a.a 2015-16

Tecniche di Analisi di Dati I, prof. N. Bruno

Concetti fondamentali - elenco e definizioni

(in ordine di apparizione, raggruppati per macroargomenti)
(documento informale, segnalazioni di errori o imprecisioni sono benvenute)

Elementi di teoria della misura

Misura. Qualsiasi processo sistematico per associare numeri a oggetti in modo che relazioni fra gli oggetti, comunque definite, siano rappresentate da corrispondenti relazioni fra i numeri.

Dati. I numeri definiti da un processo di misura. Le proprietà (o se volete il significato) dei numeri stessi dipende dalla natura del processo di misura

Variabile. La proprietà che caratterizza un insieme di dati, ossia ciò che quell'insieme di dati ha in comune.

Variabili dipendenti e indipendenti. Per variabili indipendenti (x) si intendono i predittori di un fenomeno di interesse, mentre per variabili dipendenti (y) si intendono le variabili che dipendono dai predittori, ossia quelle che ci interessa predire o spiegare. Tipicamente dunque si studia la variazione della VD in funzione della VI, ossia $y \sim x$.

Formato rettangolare standard. Formato per la registrazione dei dati in cui ogni riga corrisponde a un caso e ogni colonna corrisponde ad una variabile.

Variabili within e between. I ivelli di una variabile possono variare fra (between) o all'interno (within) delle unità oggetto di misura. In psicologia, tipicamente le unità oggetti di misura sono persone (partecipanti). Una variabile between ha livelli in cui vengono misurati gruppi diversi di partecipanti: ad esempio, il gruppo A comprende i partecipanti 1, 2 e 3 mentre il gruppo B comprende i partecipanti 4, 5 e 6. Le misure del gruppo A in questo senso sono independenti da quelle del gruppo B. La differenza fra i due gruppi dipende dall'effetto della variabile ma

anche dalle differenze individuali fra 1, 2 e 3 da un lato e 4, 5 e 6 dall'altro. Nel formato rettangolare standard dunque ogni caso coincide con uno dei partecipanti. Una variabile within ha livelli costituiti da misure ripetute sullo stesso partecipante. Ad esempio, ogni partecipante viene testato "prima" e "dopo" la presentazione di uno stimolo. In questo caso, la variabile " test" ha due livelli, "prima" e "dopo" e le misure corrispondenti sono associate anziché indipendenti. Nel formato rettangolare standard, ogni caso corrisponde a una misura, e sarà necessario includere una variabile che identifichi i partecipanti.

Fonti di errore nella misura. Ogni misura è affetta da due tipi di errore. Si dice errore variabile o casuale (variable error, random error) l'errore il cui valore dipende da un processo random, definito da un modello statistico. Il modello più comune è il modello gaussiano, secondo il quale la distribuzione dell'errore variabile è simmetrica con forma a campana. Se vale tale modello, l'errore variabile è "egualitario"; errori in eccesso o in difetto del valore reale da misurare hanno uguale probabilità di verificarsi. In un numero sufficientemente grande di misure, pertanto, gli errori di sovra- e sottostima tendono a controbilanciarsi. Si dice invece errore costante o sistematico (constant error, bias) l'errore che si verifica sempre nella stessa maniera ad ogni misurazione. L'errore costante non si controbilancia ma diventa invece sempre più evidente al crescere del numero di misurazioni.

Precisione e accuratezza. Ogni misura è caratterizzata da un certo grado di accuratezza e di precisione. Si dice precisa una misura che ha un piccolo errore variabile. Si dice accurata una misura che ha piccolo errore costante. La precisione è quindi il reciproco o inverso dell'errore variabile, mentre l'accuratezza è il reciproco o inverso dell'errore costante.

Scale o livelli di misura. Proprietà che definisce quali relazioni fra gli oggetti misurati sono preservate dai numeri che li misurano. La distinzione classica è quella quadripartita di S.S. Stevens (scale nominali, ordinali, a intervalli e a rapporto). John Tukey ha proposto una classificazione che motli oggi considerano più utile (nomi, gradazioni, ranghi, frazioni counted, conteggi, quantità, bilanci).

<u>Statistica descrittiva univariata</u>

Classe di una variabile. Un sottoinsieme dei dati in una variabile. Nel caso di una

variabile numerica, una classe è definita da un limite inferiore e da un limite superiore e rappresenta un intervallo (intervallo di classe). Nel caso di una variabile categoriale, una classe è definita dall'etichetta che identifica una delle categorie. Nelle variabili categoriali, la classe viene più spesso chiamata "livello" della variabile.

Distribuzione univariata. La distribuzione di una variabile (univariata) è una rappresentazione in cui le classi della variabile sono associate alle corrispondenti frequenze relative. La distribuzione rappresenta quindi quanti casi sono presenti, in proporzione al totale dei casi, in ogni intervallo di classe. La distribuzione può essere presentata sotto forma di tabella, oppure in forma grafica. I grafici usati più comunemente per presentare una distribuzione univariata sono l'istogramma e il box-plot.

Istogramma. Rappresentazione grafica della distribuzione di una variabile in cui la frequenza relativa associata ad ogni intervallo di classe (variabile numerica) o ad ogni classe categoriale (variabile categoriale) viene rappresentata per mezzo dell'area di un rettangolo. L'altezza di ogni rettangolo rappresenta la densità della classe. L'area totale dell'istogramma rappresenta il 100% dei casi (frequenza relativa = 1). Quando le classi sono categorie, la base del rettangolo non ha significato, quindi la densità della classe e la sua frequenza relativa coincidono. Dal punto di vista grafico, quindi, la larghezza del rettangolo in grafico è arbitraria e altrettanto arbitraria è la posizione di ogni classe sull'asse dell'istogramma.

Quantili. Il quantile q di ordine p $[0 < p < 1]$ è il valore di una distribuzione che ha la proprietà di dividerla in due parti le cui frequenze relative sono p e $1 - p$.

Quartili e percentili. Alcuni quantili corrispondono a posizioni di interesse particolare e per questo motivo prendono dei nomi specifici. Fra questi, i più usati sono i quartili (attenzione, quaRtili è diverso da quaNtili) e i percentili. I quartili sono Q1, il quantile di ordine 1/4 o primo quartile, che divide la distribuzione secondo le proporzioni 0.25 e 0.75; Q2, il quartile di ordine 2/4 o secondo quartile (05. e 0.5); e Q3, il quantile di ordine 3/4 o terzo quartile (0.75, 0.25). Analogamente un percentile di ordine c o Pc divide la distribuzione secondo le proporzioni c e $100 - c$. In maniera analoga possono essere definiti i decili, i ventili, e così via (meno spesso utilizzati).

Five-number summary. Il riassunto della distribuzione con cinque indicatori di posizione: MIN, Q1, Q2, Q3, MAX.

Box-plot (o box-and-whiskers plot). Rappresentazione grafica della distribuzione di una variabile basata sul five-number summary. Viene tracciata una scatola (box) i cui limiti inferiori e superiori sono allineati con Q1 e Q3. All'interno della scatola viene tracciato un segmento all'altezza di Q2. A partire dal box vengono tracciate due linee verso l'alto e verso il basso fino a raggiungere rispettivamente l'altezza del MIN e del MAX della distribuzione (whiskers). Eventuali valori estremi (outliers) vengono identificati mediante criteri opportuni e plottati esternamente ai limiti dei whiskers.

Tendenza centrale. La tendenza dei dati quantitativi a raggrupparsi attorno a un valore tipico. Controintuitivamente, questo valore non rappresenta necessariamente il centro geometrico o di massa della distribuzione, ma dipende invece dalla definizione di "tipico" che viene adottata.

Misure di tendenza centrale. Le più comuni sono la moda (il caso più frequente), la mediana (il secondo quartile) e la media aritmetica (il centro di massa della distribuzione, calcolato come somma dei casi diviso il numero di casi). Utili sono anche la media quadratica (radice quadrata della media aritmetica dei quadrati dei casi), la media geometrica (la radice n-esima della produttoria di n dati, ricavabile anche come antilogaritmo della media aritmetica dei logaritmi dei dati) e la media armonica (reciproco della media aritmetica dei reciproci dei dati). Ogni misura corrisponde a una nozione diversa di tipicità.

Scarti. Dato un vettore X numerico composto da n elementi, gli scarti di X sono le n differenze fra ogni elemento x di X e la media aritmetica di X. Gli scarti misurano la distanza segnata di ogni X dalla media di X, nella direzione negativa ($x < Mx$) o positiva ($x > Mx$).

Dispersione. La tendenza degli scarti a raggrupparsi attorno a un valore tipico. La tendenza centrale della distribuzione degli scarti.

Misure di dispersione. Le più comuni sono la gamma interquartile (la differenza fra Q3 e Q1) e la deviazione standard (la media quadratica degli scarti). In alcuni casi può essere utilizzare la deviazione mediana assoluta (la mediana degli scarti in

valore assoluto).

Skewness. Misura di asimmetria di una distribuzione. Esistono diverse maniere di misurare la skewness, ma tutte seguono la convenzione per cui un valore nullo indica simmetria perfetta, un segno negativo distribuzione con coda a sinistra, e un segno positivo una distribuzione con coda a destra.

Kurtosi. Misura della differenza fra la forma di una distribuzione e la curva gaussiana ("normale"). Una kurtosi pari a 3 indica perfetta coincidenza con la gaussiana (distribuzione mesokurtica); minore di 3, eccesso relativo di osservazioni nella nelle zone intermedia a destra e a sinistra del centro, scarsità relativa al centro e nelle code estreme (platikurtica); maggiore di 3, scarsità relativa di osservazioni a destra e a sinistra del centro, eccesso relativo nella zona centrale (leptokurtica). In alcuni casi la kurtosi viene riespressa sottraendo la costante 3, in modo che la distribuzione mesokurtica corrisponda ad una kurtosi pari a 0, platikurtica a una kurtosi negativa, e leptokurtica ad una kurtosi positiva.

Diagramma a barre. Un diagramma a barre è un grafico in cui quantità (ad esempio, la media) associate a classi di una variabile vengono rappresentate per mezzo dell'altezza di una barra (tipicamente una linea, o un rettangolo sottile). Un diagramma a barre non va confuso con un istogramma. Quasi sempre un grafico con due o più box-plot adiacenti è preferibile rispetto a un diagramma a barre.

Diagramma a torta. Rappresentazione grafica della distribuzione di un variabile categoriale, secondo la convenzione che l'angolo giro (360 gradi ossia l'area totale della torta) rappresenta il 100% dei casi, e il rapporto fra l'angolo o fetta della torta associato ad ogni classe e l'angolo giro rappresenta la frequenza relativa della classe stessa. Pur utilizzando una convenzione simile a quella degli istogrammi, i diagrammi a torta sono quasi sempre una maniera inadeguata di rappresentare una distribuzione, soprattutto se sono presenti molte categorie diverse. Inoltre in ogni caso vanno evitati i diagrammi a torta a tre dimensioni presentati in prospettiva.

<u>Cenni sulle trasformazioni di dati</u>

Trasformazione. Qualsiasi riespressione del dato originario volta a renderlo maggiormente informativo del fenomeno oggetto di studio. Le trasformazioni

tipicamente utilizzate nell'analisi dei dati sono trasformazioni monotoniche (o monotone), ossia trasformazioni che non modificano i ranghi dei dati, e sono invertibili (la funzione applicata per la trasformazione ha una funzione inversa).

Standardizzazione (o trasformazione z). Consiste in una traslazione e in un cambio di scala dei dati. La traslazione ha l'effetto di centrare la distribuzione sullo zero invece che sulla media. Il cambio di scala, di esprimere i dati in unità di deviazione standard. I punteggi standardizzati o punteggi z si calcolano dunque sottraendo da ogni dato la media e dividendo per la deviziazione standard. Il segno del punteggio z ci dice se il dato originario era sotto o sopra la sua media. La grandezza del punteggio z ci dice di quante deviazioni standard.

Trasformazione logaritmica. Il logaritimo in base *a* di *x* è l'esponente a cui va elevata la base per ottenere *x.* Ad esempio, se *a* = 10 e x = 100, log10(100) = 2 perché 10^2 = 100. Con qualsiasi base, log(1) = 0 e se 0 < *x* < 1 log(x) è negativo. Il logartimo di zero non è definito perché non esiste alcun *x* tale che *a^x* = 0. Allo stesso modo non sono definiti i logaritmi di numeri reali negativi. La trasformazione logaritmica è un cambio di scala, in cui a uguali intervalli nella scala lineare di x corrispondono uguali rapporti nella scala logaritmica di log(x). L'effetto della trasformazione logaritmica è espandere la scala per i valori più piccoli di x e contrarla per quelli più grandi.

<u>Statistica descrittiva bivariata</u>

Distribuzione bivariata. La distribuzione congiunta di due variabili x e y, ossia le frequenze relative associate ad ogni combinazione degli intervalli di classe di x e y. Se x e y sono variabili categoriali, la maniera più frequente di visualizzare la distribuzione bivariata è per mezzo della tabella di contingenza. Se x e y sono variabili numeriche, la maniera più frequente è per mezzo di un diagramma di dispersione.

Associazione. Assenza di indipendenza. Due eventi A e B sono indipendenti se p(A & B) = p(A) x p(B), ossia se p(A|B) = p(A) e p(B|A) = p(B).

V di Cramer. Misura di associazione fra variabili categoriali. V misura l'associazione con un numero fra 0 (assenza di associazione) e 1 (associazione perfetta).

Coefficiente di correlazione lineare. Misura di associazione fra variabili numeriche nel caso in cui vale l'assunzione di normalità bivariata (il diagramma di dispersione ha più o meno la forma di un pallone da rugby). Di solito indicata con r. Il coefficiente r misura la forza dell'associazione con un numero fra 0 (assenza di associazione) e 1 (associazione perfetta), e la direzione dell'associazione con il segno (negativo, y decresce al crescere di x; positivo, y cresce al crescere di x).

Tabella di contingenza. Rappresentazione della distribuzione bivariata di x e y quando questi sono categorie. Le frequenze assolute di ogni combinazione fra i livelli di x e di y vengono scritte in ogni cella della tabella. I totali per riga e per colonna rappresentano le distribuzioni marginali di x e y, la somma dei totali per riga = somma dei totali per colonna = numero totale di osservazioni.

Diagramma di dispersione. Rappresentazione grafica di una distribuzione bivariata in cui ogni coppia di osservazioni x, y viene rappresentata da un punto su un piano. Anche se segue la stessa convenzione di un diagramma cartesiano, un diagramma di dispersione è concettualmente diverso da un diagramma cartesiano perché le unità di misura di x e y non sono necessariamente le stesse e il punto in cui, nel grafico, gli assi si incontrano non è necessariamente il punto che ha coordinate 0, 0 (l'origine degli assi nel senso cartesiano).

Box-plot multipli. Una distribuzione bivariata può essere rappresentata per mezzo di box-plot multipli di y in funzione dei livelli di x. Questo tipo di rappresentazione grafica è particolarmente utile per visualizzare la relazione fra y numerico e x categoriale.

Regressione lineare. L'adattamento (fit) di un modello lineare alla distribuzione bivariata di y in funzione di x. Consiste nello stimare la retta che minimizza la somma dei quadrati degli scarti di x dal valore stimato corrispondente. La retta di regressione rappresenta la tendenza in media di y a variare in funzione di x. In quanto tale rappresenta il modello adatto per prevedere nuovi valori di y, noto x, se la covariazione di y e x è lineare.

Intercetta. Uno dei parametri del modello lineare. L'intercetta (o costante) è il valore previsto per y quando x = 0.

Pendenza. Uno dei parametri del modello lineare. La pendenza (o coefficiente angolare) misura di quanto cresce y quando x cresce di 1, ossia il tasso di crescita di y in funzione di x.

Residui. I residui della regressione sono gli scarti fra ogni y e il corrispondente valore previsto in base al metodo della regressione, ossia i valori che stanno sulla retta.

Errore standard dei residui. L'errore standard residuale (residual standard error) o RMSE (root mean square error della regressione) è la media quadratica dei residui. Rappresenta l'errore tipico di previsione quando si utilizza il modello lineare per prevedere y in funzione di x.

Cenni di teoria della probabilità

Probabilità. Un numero reale compreso fra 0 e 1, dove zero corrisponde all'evento impossibile e 1 all'evento certo. La definizione di probabilità dipende dall'approccio teorico adottato. Secondo la definizione soggettivista, la probabilità è il grado di certezza soggettiva che un individuo ha in merito alla possibilità che un evento si verifichi. Secondo la definizione frequentista, invece, la probabilità è il limite del rapporto fra casi favorevoli e casi possiibili. Nella definizione frequentista, la probabilità è oggettiva e misurabile ma solo per eventi ripetibili che si possono contare. Quindi l'affermazione "la probabilità che uno studente di psicologia sia donna è 0.65" è sensata perché posso contare gli studenti e contare quanti di questi sono donne, almeno in linea di principio. L'affermazione "la probabilità che l'ipotesi H sia vera è 0.95" invece non ha senso, perché l'ipotesi non è qualcosa di ripetivile ma una affermazione che o è vera o non lo è. Nella definizione soggettivista, la probabilità è una credenza o se volete uno stato mentale. In questo caso quindi è sensato parlare anche di probabiità di eventi singoli. Le regole di calcolo sono comunque le stesse in entrambi gli approcci.

Distribuzioni teoriche di probabilità. Il modello matematico di un processo stocastico (governato dal caso). Una DTP descrive i valori che la variabile generata dal processo può assumere, e li associa a densità di probabilità mediante una funzione. Una distribuzione teorica di probabilità è quindi una equazione che esprime la densità di probabilità in funzione dei valori di una variabile. La variabile può essere continua o discreta. Nel caso di una variabile continua, la distribuzione

si rappresenta con una curva, l'altezza in y della curva ad ogni valore di x rappresenta la densità di x in quel punto, e l'area sotto la curva sottesa da un intervallo in x rappresenta la probabilità di quell'intervallo. Nel caso di un variable discreta, la distribuzione è più propriamente rappresentata da una serie di barre verticali, la cui altezza rappresenta la densità di probabilità discreta associata ad ogni valore di x. A lezione sono state analizzate in dettaglio le proprietà di quattro distribuzioni teoriche: la distribuzione uniforme (continua o discreta), la distribuzione normale (continua), la distribuzione binomiale (discreta), la distribuzione t di Student e la distribuzione chi-quadrato (entrambe continue).

Distribuzioni empiriche di probabilità. La distribuzione delle probabilità associate ai valori che può assumere una variabile, stimata per via empirica utilizzando le frequenze relative osservate.

Calcolo combinatorio. Il settore della matematica che studia i modi per raggruppare o ordinare secondo regole date gli elementi di insiemi finiti. Il calcolo combinatorio serve in particolare per contare quanti sono questi ordinamenti. A lezione abbiamo studiato tre regole: quelle per le permutazioni, le disposizioni, e le combinazioni. Queste ultime sono utili per il calcolo della funzione di densità discreta della distribuzione binomiale.

<u>Stima</u>

Campione. Qualsiasi insieme di dati estratti da una popolazione non nota.

Campione casuale semplice. Un campione in cui la probabiità per ogni caso potenziale di entrare nel campione è costante ed è nota in quanto dipende solo da un processo stocastico (governato dal caso).

Campione di convenienza. Un campione in cui i fattori che determinano la probabilità per ogni caso potenziale di entrare nel campione non sono noti.

Statistiche. Grandezze che esprimono caratteristiche strutturali di una distribuzione empirica, come la sua tendenza centrale, la sua dispersione, la sua forma, o, nel caso di una distribuzione bivariata, il grado di associazione. Si noti che le statistiche sono variabili casuali, in quanto sono misurate in un campione di dati soggetto per definizione all'errore variabile. Ciò significa che ogni nuovo

campione, anche estratto dalla stessa popolazione, avrà statistiche diverse dal precedente.

Parametri. Grandezze che esprimono caratteristiche strutturali di una distribuzione teorica o di una popolazione e che sono utilizzate nella formulazione di modelli matematici di tali distrbuzioni. Ad esempio, nell'equazione della distribuzione normale troviamo i parametri media e deviazione standard. Si noti che i parametri non sono variabili casuali ma proprietà costanti di uno stato di cose del mondo (un processo che genera dati, una popolazione reale o potenziale di misure, e così via).

Stimatori. Statistiche che possono essere utilizzate per stimare parametri. Due importanti distinzioni sono quelle fra stimatori privi di errore stistematico e stimatori affetti da errore sistematico. Ad esempio, la statistica M (la media di un campione) è uno stimatore privo di errore sistematico del parametro μ (la media di una popolazione); mentre la DS (la deviazione standard di un campione) è uno stimatore affetto da errore sistematico del parametro σ (la deviazione standard di una popolazione).

Distribuzioni campionarie. Traduzione infelice dell'inglese *sampling distribution.* Sono distribuzioni teoriche di statistiche campionarie, ossia distribuzioni che risultano dall'estrazione di ripetuti campioni, cioè dall'effettuare ripetutamente una operazione di *sampling.* Ad esempio, estraggo ripetuti campioni di grandezza *n* e per ogni campione calcolo la media aritmetica. L'insieme delle medie (teoriche) che ottengo ha essa stessa una distribuzione, che può essere approssimata usando leggi probabilistiche o stimata grazie ad altri metodi.

Teorema del limite centrale. Data una popolazione qualsiasi con media aritmetica μ e deviazione standard σ, la distribuzione campionaria teorica delle medie aritmetiche di campioni di numerosità *n* tende a Norm(μ, $\sigma/\sqrt{n}$) per $n \to \infty$. Ossia, al crescere della numerosità campionaria, la distribuzione teorica della media assomiglia sempre di più alla distribuzione normale, con tendenza centrale uguale alla media della popolazione da cui estraggo i campioni ma disperisione ridotta in proporzione alla radice quadrata della numerosità.

Legge dei grandi numeri. Quando estraggo un campione da una popolazione qualsiasi con media aritmetica μ, la media del campione M -> μ per -> ∞. Ossia, al

crescere della sua numerosità, un campione assomiglia sempre più alla popolazione da cui è estratto.

Intervallo di fiducia. L'intervallo all'interno del quale, con un certo livello di fiducia (ossia una certa probabilità), posso ritenere ricada un parametro il cui valore desidero stimare, a partire da una statistica. La banda di incertezza della stima.

Introduzione all'inferenza statistica

Bontà dell'adattamento. Una misura di quanto la distribuzione dei dati è in accordo con una distribuzione teorica, ad esempio, quella prevista da una teoria. Una misura molto utilizzata è basata sul chi-quadrato. In un modello lineare, una misura di bontà dell'adattamento è data dal coefficiente di correlazione elevato al quadrato.

Effetto. L'effetto (lineare) di x (variabile indipendente) su y (variabile dipendente) è la pendenza della retta di regressione $y \sim x$. Quando x è una categoria a due livelli, l'effetto coincide con la differenza fra le medie.

Test statistico dell'ipotesi nulla. Procedura di analisi in cui il dato osservato (ad esempio, la differenza fra le medie di due gruppi) viene confrontato con la previsione di una *ipotesi nulla,* ossia con un modello statistico che assume che il dato dipenda dal solo errore casuale (ad esempio, che la differenza osservata fra le due medie sia dovuta al solo errore di campionamento). Il modello specifica quale dovrebbe essere la distribuzione campionaria della statistica osservata, se è vera l'ipotesi nulla. Da tale distribuzione è possibile calcolare la probabilità di ottenere il dato osservato, o più estremo, trovando l'area sottesa dal quantile corrispondente (ad esempio, il quantile che corrisponde alla differenza osservata, nella distribuzione t di Student con il numero appropriato di gradi di libertà), se l'ipotesi nulla è vera. Se questa probabilità è minore di un criterio (di solito, $p < 0.05$), l'ipotesi nulla viene rifiutata. Il test statistico dell'ipotesi nulla è ancora la procedura di analisi maggiormente utilizzata in psicologia e in molte altre discipline, ma è anche una procedura controversa -- talmente controversa che molti metodologi contemporanei ritengono che andrebbe abbandonata. In ogni caso, è importante ricordare che la probabilità calcolata (il cosiddetto p-value osservato) è solo la probabilità di un errore del primo tipo (vedi), non la probabiiltà che l'ipotesi nulla sia vera o che il risultato sia replicabile, né tantomeno un

indicatore della bontà o della forza del risultato ottenuto. Inoltre, non va mai dimenticato che la decisione di rifiutare o non rifiutare l'ipotesi nulla è di natura probabilistica e può sempre essere sbagliata. Questo è specialmente vero se l'ipotesi nulla è in realtà falsa, dato che il calcolo si basa sull'assunzione contraria che sia vera. Infine, un limite evidente della procedura è che quando l'esito è il mancato rifiuto del'ipotesi nulla non è possibile trarre alcuna conclusione.

Errore del primo tipo. Nel test statistico dell'ipotesi nulla, la probabilità di rifiutare l'ipotesi nulla quando questa è vera. Quando l'ipotesi nulla è falsa, è impossiible commettere un errore del primo tipo.

Errore del secondo tipo. Nel test statistico dell'ipotesi nulla, la probabilità di non rifiutare l'ipotesi nulla quando questa è falsa. Quando l'ipotesi nulla è vera, è impossiible commettere un errore del secondo tipo.

Potenza del test. Nel test statistico dell'ipotesi nulla, si dice potenza del test la probabilità di rifiutare l'ipotesi nulla quando questa è falsa. La potenza del test è 1 - la probabilità dell'errore del secondo tipo.