione: tutti gli abitanti dell'Italia

Campione: il mio campione dei dati normativi per la taratura del test



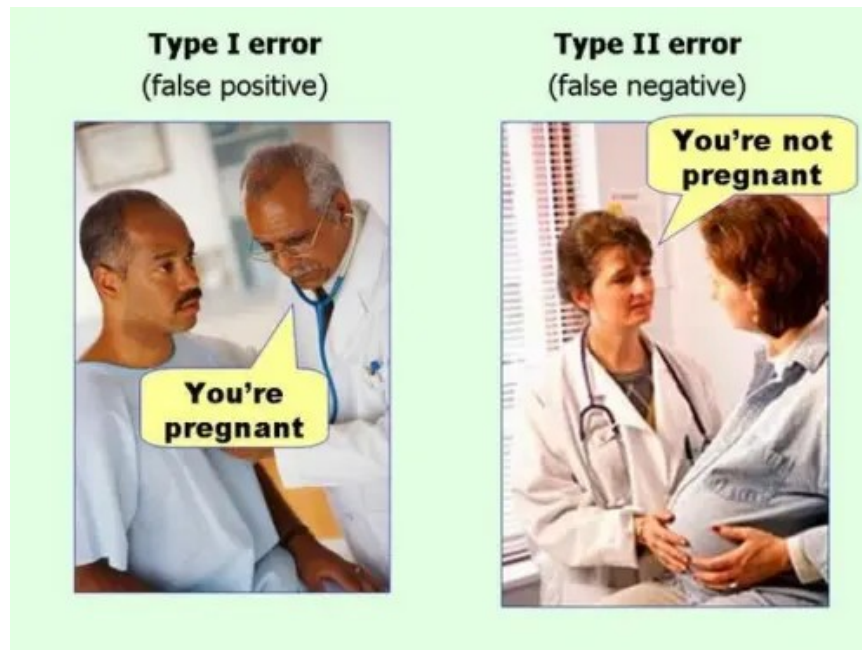
l'interesse nella statistica inferenziale non è nel campione, ma nella popolazione da cui è estratto.

Errore di tipo I ed errore di Tipo II

Nella statistica inferenziale è possibile fare due principali errori;

L'errore di **tipo 1** è l'errore di rifiutare l'ipotesi nulla quando essa è vera (detto anche alpha)

L'errore di **tipo 2** è l'errore di accettare l'ipotesi nulla quando essa è falsa.
Il complementare dell'errore di primo tipo è detta "Potenza"



Errore di tipo I ed errore di Tipo II

Il rischio di commettere degli errori è legato all'utilizzo di soglie (es. Valori critici di t o di z) , che hanno un effetto complementare sul rischio di errore.

Più e basso l'errore di tipo 1, più alto è l'errore di tipo 2, viceversa più alto è errore tipo 1, più e basso errore di tipo 2

Si vedano slides su sensibilità/specificità in cui (la conoscenza di entrambe le condizioni) facilita la comprensione dell'effetto di spostare le soglie e cambiare le probabilità di errore.

Vedi anche link per relazione tra errore di primo tipo e Power (cioè $1 - \text{errore di tipo II}$)

<https://rpsychologist.com/d3/nhst/>

Errore di tipo I ed errore di Tipo II

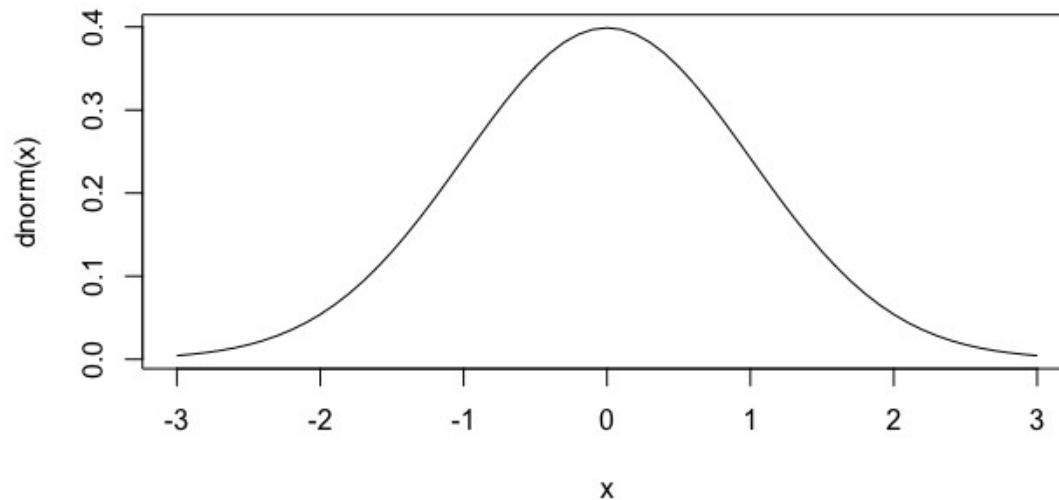
Nel contesto della psicologia forense **l'ipotesi nulla è quasi sempre che il soggetto esaminato sia sano**

Nel contesto della psicologia forense **l'ipotesi alternativa è che il paziente abbia un deficit o un danno.**

Dal momento che è molto difficile (proprio a livello epistemologico/metodologico) capire cosa significa avere un deficit o un danno (vedi anche altre slides), in neuropsicologia clinica/forense spesso l'attenzione è sul controllare adeguatamente l'errore di tipo 1

La distribuzione normale

Tra le distribuzioni di dati una delle più comuni è la distribuzione normale (o gaussiana)



La distribuzione normale

È una distribuzione di probabilità definita (da un punto di vista matematico) da questa formula.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$f(x)$ = probability density function

σ = standard deviation

μ = mean

$f(x)$, oltre ai valori di x dipende da due soli altri parametri: μ = media, e σ = deviazione standard (e e π , le altre sono costanti).

https://en.wikipedia.org/wiki/Normal_distribution

La distribuzione normale

È una distribuzione rilevante perché molto presente in vari fenomeni naturali.

In psicometria (e quindi in neuropsicologia) è rilevante invari ambiti, ma soprattutto quando si parla di errore di misurazione

L'errore di misurazione è spesso *assunto* avere distribuzione normale con media pari a 0 e varianza σ^2 (sigma²)

$$E = N(0, \sigma^2)$$

La distribuzione normale

In R per simulare n valori da una distribuzione normale con media m e deviazione standard* s

```
rnorm(n, mean=m, sd = s)
```

Esempio con 100 valori a media zero e deviazione standard 2

```
rnorm(100, mean=0, sd = 2)
```

* NOTA: deviazione standard (s) è uguale a radice quadrata varianza $s = \sqrt{s^2}$

La correlazione

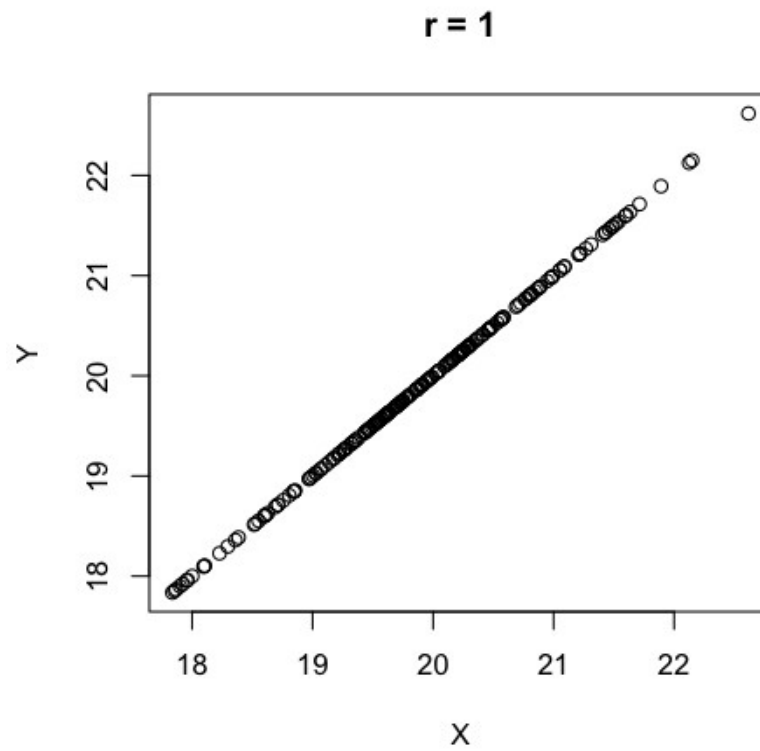
È una delle statistiche più importanti perché nella sua semplicità contiene molte informazioni

La correlazione esprime l'associazione tra due variabili continue.

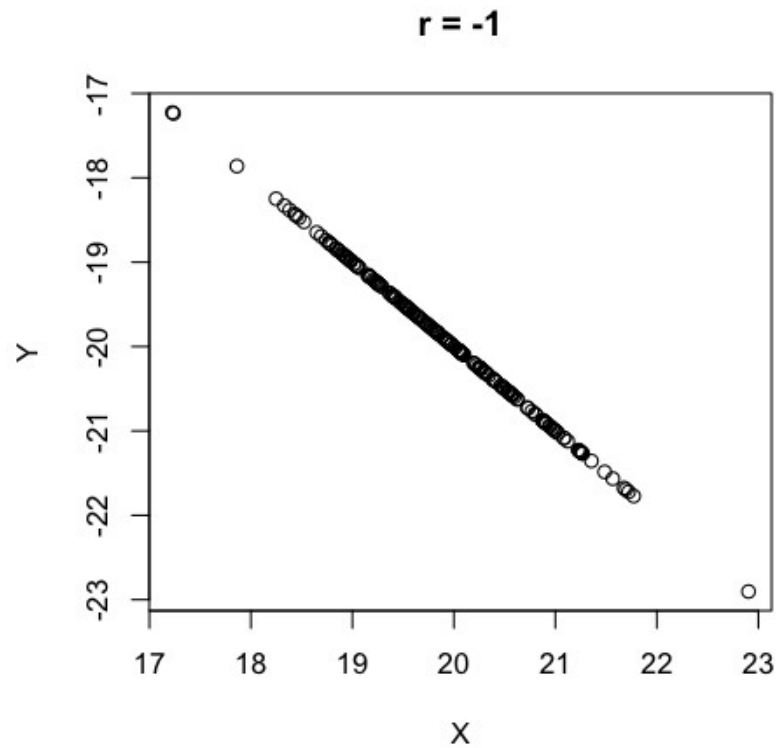
Ha un valore che va da -1 a 1.

- Un valore di 1 indica un correlazione massima positiva: all'aumentare di una variabile aumenta l'altra.
- Un valore di 0 indica nessuna correlazione. I valori delle due variabili non sono associati
- Un valore di -1 indica correlazione massima negativa. All'aumentare dei valori di una variabile diminuisce l'altra

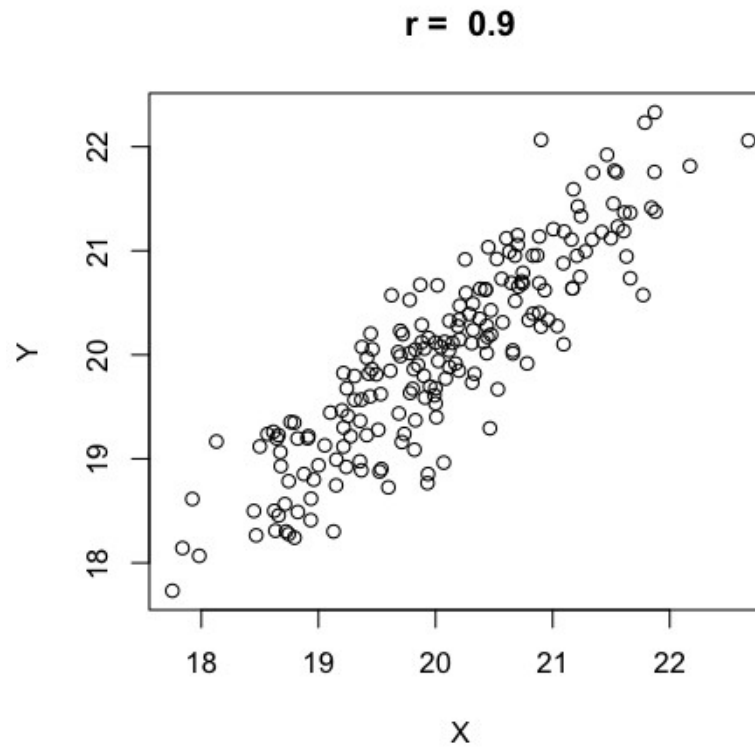
La correlazione



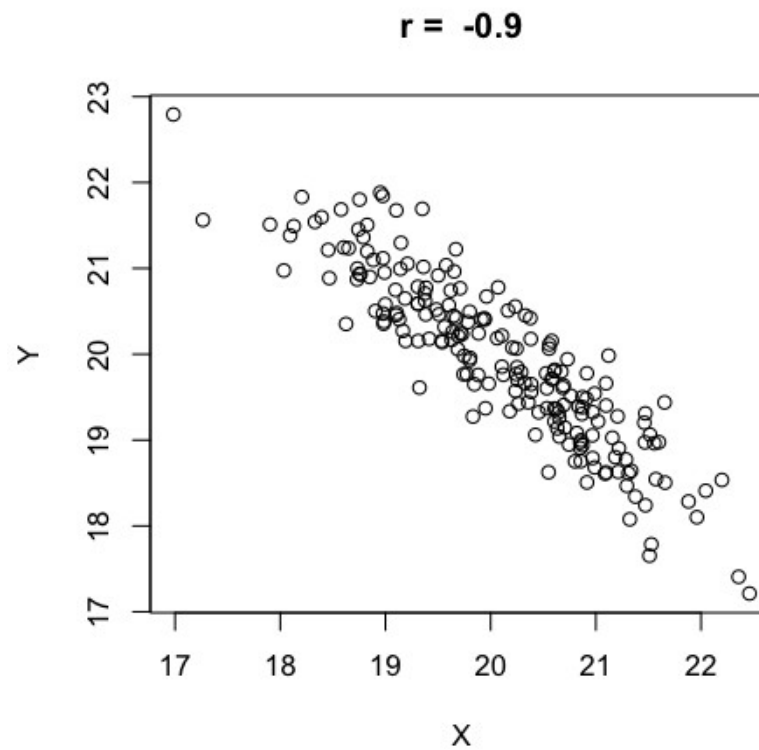
La correlazione



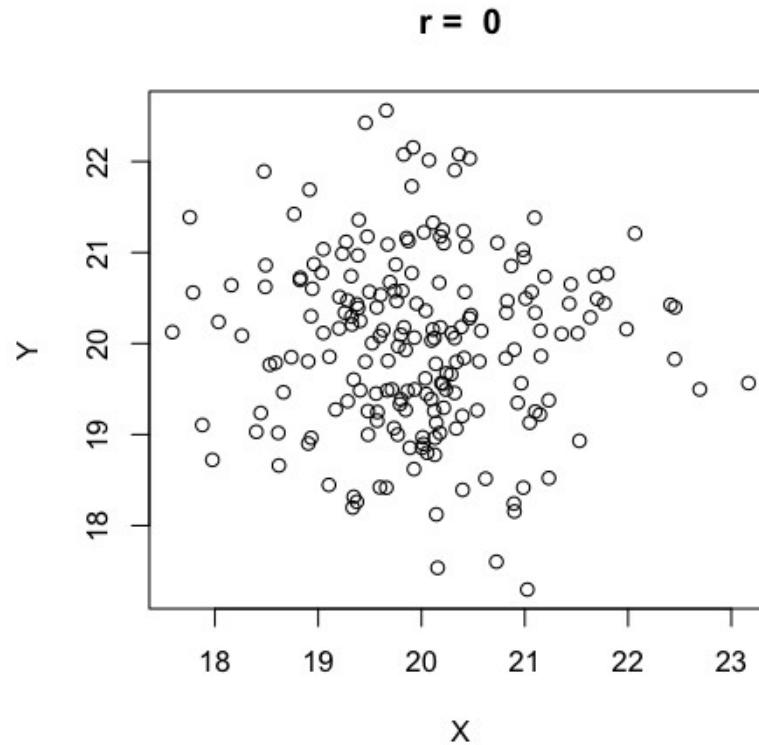
La correlazione



La correlazione



La correlazione



Guardare codice correlato per vedere effetti di diverse correlazioni

La correlazione

Per l'interpretazione della correlazione esistono delle tabelle di riferimento
Relative all'effect size

(Cohen, 1988), Table 6

Valore di correlazione r	Effect size
$0.10 < r < 0.29$	weak
$0.30 < r < 0.49$	moderate
$0.50 < r < 1.0$	high

La correlazione

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

\bar{x} = mean of the values of the x-variable

y_i = values of the y-variable in a sample

\bar{y} = mean of the values of the y-variable

oppure

$$\rho (X,Y) = \text{cov} (X,Y) / \sigma_X \sigma_Y$$

oppure

$$r = Z_x * Z_y$$

Dove $Z_n = \frac{n - \bar{n}}{\sigma_n}$

La correlazione

La correlazione di Spearman è semplicemente la correlazione di Pearson effettuata sui *ranghi* e non sui valori grezzi.

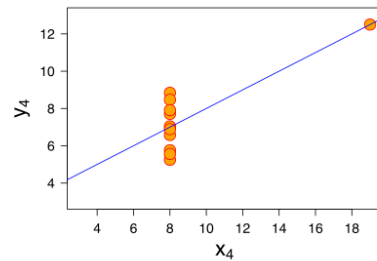
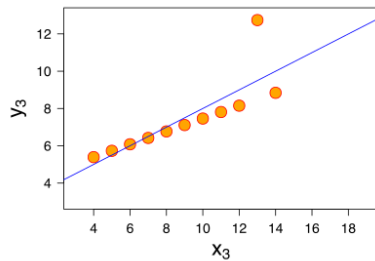
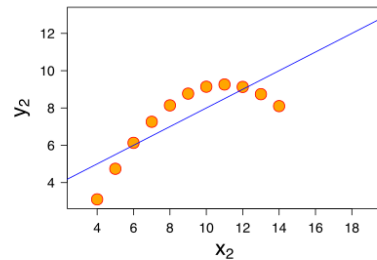
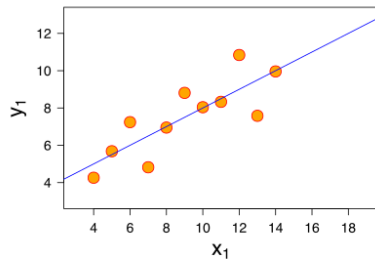
Il rango indica la posizione ordinale dei valori, se ordinati dal più basso al più alto) (es. Dati 100 elementi, il valore più basso avrà rango 1, il valore più alto rango 100).

La correlazione di Spearman va usata quando i dati sono su scala ordinale o quando sono presenti outliers

La correlazione

Limiti della correlazione

L'Anscombe Quartet è un classico esempio per mostrare che a parità di correlazione possono esserci dati molto diversi.



Tutti queste coppie di variabil hanno lo Stesso valore di correlazione r .

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

La correlazione

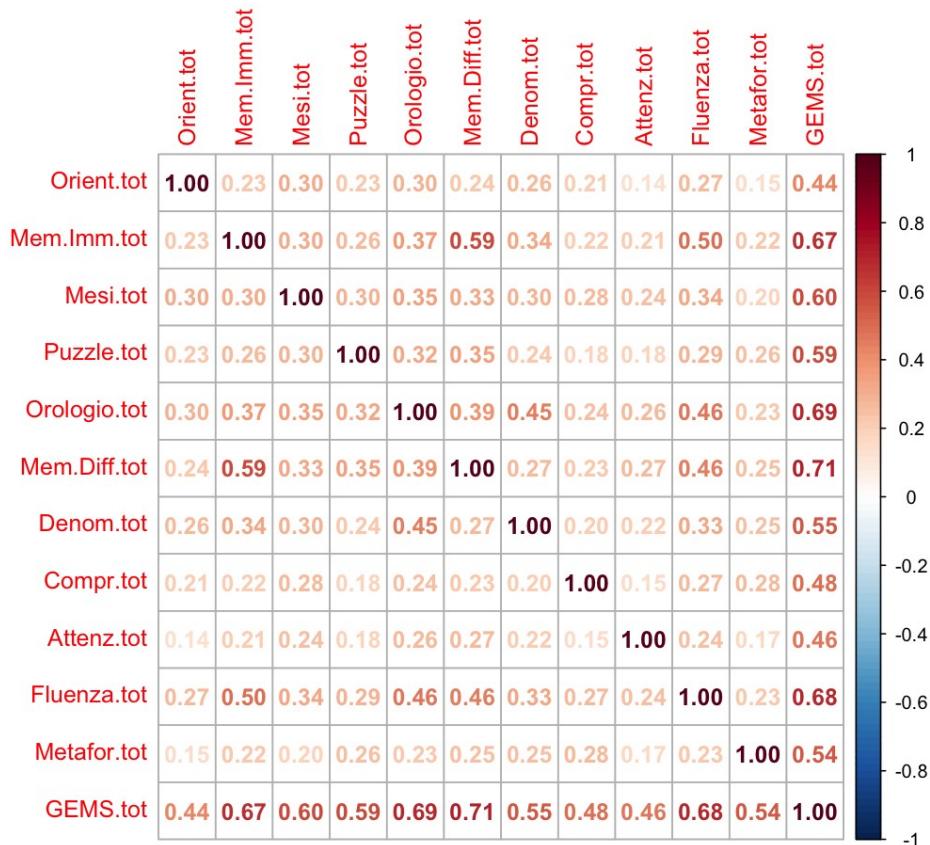
Altro limite correlazioni

La correlazione è sempre una misura che prende in considerazione solo due variabili alla volta. Queste potrebbero però riflettere pattern di associazioni legate ad uno (o più) fattori latenti. Per trovare queste relazioni servono altre tipologie di analisi.

(es. Analisi Fattoriale, Modelli di Equazioni strutturali etc.)

La correlazione

Altro limite correlazioni



Esempio:

Questa matrice mostra molte correlazioni positive, sono scollegate fra loro? Oppure c'è un pattern?

(Serve analisi apposite per rispondere, vedi nelle slides su validità e affidabilità formule, analisi fattoriale)

La correlazione e Il metodo correlazionale

La correlazione, intesa come analisi statistica, è quasi sempre legata al **metodo correlazionale** di indagare l'associazione tra due variabili.

Questo metodo si basa sullo studio dell'osservazione di due (o più variabili) senza una loro diretta manipolazione in genere seguita da analisi statistica. L'analisi usata è appunto spesso correlazione, ma possono esserne usate anche altre (es. regressione)

Il metodo correlazionale ha una serie di limiti inerenti sul tipo di evidenze che si possono trarre. I dati che si utilizzano in studi che sviluppano i test (e quindi indagano validità o affidabilità) non possono che usare il metodo correlazionale

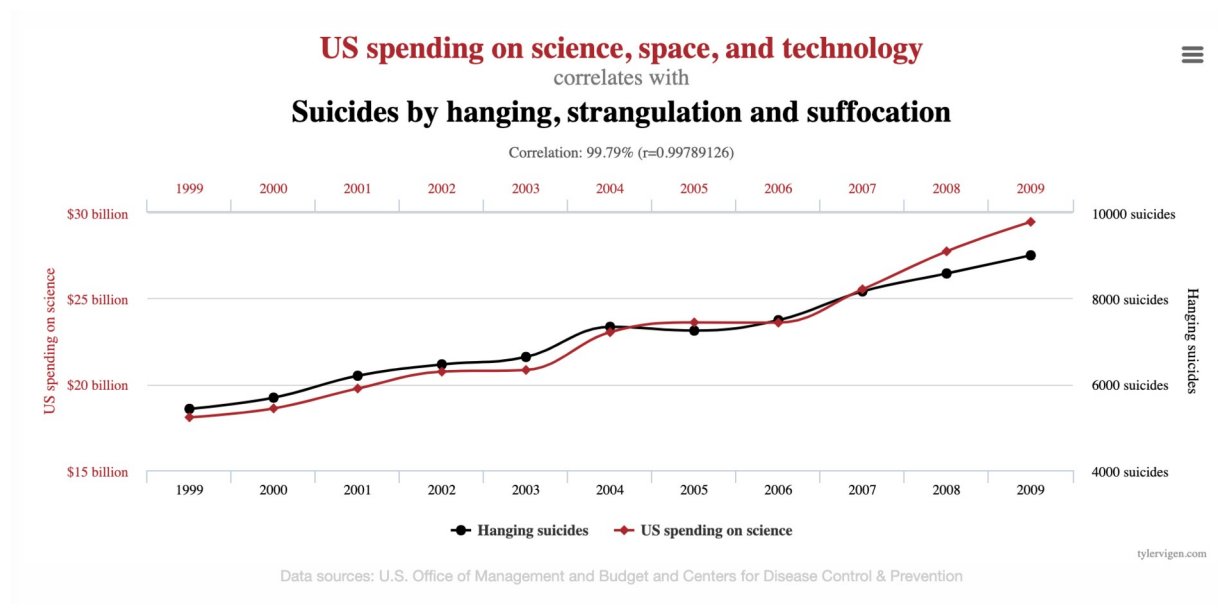
Attenzione che talvolta quando si parla di correlazione si parla dell'analisi statistica, mentre a volte si parla del metodo correlazionale (o di evidenze di natura correlazionale) ed è facile confondere i due termini.

Quasi sempre le evidenze che si hanno a disposizione in neuropsicologia clinica (e forense) sono di tipo correlazionale, dal momento che non si possono manipolare le variabili, ma semplicemente osservarle e studiarne le relazioni/associazioni.

La correlazione

Correlazione non implica una relazione causa-effetto di ciò che è osservato.

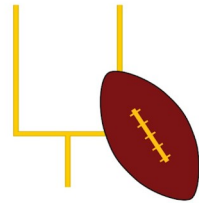
https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation



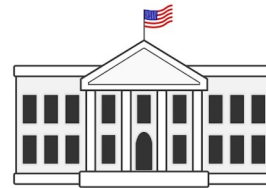
<https://www.tylervigen.com/spurious-correlations>

La correlazione

① Random coincidence



Football team's record

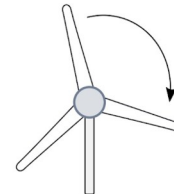


Presidential election

② Reverse causality



Wind

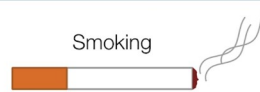


Spinning windmill

③ Confounding variable



Alcohol consumption

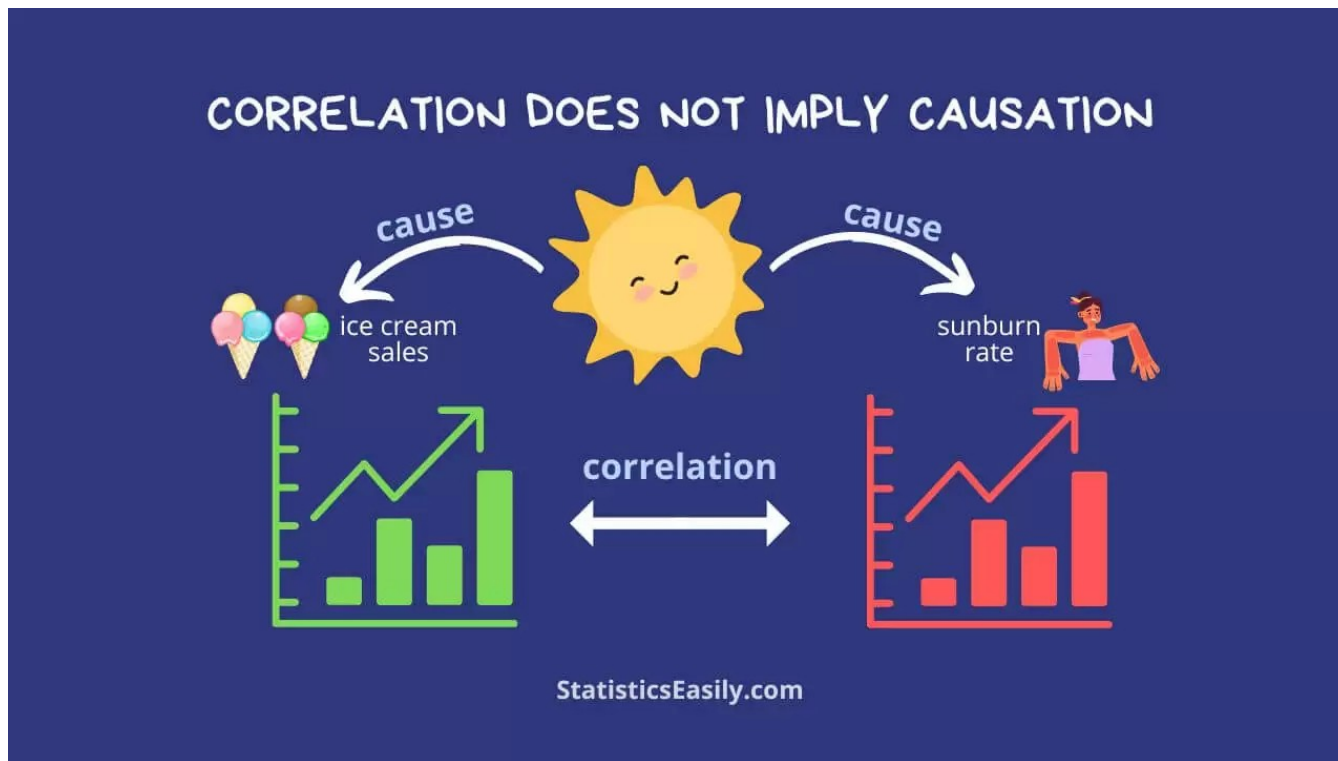


Smoking



Lung cancer

La correlazione



<https://statisticseasily.com/correlation-vs-causality/>

La correlazione

In neuropsicologia forense bisogna stare attenti alla *plausibilità* di un nesso causale con la sua attuale dimostrazione.

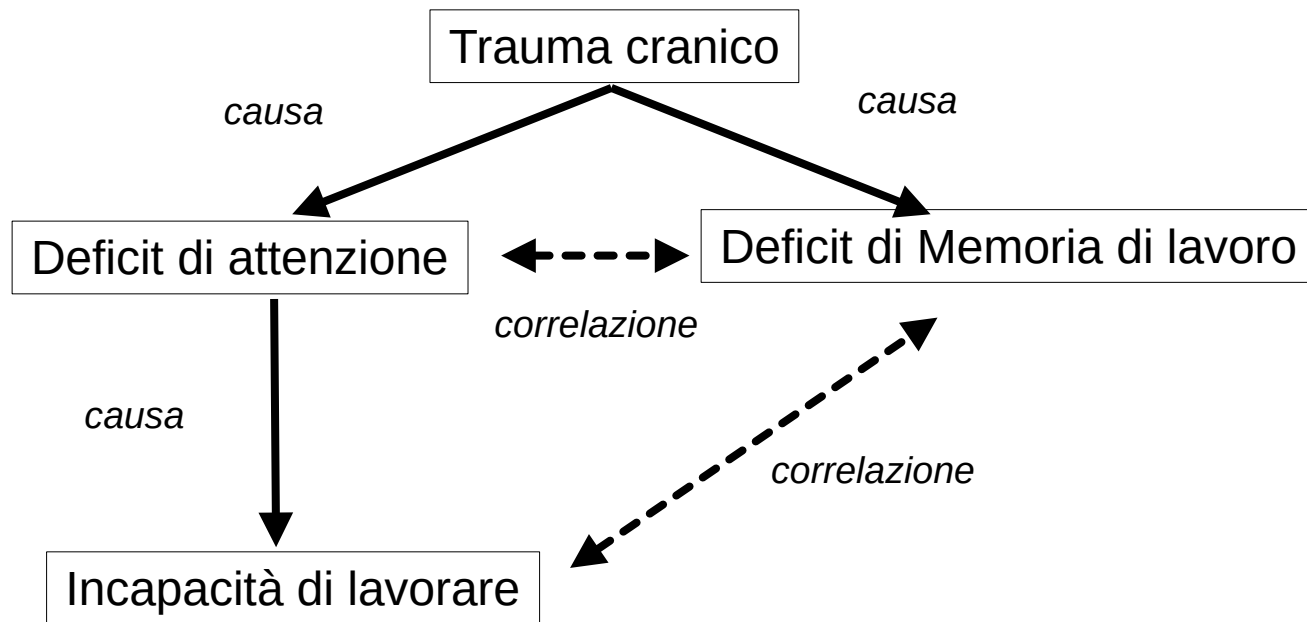
La più comune inferenza è quello che un deficit di base (es. di memoria di lavoro), abbia una *causa* su un aspetto della vita quotidiana (es. Incapacità di lavorare).

Questo tipo di inferenze è talvolta supportato da evidenze in studi sperimentali. Che mostrano *validità ecologica* e che spesso riportano *correlazioni* tra la variabile più di base (es. Attenzione) con quella funzionale (Es. Capacità di lavorare).

Occorre ricordarsi che in tal caso la relazione è correlazionale, potrebbe esserci stata una terza variabile che determinava entrambe (es. Un trauma cranico che creava un deficit anche di attenzione oltre che di memoria di lavoro ed è il deficit di attenzione a creare incapacità di lavorare)

La correlazione

Un'ipotetica correlazione spuria per un caso forense



*Questo esempio leggermente più complesso mostra un'ipotetica correlazione spuria
Tra un deficit di memoria di lavoro e l'incapacità di lavorare, quando la causa è un'altra*

La correlazione

NOTA IMPORTANTE 1

Nota che il fatto che la correlazioni non implichi causalità non vuol dire che ogni cosa
Relazione che mostriamo tramite correlazione *non* è causale.

Potrebbe esserlo (ma è importante ricordare che non è la correlazione a dimostrarlo)

Se c'è relazione causale tra due variabili, ci sarà correlazione tra loro (OK)

Se c'è correlazione tra due variabili, c'è relazione causale tra loro (FALLACIA)

(In termini logici questa è una fallacia, quella dell'*affermazione del conseguente*)
https://en.wikipedia.org/wiki/List_of_fallacies

La correlazione

NOTA IMPORTANTE 2

Il concetto di causa è molto complesso ed è molto importante per il neuropsicologo forense che spesso è tenuto a dimostrare o capire relazioni causali (es. Incidente ha causato il danno cognitivo)

Spesso il desumere una relazione causale è un' *inferenza* a partire da dati correlazionali.



Intervalli di confidenza

Un concetto importante nella statistica (e spesso causa di confusione) è quello di intervalli di confidenza.

Gli **intervalli di confidenza** sono intervalli numerici che definisco un range di confidenza attorno ad un parametro stimato.

Quanto si stima un parametro (che sia media, un valore di correlazione, il coefficiente di una regressione, etc.), il valore numerico è una *stima puntuale* del parametro, cioè un singolo valore numerico.

É però verosimile che la nostra stima non sia perfettamente precisa e gli intervalli di confidenza sono (come dice il nome stesso) intervalli, spesso simmetrici intorno al parametro, che ci danno un'indicazione di quanta è l'incertezza nella nostra stima. Nel farlo sono associati ad una probabilità (spesso 95%, o 99%).

Intervalli di confidenza

Un concetto importante nella statistica (e spesso causa di confusione) è quello di intervalli di confidenza.

Gli **intervalli di confidenza** sono intervalli numerici che definisco un range di confidenza attorno ad un parametro stimato.

Quanto si stima un parametro (che sia media, un valore di correlazione, il coefficiente di una regressione, etc.), il valore numerico ottenuto è una *stima puntuale* del parametro, cioè un singolo numero.

É però verosimile che la nostra *stima campionaria* non sia perfetto nello stimare il *parametro della popolazione* (vedi slide precedenti su campione vs popolazione).

Gli intervalli di confidenza sono (come dice il nome stesso) intervalli, spesso simmetrici intorno al parametro, che ci danno un'indicazione di quanta è l'incertezza nella nostra stima. Nel farlo sono associati ad una probabilità (spesso 95%, o 99%).

Intervalli di confidenza

Esempio

$$r_1 = 0.4 \text{ [CI 95\% = 0.25 - 0.65]}$$

$$r_2 = 0.4 \text{ [CI 95\% = 0.38 - 0.42]}$$

r_1 ed r_2 sono due valori di correlazione identici, ma con intervalli di confidenza diversi. L'intervallo ci dà indicazione di quanto siamo precisi nella nostra stima.

Intervalli di confidenza

L'errore classico nell'interpretare gli intervalli di confidenza

Anche se sembrano molto intuitivi, gli intervalli di confidenza andrebbero interpretati in maniera controintuitiva e l'interpretazione più naturale non è corretta. Vediamo un esempio

$$r_2 = 0.4 \text{ [CI 95\% = 0.38 - 0.42]}$$

INTERPRETAZIONE SBAGLIATA: C'è il 95% di probabilità che il parametro a livello di popolazione sia all'interno dell'intervallo 0.38 - 0.42

INTERPRETAZIONE CORRETTA: Costruendo intervalli come quello riportato, nel 95% dei casi il mio intervallo conterrà il parametro della popolazione.

Intervalli di confidenza

Perché è sbagliato interpretare gli intervalli di confidenza nella maniera più intuitiva?

Esaminiamo la frase (sbagliata)

C'è il 95% di probabilità che il parametro a livello di popolazione sia all'interno dell'intervallo 0.38 – 0.42

Essa fa una certa confusione tra cosa effettivamente può variare. Quello che può variare non è certamente il valore della popolazione (quello, per quanto teorico, è un valore fisso), ma il valore del campione. Quindi la verità è che considerando il singolo intervallo le possibilità sono solo due: o contiene il parametro della popolazione, o non lo contiene. Il 95% si riferisce alla probabilità associata *a infiniti intervalli* costruiti con lo stesso metodo,

Questa visualizzazione online spiega bene il concetto

<https://rpsychologist.com/d3/ci/>