Statistica

domenica 11 ottobre 2020
13:26

1. Definire problema;
2. Formulare ipotesi precisa;
3. Impostare disegno sperimentale;
4. Raccogliere e analizzare i dati.

Raccoglie dati (dati grezzi), elaborazione dati (tabelle, grafici…), informazioni sui dati, valutazione critica dati e presentazione dati. La statistica applicata alle scienze biologiche è chiamata **biometrica.** Ha due branchie principali:

- **Descrittiva:** tabelle, rappresentazioni grafica dati, sintesi dati in pochi indici informativi;
- **Inferenziale:** definizione caratteristiche della popolazione attraverso studio di un campione da essa estratto.

Descrittiva

L'insieme è la collezione di oggetti, individui o animali con almeno una caratteristica comune. La variabile è una caratteristica che può assumere diversi valori:

1. Qualitativa: no rappresentate da numero, suddivisibile in categorie (colore mantello) distinte in base a caratteristiche non numeriche;
2. Quantitativa: misura a cui è associato un numero su scala numerica.

Variabili qualitative, Scale di misura (diagrammi a barre):

- **Nominale**: valori rientrano in categorie o classi non ordinate, le caratteristiche hanno dei nomi (maschio/femmina);
- **Ordinale**: ordine importante, assegnazione ordine logico, (colore mantello da più chiaro a scuro);

Variabili quantitative:

a. **Discreta:** assume solo valori isolati;
b. **Continua:** assume tutti i possibili valori.

Le variabili continue possono essere raccolte in classi facendo la **categorizzazione**:

a. Ordinare i valori dal piccolo al grande;
b. Dividere in intervalli, ciascuno con limite inferiore e superiore. Il limite superiore di una classe non deve coincidere con quello inferiore della classe successiva.

Le misurazioni non sono identiche ma variano, dovuto a variazione biologica o errori di misurazione. Quando si osservano i dati è importante controllare il range e vedere se c'è qualche dato eccessivamente al di fuori dai limiti. L'accuratezza il grado di accordo tra valore osservato e valore vero (valore ottenuto - valore vero). L'inaccuratezza è data da rapporto tra accuratezza e valore vero. La precisione rappresenta la somiglianza delle misurazioni che faccio ripetutamente per un parametro:

- Ripetibilità: misure ottenute con stesso metodo nello stesso laboratorio;
- Riproducibilità: misure ottenute con stesso metodo in laboratori diversi.

La precisione è data dal **coefficiente di variabilità**, dato da deviazione standard su valore vero.

POPOLAZIONE: insieme di elementi su cui si fa l'analisi. Un sui sottoinsieme è un CAMPIONE:

- Scelto con accuratezza (evitare bias cioè distorsione dello stimatore), scelta casuale per ridurre l'errore sistematico;
- Utilizzare un adeguato numero di unità campionarie per ridurre errore causale;
- Elaborare in modo corretto gli stimatori.

Variabili categoriche: • Tabelle; • Diagrammi a barre; • Torte.	**Variabili quantitative:** • Tabelle; • Istogramma; • Box plot
Associazione tra variabili categoriche: • Tabelle di contingenza; • Diagramma a barre raggruppate	**Associazione variabili quantitative:** • Diagramma a dispersione; • Diagramma a linee.

I dati devono essere riordinati, classificati e sintetizzati mediante tabelle e grafici di FREQUENZA. Il numero di osservazioni di un determinato valore di misura è chiamato FREQUENZA ASSOLUTA. La classe è la suddivisione delle osservazioni in intervalli di determinata ampiezza. La DISTRIBUZIONE DI FREQUENZA è l'ordinamento dei dati secondo le classi e secondo le corrispondenti frequenza. La FREQUENZA RELATIVA di una classe è la frequenza della classe divisa per il totale delle frequenza di tutte le classi. La classe modale è quella che compare più volte.

La FREQUENZA CUMULATIVA è la somma delle frequenza di un intervallo a quelle dell'intervallo precedente.

La serie di dati può essere divisa in 4 zone equivalente (**QUARTILI**), i valori sono:

- **1° quartile o 25° percentile**: valore al di sotto del quale c'è il 25% delle osservazioni;
- **2° quartile o 50° percentile**: valore al di sotto del quale c'è il 50% delle osservazioni (MEDIANA);
- **3° quartile o 75° percentile**: valore al di sotto del quale c'è il 75% delle osservazioni;

I percentili dividono i dati ordinati in 100 parti uguali. Danno un'indicazione della distribuzione. Con due curve si possono andare a confrontare le varie classi.
Per calcolare i quartili di una distribuzione:

- Ordino gli n dati in ordine crescente;
- P=0,25 per il 25° percentile e calcolo **k=np;**
- Se k è intero, il quartile si ottiene facendo la media del k-esimo e del (k+1)-esimo valore dei dati;

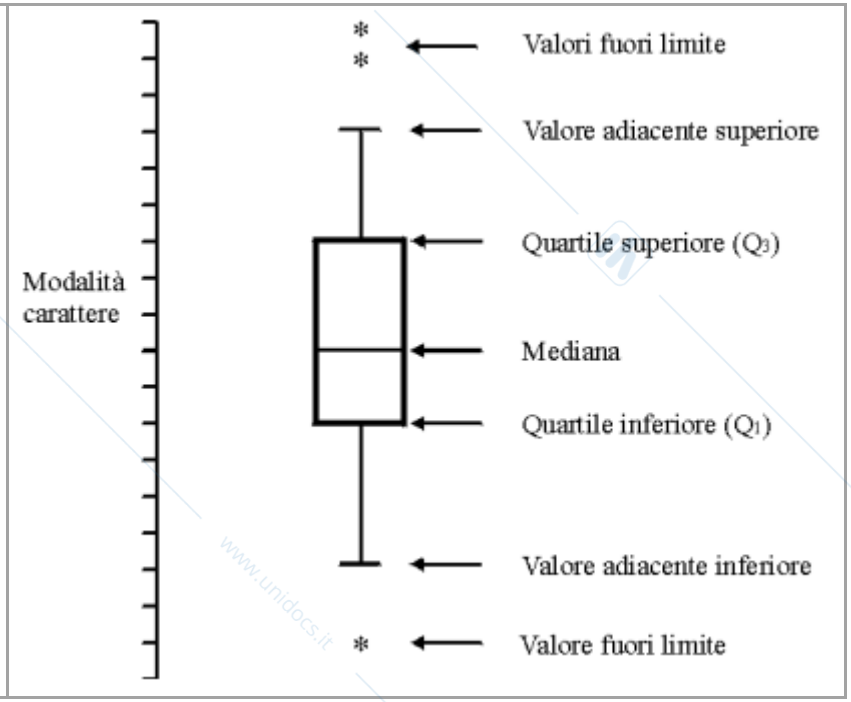
- Se k non è intero si arrotonda k per eccesso al primo intero successivo e si sceglie come quartile il corrispondente valore dei dati.

Box plot

Fornisce una descrizione sintetica della distribuzione dei valori. Riporta:

Mediana;
1° e 3° quartile estremi del rettangolo;
Valori adiacenti superiore e inferiore sono i baffi
Valore minimo e massimo.
Differenza interquartilica = terzo quartile - primo quartile = r
I valori anomali, fuori dal limite sono:
> Q3 + (1.5*r)
< Q1 - (1.5*r)

Si evidenziano anche i valori fuori dal limite. Se sono troppo estremi si vanno a eliminare per non andare ad risentire su altri valori come la media. Dall'estensione dei baffi si ricava la dispersione dei dati. Nel box ci sono i quartili, contenenti il 50% dei dati. Più è lungo il rettangolo e più la distribuzione è dispersa. La mediana non è per forza al centro ma spostata verso un quartile, la distribuzione è asimmetrica. Si usano per confrontare più distribuzioni. Se la distribuzione è simmetrica mediana e media coincidono.



- La distribuzione dei dati può essere rappresentata da una curva che può essere simmetrica, asimmetrica a sinistra (negativa) o asimmetrica a destra (positiva). Si possono avere classi con frequenze elevate (classi modali), se una è elevata la distribuzione è unimodale o curve con più classi modali e si dicono bimodali o trimodali.

- Per associare due variabili continue ci utilizza il grafico a dispersione, ci dice come variano le due variabili. Se tutti i punti sono dispersi non c'è relazioni tra le variabili.

- I diagrammi a linee si utilizzano per variabili continue con una variabile temporale. Il grafico combinato contiene più variabili che variano nel tempo.

- Possono essere associate due variabili categoriche, rappresentate con tabelle di contingenza

I grafici devono avere un titolo, indicazione delle variabili e l'unità di misura. Per le barre si deve riportare cosa indicano. Per le tabelle ci deve essere titolo, intestazione colonne e righe con unità di misura.

Indici di disposizione e dispersione

domenica 11 ottobre 2020
20:46

Come si dispongono i dati nella disposizione.

$\sum_{i=1}^{n} Xi$ sommatoria per i che va da 1 a n, n è il numero di soggetti.

Media

Somma tutte osservazione diviso il numero di osservazioni $\mu = \overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$. Se i dati rappresentano l'intera popolazione si usa la lettera greca, se sono di un campione di n osservazioni di usa $\bar{x}$. lLunità di misura è la stessa della variabile. Risente molto i valori anomali.

Esistono due tipi di medie:

- Ponderata: alcuni dati contribuiscono più degli altri $\bar{x} = \frac{\sum(W*x)}{\sum x} = \frac{\text{sommatoria ciscun valore *il sui peso}}{\text{somma dei pesi}}$;
- Geometrica: per distribuzione dati asimmetrica, trasformazione logaritmica dei dati per farla diventare più simmetrica. È l'antilogaritmo della media dei logaritmi dei singoli valori.

Mediana

Divide la distribuzione in due parti equivalenti. Se il numero di osservazioni sono dispari, si ordinano e si trova quello he sta in centro $i = \frac{n+1}{2}$.

Se sono pari la mediana è data dalla media dei valori centrali, trovo due numeri $\frac{n}{2}$ e $\frac{n}{2} + 1$ e faccio la media. Stessa unità di misura della variabile. Non si adatta a manipolazioni matematiche.

Moda

Valore con maggiore frequenza. C'è ne possono essere più di una.

Indici di dispersione

Gli **indici di dispersione** indicano come questi valori si disperdono sull'asse.

Range (intervallo di variazione)

Differenza tra valore massimo e valore minimo.

Varianza (MS)

Misura dispersione dati intorno alla media. Per ogni dato ci interessa sapere lo scostamento dalla media. Se non è elevato i dati sono raggruppati vicino alla media. La somma degli scostamenti dalla media è sempre 0 quindi si elevano al quadrato.

Popolazione	**Campione**
$\sigma^2_x = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$ μ indica la media $\sigma^2_x = \frac{\sum_{i=1}^{n} x^2_i - \frac{(\sum x_i)^2}{n}}{n}$	$S^2_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ Il denominatore è chiamato GRADO DI LIBERTA' = numero osservazioni - 1. Se ci fosse n la varianza della popolazione e del campione sarebbero uguali. Differenza fra numero osservazioni e numero parametri che dobbiamo stimare (in questo caso solo media quindi -1) $S^2_x = \frac{\sum_{i=1}^{n} x^2_i - \frac{(\sum x_i)^2}{n}}{n-1}$

Il numeratore della varianza si chiama DEVIANZA (SS)

Deviazione standard

Radice quadrata della varianza. Indica lo scostamento medio dalla media. Più è grande e più la distribuzione è dispersa. (si usa quando voglio far vedere la variabilità, per far vedere la precisione di usa l'errore standard).

$\sigma_x = \sqrt{\sigma^2_x}$ $S_x = \sqrt{S^2_x}$

Coefficiente di variabilità (C.V.)

Misura adimensionale, indica la variabilità di un carattere.

$$\frac{\text{deviazione standard}}{\text{media}}$$

Per trovare il numero delle classi si usa la regola di Sturges: $k = 1 + \log_2(n)$
Per trovare l'intervallo di variazione delle classi si fa range diviso numero classi.

Le variabili continue in genere hanno una distribuzione simmetrica, moda mediana e media coincidono. In quelle asimmetriche le tre misure non coincidono, c'è un prolungamento di una delle due code di dati (a sinistra o a destra). A sinistra è negativa a destra positivo. La punta massima della curva si chiama CURTOSI, se i dati sono concentrati in un valore alto e l'indice è positivo (leptocurtica), se è negativa i dati non sono concentrati e la distribuzione è relativamente piatta (platicurtica).

Probabilità

lunedì 12 ottobre 2020
14:39

La prova è un esperimento che si fa che ha 2 o più risultati che non possono essere predetti. L'evento è il risultato dell'esperimento, lo spazio campionario è un esperimento composto da tutti gli eventi.
La probabilità di un dato evento è il rapporto tra casi favorevoli e casi possibili. La probabilità che non si manifesti l'evento è 1-probabilità evento che si manifesta. Quello che si manifesta è p quello che non si manifesta è q. probabilità è numero compreso tra 0 e 1,come percentuale o come rapporto.

Due eventi sono indipendenti quando il verificarsi di un evento non ha influenza sull'altro.
La probabilità che si verificano entrambi è dato dal prodotto delle probabilità dei due eventi (PRINCIPIO DEL PRODOTTO).

Due eventi mutuamente esclusivi non possono verificarsi contemporaneamente. La probabilità che si presenti l'uno o l'altro è data dalla somma delle probabilità (PRINCIPIO DELLA SOMMA).
Qual è la probabilità di estrarre un asso (evento A) o un re (evento B) da un mazzo di carte in una sola estrazione?
Pr{A} = 4/52 (4 assi in un mazzo di 52 carte) = 1/13
Pr{B} = 4/52 (4 re in un mazzo di 52 carte) = 1/13
Pr{A + B} = 1/13 + 1/13 = 2/13 = 0.15

Due eventi non sono mutuamente esclusivi se il presentarsi di uno non esclude il presentarsi dell'altro. La probabilità del verificarsi dell'uno o dell'altro o entrambi è uguale alla somma delle probabilità dei due eventi meno la probabilità che si verifichino contemporaneamente.
Qual è la probabilità di estrarre o un asso o una carta di cuori da un mazzo di carte?
Gli eventi A (estrazione di un asso) e B (estrazione di una carta di cuori) non si escludono a vicenda, dato che può essere estratto l'asso di cuori. Quindi
Pr{A o B} = Pr{A} + Pr{B} - Pr{AB}
Pr{A} = 4/52 (4 assi in un mazzo di 52 carte) = 1/13
Pr{B} = 13/52 (13 carte di cuori in un mazzo di 52 carte) = 1/4

Pr{AB} = 4/52 * 1/4 = 1/52
Pr{A o B} = 4/52 + 1/4 = 0,3269 1/52 = 16/52

La variabile casuale può assumere numerosi valori, dovuti al caso:

- Discreta, assume un numero finiti di risultati;
- Continua, può assumere qualsiasi valore.

Le distribuzioni di probabilità di una variabile discreta specifica tutti i possibili risultati della variabile (sulle x) insieme alla probabilità che si verifichino (asse y). Possono essere di due tipi:

Distribuzioni di probabilità discreta

- Distribuzione binomiale
- Distribuzione di Poisson

Distribuzioni di probabilità continua

- Distribuzione normale
- Distribuzione χ^2
- Distribuzione *t* di Student
- Distribuzione F

• Distribuzione binomiale

Si riferisce ad un insieme di prove ripetute ciascuna con due esiti mutuamente esclusivi (maschio/femmina o positivo/negativo). I risultati sono indipendenti. Se p è la probabilità che si presenti un cero evento in una prova e q=1-p la probabilità che non si presenti, allora la probabilità che l'evento si presenti x volte in n prove è data da:
(probabilità che su tre soggetti ne trovo 1 sano oppure 2 sani….)

$$\frac{n!}{x!*(n-x)!} * p^x * q^{n-x}$$

Se q=p la distribuzione è simmetrica, se n cresce le combinazioni diventano sempre meno probabili.

• Distribuzione di Poisson

Per eventi che si verificano raramente nel tempo e nello spazio. Probabilità piccole con n molto grandi. λ è il numero medio di volte che si verifica l'evento. La probabilità che l'evento si verifichi x volte è pari a:

$$\frac{\lambda^x}{x!}e^{-\lambda} \quad \lambda = n * probabilità$$

Le variabili continue possono assumere un infinito numero di valori, impossibile tracciare linee di separazione, non classi. La distribuzione è descritta da una curva e l'altezza è detta densità di probabilità, l'area sotto la curva è uguale a 1. tramite gli integrali si può calcolare l'area tra due punti.

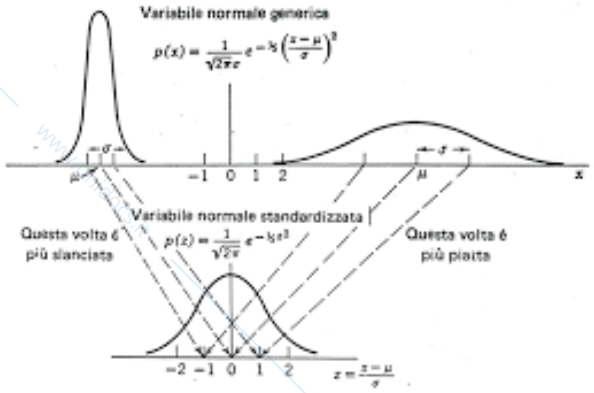
- ## Distribuzione normale (curva di Gauss)

La maggior parte delle variabili sugli animali sono misure quantitative che descrivono una curva chiamata *curva di Gauss, a campana o curva di distribuzione normale*. È una curva simmetrica, l'asse centrale è mediana, moda e media. Selle ascisse ci sono le variabile studio e sulle ordinate la probabilità. Più ci allontaniamo dalla mediana più diventano rari i valori. La funzione che la descrive contiene media e deviazione standard:

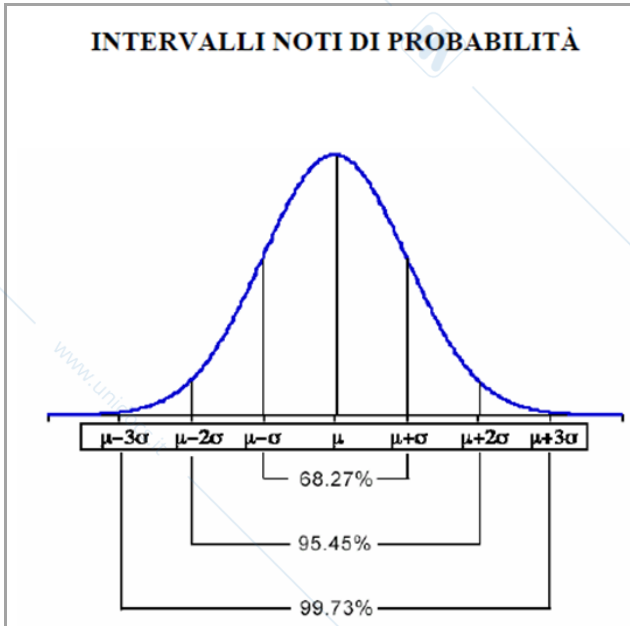
$$F(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Nei valori si indica o deviazione standard (indica variabilità) o errore di misura. La deviazione standard è il segmento che va dalla media al punto di flesso della curva. Considerando più curve, se varia solo la deviazione ma non la media, la curva si appiattisce all'aumentare della deviazione, diventa più appuntita quando diminuisce la deviazione. È una funzione che tende a 0 a + e - infinito.
Tutte le possibili distribuzioni possono essere confrontate tra loro riportandole ad una curva di Gauss standardizzata.
La STANDARDIZZAZIONE è (deviazione dalla media della popolazione)/deviazione standard. Valore ciascuna variabile - media distribuzione diviso deviazione standard:

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



Standardizzare una variabile serve a renderla indipendente dalla grandezza fisica misurata. La curva standardizzata ha media=0 e deviazione standard=1, rappresentata da una sola curva, mentre quella normale ha infine curve. Ci permette di calcolare la probabilità di diversi intervalli, utilizzando tavole specifiche. A sinistra ci sono i valori di Z e nelle prima riga i decimali.



- $\mu \pm 1\sigma$ include il 68% circa delle osservazioni [1 – (2*0.1587) = 0.6826]
- $\mu \pm 2\sigma$ include il 95% circa delle osservazioni [1 – (2*0.0228) = 0.9544]
- $\mu \pm 3\sigma$ include il 99% circa delle osservazioni [1 – (2*0.0013) = 0.9974]
- $z = 1.96$ esclude il 2.5% a destra della media
- $z = -1.96$ esclude il 2.5% a sinistra della media
- $-1.96 \leq z \leq 1.96$ esclude il 5% della distribuzione normale

Data la probabilità si può calcolare X con la formula inversa.

Per identificare la non normalità:

- Trasformazione logaritmica Y'= ln(y);
- Trasformazione arcoseno $p' = \arcsin(\sqrt{p})$
- Trasformazione radice quadrata $Y' = \sqrt{Y} + 1/2$
- Trasformazione quadratica
- Trasformazione reciproca Y'=1/Y

Area tra -1.20 e 2.40: P(Z=-1,20) = 0,1151 P(Z=2,40) = 0,0082 P (-1.20 ≤ Z ≤ 2,40) = 1 - (0,1151 + 0,0082) = 0,8767	Area tra 1,23 e 1,87: P(Z=-1,23) = 0,1093 P(Z=1,87) = 0,0307 P (1,23 ≤ Z ≤ 1,87) = 0,1093 - 0,0307 = 0,0786

Media = 0,37
Dv= 0,066

- 0,40 sopra Z=0,45 area 0,3264
- 0,30 sotto Z=-1,06 area 0,1446
- Tra 0,30 e 0,40 area = 1- 0,471=0,529
- Valori che escludono 5% a sinistra e 5% a destra, si trova Z, -1,645 e 1,645. Si ricava X.

Intervallo di confidenza

lunedì 26 ottobre 2020
13:52

Errore standard

Statistica Inferenziale: definizione caratteristiche della popolazione attraverso studio di un campione da essa estratto.
Il campionamento comporta due tipi di errori riducibili ma ineliminabili

- **Errore sistematico**: variano in modo imprevedibile da una misura all'altra e influenzano il risultato qualche volta per eccesso, qualche altra volta per difetto.
- **Errore casuale**: avvengono sempre nello stesso senso, o sempre per eccesso o sempre per difetto.

Non si può quindi affermare con certezza che i risultati ottenuti su un campione siano trasferibili ad una popolazione. Ma si può generalizzare in termini medi utilizzando opportuni metodi statistici.

POPOLAZIONE: insieme di elementi su cui si fa l'analisi. Un sui sottoinsieme è un CAMPIONE:

- Scelto con accuratezza (evitare bias cioè distorsione dello stimatore), scelta casuale per ridurre l'errore sistematico (errore che si ripete in tutte le misurazioni);
- Utilizzare un adeguato numero di unità campionarie per ridurre errore causale;
- Elaborare in modo corretto gli stimatori.

Per ridurre questi errori si deve scegliere random il campioni e un adeguato numero di unità campionarie.

Se da una popolazione estraiamo vari campioni di uguale numerosità ognuno con una media diversa, si può fare la distribuzione delle medie campionarie:

1. Forma della distribuzione è normale
2. La media è uguale alla media della popolazione
3. La deviazione standard è chiamato **errore standard** = sigma/radice di n $\frac{\sigma}{\sqrt{n}}$ aumentando le osservazioni l'errore standard diminuisce

L'errore standard è la distribuzione standard di una media campionaria $\sigma_{\bar{x}}$ invece per campione $S_{\bar{x}}$.

TEOREMA DEL LIMITE CENTRALE: Se n è sufficientemente grande, la distribuzione delle medie campionarie è approssimativamente normale. Visto che è normale le medie campionarie possono essere standardizzate (media campionaria standardizzata): $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Quando si analizzano singoli valori provenienti da una distribuzione normale occorre utilizzare la variabile standardizzata: $Z = \frac{X-\mu}{\sigma}$

L'intervallo di confidenza della media è l'intervallo entro cui è compresa la media della popolazione (limite massimo e minimo). Se questo intervallo è stretto, la media è abbastanza buono, se è largo la media non è una stima molto buona. Escludiamo il 2,5% a destra e a sinistra perché nelle code ci sono medie campionare che si verificano poche volte, interessa invece la regione centrale dove le medie compaiono più spesso. La somma delle due aree che elimino si chiama alfa.

I limiti di confidenza sono $\bar{x}(\pm)1{,}96 * \frac{\sigma}{\sqrt{n}}$

I limiti di confidenza sono gli estremi di intervallo dove si è confidenti che si trovi il 95% della media della popolazione.

Per calcolare l'**errore standard del campione** si una la deviazione standard del campione (S): $\frac{S}{\sqrt{n}}$

Media ± DS	Media ± ES	
* La deviazione standard misura lo scostamento dalla media	* È una misura della precisione della media del campione come stima della media della popolazione	Se metto la deviazione indico la variabilità, invece se metto errore standard indico la precisione della misura Media (+-) deviazione o media (+ -) errore standard
* Dà un'indicazione di quanto vicine alla media siano le osservazioni	* Dà un'indicazione di quanto vicino alla media della popolazione sia la media del campione	
* È utilizzata per costruire un range in cui si trova la maggior parte delle osservazioni di una popolazione	* È utilizzata per calcolare l'intervallo di confidenza che permette di giudicare la precisione della stima della media della popolazione	

Distribuzione t di student (sigma non noto)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Se **sigma non è noto**, lo sostituisco con S che è la deviazione standard del campione. Il rapporto non si chiama Z ma si chiama t che non segue la distribuzione standardizzata ma segue la distribuzione t di student. Ne esiste una per ogni grado di libertà, in questo caso n-1. Simmetrica, più piatta e le code sono più grosse, l'area della curva è 1. Tutte le tabelle si sintetizzano in un'unica, a sinistra una colonna con i gradi di libertà, sopra l'area da escludere. I valori che trovo sono gli estremi delle aree che vado a eliminare.

I limiti sono $\bar{X}\ (\pm)\ t_{n-1} * S_{\bar{X}}$

dove indica il valore di t che esclude il 5% della distribuzione (2.5% per parte).

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

INTERVALLO DI CONDIFENZA DI UNA PROPORZIONE

La distribuzione di una proporzione è approssimativamente normale. La proporzione del campione (p) è una stima non viziata della proporzione della popolazione (π), l'intervallo di confidenza è:

$p \pm 1{,}96 * \sqrt{\frac{p(1-p)}{n}}$	$p = \frac{parte\ di\ campione}{totale\ campione}$

Un campione di 115 bovini è selezionato in una determinata zona e i campioni di sangue sono testati per la presenza di anticorpi verso la Leptospira. 36 bovini sono risultati positivi.
P = 36/115 = 0.313

Test d'ipotesi

Il test d'ipotesi è un processo mediante il quale si traggono conclusioni su di una popolazione utilizzando un campione tratto da essa. Le conclusioni che otteniamo non siamo assolutamente sicuri che siano giusti. Possiamo esprimere le conclusioni attraverso la **probabilità**.
Si applica un test ai dati, il valore che si ottiene si vede con che probabilità si verifica, se ha alta probabilità si accetta l'ipotesi, se la probabilità è bassa allora rifiutiamo l'ipotesi che abbiamo fatto. 5% di solito è la soglia in cui la statistica si presenta e si va a rifiutare.
l'ipotesi da formulare sono:

- **Ipotesi di differenza**: 2 gruppi possono differire in diversi modi, infinite possibilità
- **Ipotesi di uguaglianza**: i due gruppi sono uguali, c'è solo una possibilità. L'ipotesi di uguaglianza si divide in ipotesi nulla non c'è differenza tra parametri campione e popolazione (H_0) $\boldsymbol{x}$ medio $= \mu$, mentre nell'ipotesi alternativa sono differenti (H_1) $\boldsymbol{x}$ medio diverso da μ.

Quando trovo dei valori di statistica che ha una probabilità di verificarsi del 95% accetto l'ipotesi, se invece è sotto il 5% la rifiuto.

1. Formulo ipotesi di uguaglianza
2. Si decide soglia oltre la quale si accetta o rifiuta H_0
3. Si usa un test statistico
4. Si ottiene una statistica test = grandezza calcolata a partire dai dati per valutare il grado di compatibilità dei risultati con l'ipotesi nulla
5. Esiste per ciascuna ipotesi un probabilità (p-value, ossia la probabilità di ottenere i dati osservati se H0 fosse vera), se è piccolo (< 0,05) l'ipotesi nulla non si accetta e ma H1 si. Se è uguale a 0,05 è ambiguo. 0,05 è chiamato livello di significabilità e si indica con alfa.

Errore di I e di II tipo

Si possono fare però degli errori, ce nel sono di due tipi:

1. Rifiuto H 0 quando è vera ►ERRORE DI I TIPO
2. Accetto H 0 quando è falsa ►ERRORE DI II TIPO

Decido:	H_0 vera	H_0 falsa
Accetto l'ipotesi	OK! Corretto! P = 1 – α Livello di protezione	ERRORE di II TIPO P = β
Rifiuto l'ipotesi	ERRORE di I TIPO P = α Livello di significatività	OK! Corretto! P = 1 - β Potenza del test

- Se l'ipotesi nulla è vera e si rifiuta, si commette un **errore di I tipo**
- La probabilità di commettere un errore di I tipo è pari ad **α**
- Il **livello di protezione** del test contro la possibilità di commettere un errore di I tipo è pari a **1 - α**

- Se l'ipotesi nulla è falsa e si accetta, si commette un **errore di II tipo**
- La probabilità di commettere un errore di II tipo è pari a **β**
- La probabilità che un campione casuale determini il rifiuto di un'ipotesi nulla falsa è detta **potenza del test** e si indica con **1 - β**

Se diminuisco alfa ma aumenta l'errore di tipo beta. In questo modo diminuisce la potenza del test. Le medie di H0 H1 devono essere distanti tra di loro, tanto più il test è protetto e potente.

L'ideale è che il test sia ben protetto da una parte e potente dall'altra. Quanto più le medie che definiscono H0 e H1 sono distanti tra loro, non tanto in valore assoluto quanto in termini di sigma, tanto più il test sarà protetto e potente. È necessario però che questa differenza misurata in termini di sigma ponga le due distribuzioni ben separate e con la minore sovrapposizione Possibile. Se le due distribuzioni sono ben separate, sigma è piccolo

Confronti fra medie

martedì 22 dicembre 2020
23:12

1. Media di un singolo gruppo confrontata con la media della popolazione Si possono verificare 2 casi:

σ noto
σ ignoto

2. Medie di 2 gruppi di osservazioni che possono essere

Indipendenti: raccolti per 2 popolazioni o trattamenti
Appaiati (a coppie): i dati possono essere raccolti

- Per due trattamenti a ciascuna unità campionaria

- Per due unità campionarie appaiate in base ad alcune variabili (es: età, peso, ecc

- Confronto tra medie con deviazione standard nota

Si usa Z, è una media campionaria, media campione - media popolazione diviso (sigma/radice di N).
H0 : X medio = media popolazione.

Si ottiene **P-value** dalla tavole Z. Calcolo I limiti di confidenza con $\bar{x}(\pm)1{,}96 * \frac{\sigma}{\sqrt{n}}$

Problema:
Variabile = concentrazione di urea nel sangue di bovini
Campione di 10 animali: media = 22 mg/100 ml.
In letteratura: ampia indagine → $\mu = 25$; $\sigma^2 = 45$.
Domanda: posso affermare che il campione proviene da una popolazione con media 25 e varianza 45?

Ipotesi nulla: il campione proviene dalla popolazione normalmente distribuita, con $\mu = 25$ e $\sigma^2 = 45$:

$H_0 : \mu = \mu_0 = 25$ mg

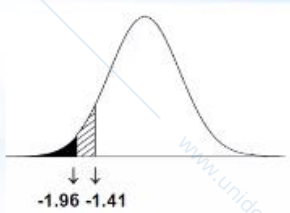
μ = media del campione, μ_0 = media della popolazione

Livello di significatività: $\alpha = 0.05$

Si utilizza il test a 2 code

Valore critico: Z = 1.96, Z = -1.96

$$\sigma_{\bar{x}} = \sqrt{\frac{45}{10}} = 2.12$$



① $z = \frac{\bar{x}-\mu}{\sigma_x} = \frac{22-25}{2.12} = -1.41$

② **P(z = -1.41) = 0.079**

Dai risultati ottenuti sia con la statistica test ① sia con il P-value ② si giunge alla medesima conclusione

① **z non supera il valore critico di -1.96 ► accetto H_0**

② **P(z = -1.41) = 0.079**

$0.079 \times 2 = 0.158 > \alpha = 0.05$ ► accetto H_0

Accetto l'ipotesi nulla (ossia la media del campione è uguale a quella della popolazione)

Gli intervalli di confidenza al 95% della media della popolazione sono:

$$\bar{X} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

$I_1 = 22 - 1.96 * 2.12 = 17.84$ mg/100 ml
$I_2 = 22 + 1.96 * 2.12 = 26.15$ mg/100 ml

Siamo fiduciosi al 95% che i limiti di 17.84 e 26.15 mg/100 ml contengono la media della popolazione che è pari a 25 mg/100ml

- Confronto tra medie con deviazione standard non nota

Si utilizza la distribuzione di t.

*In un campione di 36 suini è stato rilevato l'incremento medio ponderale giornaliero (IMPG, Daily Weight Gain = DWG) e si sono ottenuti i seguenti risultati

$\bar{X}$ = 599.194 g s = 18.656 g Es = $\frac{s}{\sqrt{n}}$ = 3.109 g

L'IMPG calcolato sulla popolazione è di 607 g.
C'è differenza tra i 2 valori?

H_0: $\bar{X} = \mu$ ossia $\bar{X} - \mu = 0$

$t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{N}}} = \frac{599.194-607}{3.109}$ =-2.51

1° modo) Dalle tavole: $t_{(\alpha=0.05;35\ GL)}$ compreso tra 2.042 e 2.021

30	0.683	0.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
40	0.681	0.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551

t =2.51 < t tabulato ►Rifiuto H_0

2° modo) Si usa la funzione di Excel: DISTR.T.2T(2.51;35)

P(t = -2.51, 35 GL) = **0.017** (=1,7% < 5%) ►Rifiuto H_0

Rifiuto H_0: la media del campione e quella della popolazione sono differenti

L'intervallo di confidenza al 95%:

$$\bar{X} \pm t_{(n-1,\alpha=0.05)} * \frac{s}{\sqrt{n}}$$

$t_{(n-1,\alpha=0.05)} = t_{(35,\ \alpha=0.05)} = 2.03$

599.10 ± 2.03*3.109 = 592.8 / 605.3 g

Sono i limiti entro cui troviamo la media della popolazione (μ = 607) da cui è stato tratto il campione di 36 suini e che non comprende il valore di 607: i 36 suini non provengono da una popolazione con un IMPG di 607 g

Confronti tra medie di 2 campioni

Sono campioni indipendenti estratti da popolazioni normalmente distribuite, con media mu1 e mu2 , e con varianza non nota.
Si valuta se le differenze osservate tra le medie di 2 campioni sono troppo grandi per essere attribuite al caso:

1. Si formula un'ipotesi nulla : nella maggior parte dei casi si testa se le medie delle 2 popolazioni sono uguali H0 : $\boldsymbol{X1=X2}$
2. Si stabilisce il livello di probabilità α
2. Si calcola la probabilità di ottenere una coppia di medie così differenti
3. Se la probabilità è più piccola di α , rifiutiamo l'ipotesi nulla

- Confronto tra medie di due campioni indipendenti

Le osservazioni sono assegnate casualmente a due trattamenti. Quando la t trovata è maggiore dei valori tabulati si rifiuta il test.

indipendenti

$$t_{[gl]} = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

gl (gradi di libertà) è la somma dei gradi di libertà dei 2 campioni →$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$

$s_{\bar{x}_1 - \bar{x}_2}$ è l'errore standard della differenza tra le medie dei 2 campioni

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\bar{s}^2 * \frac{n_1 + n_2}{n_1 * n_2}}$$

Per calcolarla, è necessario calcolare la varianza ponderata:

$$\bar{s}^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

E' stato misurato il cortisolo ematico in 18 cani, 9 maschi e 9 femmine. Ci si chiede se il sesso influenza il livello di cortisolo. Sono stati ottenuti i seguenti risultati

Sesso	n	Media	s
M	9	61,59	49,04
F	9	70,29	28,36

Domanda: le medie dei 2 sessi differiscono in modo significativo?

$H_0: \bar{X}_M = \bar{X}_F$

Livello di significatività: $\alpha = 0.05$

Utilizziamo la statistica test t- Student

Sesso	n	Media	s	s^2	gl
M	9	61,59	49,04	2404,92	8
F	9	70,29	28,36	804,29	8

$\bar{s}^2 = \frac{(2404,92x8)+(804,29x8)}{(9-1)+(9-1)} = 1604,61$

$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{1604,61 * \frac{9+9}{9*9}} = 18.88$

$t = \frac{61.59 - 70.29}{18.88} = -0.46$

il t tabulato per alfa=0,05 è 2.12, maggiore del t calcolato, quindi accetto ipotesi H0

- Confronto tra medie di due campioni appaiati

I due trattamenti sono applicati a ciascuna unità sperimentale (gruppo di cane in cui ognuno è sottoposto a due trattamenti)

Il test *t* per dati appaiati si riduce ad un semplice test t dopo aver calcolato la differenza tra le misure effettuate su ogni unità sperimentale

* H_0: $\mu_d = 0$

$$t = \frac{\bar{d}}{ES_{\bar{d}}}$$

* $\bar{d}$ è la differenza media tra le osservazioni di ogni unità
* $ES_{\bar{d}}$ è l'errore standard della differenza

* Data l'ipotesi nulla, la statistica *t* ha una distribuzione *t* con GL = n - 1

In molte specie i maschi hanno maggiori possibilità di attrarre le femmine se presentano elevati livelli di testosterone. Un'ipotesi è che i maschi con più elevati livelli di testosterone siano più vulnerabili alle malattie. Per verificare questa ipotesi, un gruppo di ricercatori hanno aumentato sperimentalmente i livelli di testosterone di 13 maschi di merlo alirosse (*Agelaius phoeniceus*), impiantando un tubicino permeabile contenente testosterone. Il livello anticorpale (ln[mDO/min]) è stato misurato prima e dopo l'impianto

ID Maschio	Produzione anticorpi PRIMA	Produzione anticorpi DOPO	d
1	4.65	4.44	0.21
4	3.91	4.30	-0.39
5	4.91	4.98	-0.07
6	4.50	4.45	0.05
9	4.80	5.00	-0.20
10	4.88	5.00	-0.12
15	4.88	5.01	-0.13
16	4.78	4.96	-0.18
17	4.98	5.02	-0.04
19	4.87	4.73	0.14
20	4.75	4.77	-0.02
23	4.70	4.60	0.10
24	4.93	5.01	-0.08

H_0: $\mu_{d0} = 0$

$\bar{d} = -0.056$

$s_d^2 = 0.025$ (varianza)

$s_d = 0.159$ (deviazione standard)

$ES_d = \frac{s_d}{\sqrt{n-1}} = \frac{0.159}{\sqrt{12}} = 0.044$

$t = \frac{\bar{d}}{ES_d} = \frac{-0.056}{0.044} = -1.27$

$t_{(\alpha=0.05;GL=12)} = 2.18$

Poiché 1.27<2.18, accetto H_0: il testosterone non ha effetto sull'immunocompetenza dei maschi

Da: Analisi statistica dei dati biologici (2010), Ed. Zanichelli

Associazione tra due variabili continue

lunedì 9 novembre 2020
14:22

Correlazione

Rappresentata con un grafico dispersione. Su ascisse c'è una variabile, sulle ordinate un'altra. La nuvola di punti permette di fare osservazioni:

- Alcune osservazioni sono estreme

c'è un coefficiente di correlazione che dice com'è la relazione tra le variabili, se l'aumentare di uno corrisponde aumentare dell'altro o una diminuzione.

COVARIANZA = variabilità congiunta di due variabili continue. Il numeratore è chiamato CODEVIANZA. La covarianza può essere positiva o negativa.

$$s_{xy} = Cov_{(x,y)} = \frac{\sum_i (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Varianza

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

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Numeratore: **DEVIANZA**

Covarianza

$$s_{xy} = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{n-1}$$

$$s_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1}$$

Numeratore: **CODEVIANZA**

La correlazione è:

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}}$$

Varia da -1 a +1:

- -1 associazione perfetta negativa, tutti i punti stanno su una retta in direzione verso il basso e verso destra
- +1 associazione perfetta positiva, tutti i punti stanno su una retta in direzione verso l'alto e verso destra
- 0 assenza di associazione (covarianza = 0)

$0 < r < 0{,}3$ correlazione debole
$0.3 < r < 0.7$ correlazione moderata
$r > 0.7$ correlazione forte.

Analisi della correlazione

H_0: $r_{xy} = 0$ ▶ assenza di correlazione

1) Si calcola la statistica t:

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

Dove:
r = coefficiente di correlazione
n-2 = GL (2 GL in meno perché bisogna stimare $\bar{X}$ e $\bar{Y}$)

2) Si confronta il t ottenuto con la distribuzione teorica di t_{n-2}

$t = r\sqrt{\frac{n-2}{1-r^2}}$ deriva da $t = \frac{r}{ES_r}$ e $ES_r = \sqrt{\frac{1-r^2}{n-2}}$

Se t trovato è maggiore del t tabulato si rifiuta l'ipotesi H0, che dice che il coefficiente di correlazione è = a 0

I coefficienti di correlazioni tra 3 o più variabili possono essere sintetizzati in una tabella con 2 entrate di n righe x n colonne, dove n è il numero di variabili, risulta una matrice (tabella speculare rispetto alla diagonale).

Matrice di correlazione

I coefficienti di correlazioni (con relativa P) tra 3 o più variabili possono essere sintetizzati in una tabella a 2 entrate di *n* righe x *n* colonne, dove *n* è il numero di variabili → MATRICE (tabella speculare rispetto alla diagonale)

Esempio: 3 variabili relative alla infestazione di Varroa d. nelle api

1. **Caduta naturale** CN
2. **Varroa foretica** VF
3. **Comportamento igienico** HB

In rosso r; in blu P – H_0: r = 0

	CN	VF	HB
CN	1	0.823 *<.0001*	0.109 *0.467*
VF	0.823 *<.0001*	1	0.079 *0.599*
HB	0.109 *0.467*	0.079 *0.599*	1

Regressione

Misura la dipendenza di una variabile y dipendente dall'altra, la x indipendente. Misura quanto varia la y al variare di un'unità della x. Il coefficiente di regressione è:

$$b = \frac{Cov(x,y)}{Var(x)}$$

Espressa nell'unità di misura della y.

La **retta migliore** è quella che minimizza gli scarti tra i punti dei dati (valori assoluti) e i punti sulla retta (valori attesi). La retta per la quale sia minima la somma di tutti i quadrati degli scarti è detta METODO MINIMO DEI QUADRATI.

La retta è descritta matematicamente dalla seguente equazione:

$$y = a + bx$$

dove

$$b = \frac{Cov_{(x,y)}}{V_x}$$

$$a = \bar{y} - b\,\bar{x}$$

a = costante = intercetta: esprime il valore di y (ordinata) quando x è uguale a 0

b : esprime la pendenza della retta di regressione ossia *la variazione di y quando x varia di un'unità:*

b di segno positivo: retta sale verso destra (le due variabili sono associate positivamente)

b di segno negativo: retta scende verso destra (le due variabili sono associate positivamente)

b uguale a zero: retta è orizzontale, parallela all'asse delle x (non c'è variazione congiunta)

Negativamente non positivamente

Modello statistico di regressione lineare semplice :

$$y_i = a + bx_i + e_i$$

Qualsiasi i-ma osservazione (y_i) è funzione (= il risultato) di
1) una *costante* (a)
2) un *coefficiente di regressione* (b) che moltiplica
   il corrispondente i-mo valore di x (x_i)
3) un ulteriore effetto casuale detto *errore* (e_i).

Il modello di regressione (come ogni altro modello) è uno *strumento di previsione*

Si può calcolare per ciascuna osservazione, y, il valore atteso E(y)

$$y_i = a + bx_i + e_i$$
$$y_i = E(y_i) + e_i$$
$$E(y_i) = a + b_i$$
$$e_i = y_i - E(y_i)$$

Data l'equazione: y =a + b(x):

$$a = \bar{y} - b(\bar{x})$$ INTERCETTA

$\bar{y}$ = 18 : 6 = 3

$\bar{x}$ = 72 : 6 = 12

$a = \bar{y} - b(\bar{x}) = 3 - (+0.8823) * (12)$= -7.59 INTERCETTA

b = 0.88

y = -7.59 + (0.88)x ▶ VALORE ATTESO = E(y)

i	y_i	x_i	$E(y_i)$	e_i
1	2	13	3.85 (-7.59+0.88*13)	-1.85 (2 - 3.85)
2	6	15	5.61 (-7.59+0.88*15)	0.39 (6-5.61)
3	1	10	1.21 (-7.59+0.88*10)	-0.21 (1 - 1.21)
4	3	12	2.97 (-7.59+0.88*12)	0.03 (3 – 2.97)
5	0	8	-0.55 (-7.59+0.88*8)	0.55 (0-(-0.55))
6	6	14	4.73 (-7.59+0.88*14)	1.27 (6 – 4.73)

Sono riportati i pesi e le età di agnelli di razza Bergamasca. Calcolare il peso di un agnello a 17 mesi di età

X	Y
Età (mesi)	Peso (kg)
1	6
2	22
3	22
4	35
5	41
6	47
7	56
8	59
9	64
10	68
11	72
12	79
13	80
14	86
15	92

$\bar{x}$= 8 mesi
$\bar{y}$ = 55.27 kg
Var(x) = 20 mesi^2
Var(y) = 666.0667 kg^2
Cov(x,y) = 113.9286 kg*mesi
b = $\frac{Cov_{xy}}{var_x}$ = 5.69
A = $\bar{y} - b * \bar{x}$ =55.27-(5.69*8) = 9.69

Equazione di previsione:
Y (peso) = 9.69 + 5.69 * x (età)

Peso previsto a 17 mesi = 9.69 + 5.69*17 =106.5

Chi quadro

lunedì 23 novembre 2020
15:44

Il valore di χ^2 ottenuto dai dati sperimentali è quindi associato ai gradi di libertà:

1) Se non si stima alcun parametro

gradi di libertà = N° classi degli eventi (o categorie) -1

2) Se si stimano alcuni parametri dai dati

gradi di libertà = N° classi degli eventi (o categorie) -1

- N° parametri stimati in base al campione

si confronta con i valori di χ^2 tabulati per quei gradi di libertà

H_0 : frequenze **o**sservate = frequenze attese

χ^2 > del valore tabulato di χ^2	RIFIUTO H_0
χ^2 < del valore tabulato di χ^2	ACCETTO H_0

Mette a confronto le frequenze, tra le frequenze osservate e le frequenze attese

$$\chi^2 = \frac{(frequenze\ osservate - frequenze\ attese)^2}{frequenze\ attese}$$

I test chi quadrato è utilizzato per verificare che le frequenze dei valori osservati si adattino alle frequenze teoriche di una distribuzione di probabilità prefissata.
Le frequenze attese sono quelle che ci attendiamo in base a una probabilità.
Se numeratore è 0 le frequenze attese coincidono con quelle osservate. Le frequenze sono sempre quelle assolute.

Chi-quadro si ottiene sommando per ogni evento o classe o categoria il quadrato della differenza tra frequenze osservate e attese diviso per le frequenze attese. Ha k-1 gradi di libertà. Si sommano per ogni classe il relativo chi-quadro
La distribuzione del chi-quadro è simile a quella di F, parte da 0 e va a infinito, è **asimmetrica**, si va a guardare la coda a destra.

Dai bollettini AIA 1996, risulta che nella popolazione delle bovine di razza Frisona italiana sottoposte a controllo funzionale per la produzione di latte, il 32% delle bovine è in prima lattazione, il 24% in seconda, le rimanenti sono in lattazioni superiori alla seconda. In un'azienda in provincia di Milano sono state osservate le frequenze riportate in tabella. Il numero di bovine per lattazione si accorda con i dati AIA?

Lattazione	N. bovine
1	1219
2	911
≥3	1021

H_0: frequenze osservate nel campione = frequenze attese in base ai dati AIA

classe	freq. osservate	freq. attese	χ^2
I latt.	1219	3151x0.32 = 1008.32	44.0198
II latt.	911	3151 x 0.24 = 756.24	31.671
> II latt.	1021	3151 x 0.44 = 1386.44	96.3232
totale	3151	3151	172.014

$\chi^2_{[2]} = 172.014 \rightarrow P<0.0000 \rightarrow$
ALTAMENTE SIGNIFICATIVO → RIFIUTO L'IPOTESI NULLA

Anche in questo caso si va a guardare il 5% della coda a destra. I gradi di libertà sono il numero delle classi confrontate - 1

In generale: quando si è interessati a stabilire se alcune serie di dati

- Seguono una distribuzione teorica (normale, binomiale, Poisson)
- Seguono una distribuzione attesa determinata da circostanze biologiche

Si usa il test del Chi Quadro

Nella specie umana il fatto di generare un maschio è indipendente dalla probabilità che anche il successivo sia maschio. Ci si attende che il N° di maschi e il numero di femmine in una famiglia con 2 figli segua una distribuzione binomiale, con n =2 e p (P di avere un maschio) =0.5106. In un'indagine su 2444 famiglie sono stati trovati i risultati riportati.

N° maschi	N° famiglie (con 2 figli)
0	530
1	1332
2	582
Totale	2444

Si adattano ad una distribuzione binomiale?

p (maschio) = 0. 5106 **q (femmina) = 1 – 0.5106 = 0.4894**

Sulla base della distribuzione binomiale si calcola la P di avere famiglie con 0, 1 e 2 maschi su 2 figli

$P(0M, n=2) = \frac{2!}{0!*(2-0)!} * 0.5106^0 * 0.4894^2 = 0.2395$

$P(1M, n=2) = \frac{2!}{1!*(2-1)!} * 0.5106^1 * 0.4894^1 = 0.4997$

$P(2M, n=2) = \frac{2!}{2!*(2-2)!} * 0.5106^2 * 0.4894^0 = 0.2607$

H_0: i dati osservati sono uguali ai dati attesi in base alla distribuzione binomiale

N° maschi	N° famiglie(con 2 figli) OSSERVATE	N° famiglie(con 2 figli) ATTESE	χ^2
0	530	(0.2395 * 2444) = 585.3	5.237
1	1332	(0.4997 * 2444) = 1221.4	10.005
2	582	(0.2607 * 2444) = 637.3	4.779
Totale	**2444**	**2444**	**20.021**

GL = 3 - 1 = 2

Valore di χ^2 tabulato per 2 gl e $\alpha = 0.05$ ▶ 5.99

Inferiore a quello trovato (= 20.021)

TEST DI CONTINGENZA

Si usa per verificare se le osservazioni sono contingenti, ossia dipendenti, dalle condizioni in cui si sono verificate.

Si usa quando si vuole vedere se esiste una relazione tra due variabili categoriche: si incrociano queste due variabili in una **tabella di contingenza**

In ogni casella della tabella troviamo il numero di osservazioni che presenta **una particolare combinazione delle categorie delle due variabili**.

Si usano le **tabelle di contingenza** (2 x 2, 2 x 3, 3 x 3 ecc.)

H_0 : non esiste dipendenza tra le variabili categoriche

TEST DI CONTINGENZA: frequenze osservate

Studio sull'efficacia dei caschi protettivi per ciclisti nella prevenzione dei traumi cranici. Su un totale di 793 ciclisti che hanno subito incidenti, 235 hanno riportato un trauma e, fra questi, 17 portavano il casco. Fra i restanti ciclisti (558) che non hanno riportato un trauma, 130 portavano il casco.

Si riportano i risultati dell'indagine (**FREQUENZE OSSERVATE**):

	Con casco	Senza casco	Totale
Trauma	17	218	235
No trauma	130	428	558
Totale	147	646	793

P di riportare traumi: 235/793 = 29.6%

P di non riportare traumi: 558/793 = 70.4%

P di portare il casco: 147/793 = 18.5%

P di non portare il casco: 646/793 = 81.5%

Esiste una relazione tra l'uso del casco nei ciclisti e la presenza di trauma in caso di incidente?

TEST DI CONTINGENZA: calcolo frequenze attese

Se non ci fosse nessuna relazione fra la presenza di un trauma e il fatto di portare il casco, le percentuali totali di riga dovrebbero essere uguali in tutte le colonne

	Con casco	Senza casco	Totale
Trauma	29.6%	29.6%	235 (29,6%)
No trauma	70.4%	70.4%	558 (70.4%)
Totale	147	646	793

CALCOLO FREQUENZE ATTESE

1) con casco

147 * 0.296 = 43.6 soggetti con traumi

147 * 0.704 = 103.4 soggetti senza traumi

2) senza casco

646 * 0.296 = 191.4 soggetti con traumi

646 * 0.704 = 454.6 soggetti senza traumi

	Con casco	Senza casco	Totale
Trauma	43.6	191.4	**235**
No trauma	103.4	454.6	**558**
Totale	**147**	**646**	**793**

TEST DI CONTINGENZA: calcolo frequenze attese

Se non ci fosse nessuna relazione fra la presenza di un trauma e il fatto di portare il casco, le percentuali totali di colonna dovrebbero essere uguali in tutte le righe

	Con casco	Senza casco	Totale
Trauma	18.5%	81.5%	235
No trauma	18.5%	81.5%	558
Totale	147 (18.5%)	646 (81.5%)	793

CALCOLO FREQUENZE ATTESE

1) Trauma

235 * 0.185 = 43.6 soggetti con casco

235 * 0.825 = 191.4 soggetti senza casco

2) No trauma

558 * 0.185 = 103.4 soggetti con casco

558 * 0.185 = 454.6 soggetti senza casco

	Con casco	Senza casco	Totale
Trauma	43.6	191.4	**235**
No trauma	103.4	454.6	**558**
Totale	**147**	**646**	**793**

	Con casco	Senza casco	Totale
Trauma	17	218	235
No trauma	130	428	558
Totale	147	646	793

Frequenze osservate

	Con casco	Senza casco	Totale
Trauma	43.6	191.4	235
No trauma	103.4	454.6	558
Totale	147	646	793

Frequenze attese

$$\chi^2 = \frac{(17-43.6)^2}{43.6} + \frac{(130-103.4)^2}{103.4} + \frac{(218-191.4)^2}{191.4} + \frac{(428-454.6)^2}{454.6}$$

$\chi^2 = 28.26$

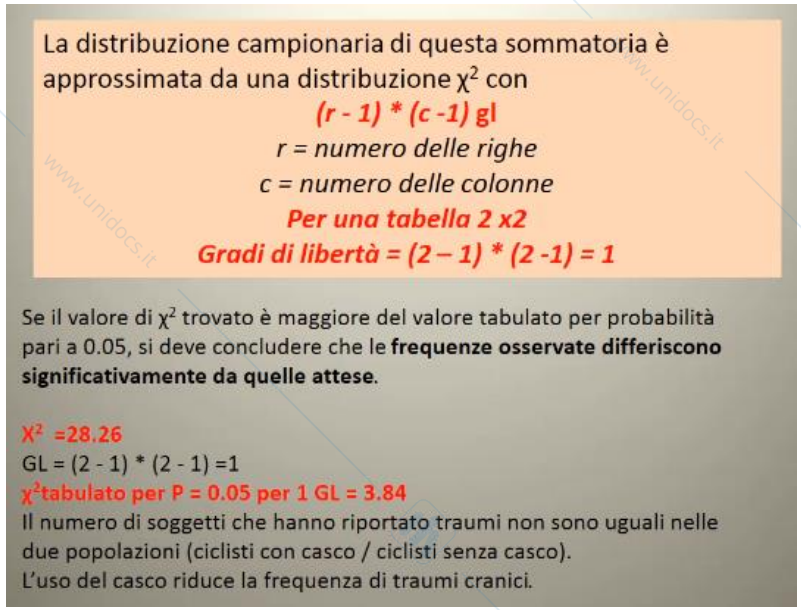
Calcolo delle frequenze attese per ciascuna cella

La frequenza attesa di ciascuna cella si può calcolare con la seguente formula

$$freq.attesa = \frac{totale\ riga\ x\ totale\ colonna}{totale\ generale}$$

	Con casco	Senza casco	Totale
Trauma	17 (235*147)/793 = 43.6	218 (235*646)/793 = 191.4	235
No trauma	130 (558*147)/793 = 103.4	428 (558*646)/793 = 454.6	558
Totale	147	646	793

Le frequenze attese si confrontano con le frequenze osservate mediante il test del chi-quadrato



Solo una categoria può variare e le altre sono legate a questa, quindi c'è 1 solo grado di libertà. Nelle tabelle 2x2 i gradi di liberà sono sempre 1.

Rifiuto ipotesi nulla, quindi visto che rifiuto il fatto che non c'era associazione, vuol dire che c'è associazione tra variabili categoriche.

Formula per calcolare direttamente il χ^2 senza calcolare le frequenze attese in una tabella 2 x 2

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d = N

a, b, c, d le frequenze osservate

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

Nell'esempio precedente

$$\chi^2 = \frac{(17*428 - 218*120)^2 * 793}{235*558*147*646} = 28.26$$

	Con casco	Senza casco	Totale
Trauma	17	218	235
No trauma	130	428	558
Totale	147	646	793

MISURA DI ASSOCIAZIONE – ODDS RATIO

La misura per valutare l'associazione tra esposizione ad un fattore di rischio ed una malattia è l'Odds Ratio (OR), cioè il rapporto tra **odds**

ODDS

Per una variabile in una singola prova sono possibili 2 risultati : successo/insuccesso

p = probabilità di successo **1 – p = probabilità di insuccesso**

L'odds di un successo è:

$$O = \frac{p}{1-p}$$

ODDS RATIO

E' il rapporto tra l'odds di un successo di un gruppo e l'odds di successo di un secondo gruppo:

$$OR = \frac{O_1}{O_2}$$

1) Calcolo dell'ODDS nei CASI

	Casi (M^+)	Controlli (M^-)	Totale
Esposti (Exp)	a	b	a + b
Non esposti (Exp^-)	c	d	c + d
Totale	a + c	b + d	

La probabilità che i malati siano stati esposti in passato al fattore di rischio è:

→a/a+c

La probabilità che i malati non siano stati esposti in passato al fattore di rischio è:

→c/a+c

$$\text{Odds esposizione nei malati} = \frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{a+c} x \frac{a+c}{c} = \frac{a}{c}$$

La probabilità che i non malati (o controlli) siano stati esposti in passato al fattore di rischio è:

→b/b+d

La probabilità che i non malati (o controlli) non siano stati esposti in passato al fattore di rischio è:

→d/b+d

$$\text{Odds esposizione nei non malati} = \frac{\frac{b}{b+d}}{\frac{d}{b+d}} = \frac{b}{b+d} x \frac{b+d}{d} = \frac{b}{d}$$

$$\text{Odds Ratio} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a}{c} x \frac{d}{b}$$

- **OR = 1→** l'esposizione al fattore di rischio non è associata alla malattia
- **OR > 1→** l'esposizione al fattore di rischio ha un effetto nocivo
- **OR < 1→** l'esposizione al fattore di rischio ha un effetto protettivo

LIMITI DI CONFIDENZA DELL'OR

IC è un range di valori che con una P del 95% contiene l'OR reale
La distribuzione campionaria dell'OR è asimmetrica a destra.
E' necessario convertire l'OR nel suo logaritmo naturale
I limiti al 95% sono

$$\ln(OR) \pm 1.96 * ES[\ln(OR)]$$

Dove

$$ES[\ln(OR)] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

E' quindi necessario riconvertire due valori logaritmici $\ln_1$ ed $\ln_2$ dell'intervallo di confidenza in OR, mediante elevamento a potenza di e = 2.71828

$$OR_1 = e^{\ln_1} = 2.71828^{\ln_1}$$
$$OR_2 = e^{\ln_2} = 2.71828^{\ln_2}$$

L' IC dell'OR consente di verificare la significatività

- Se contiene 1, H0 è accettata: mancanza di associazione
- Se non contiene 1, H0 rifiutata: associazione significativa

Calcolo dell'intervallo di confidenza

1) Convertire Or nel suo logaritmo naturale
OR = 2.19 ln(2.19) =0.784

2) Calcolo ES(lnOR)

$$ES[\ln(OR)] = \sqrt{\frac{1}{383}+\frac{1}{322}+\frac{1}{631}+\frac{1}{1165}} = 0.09$$

3) Calcolo limiti
LI 0.784 – 1.96 * 0.09 = 0.607
LI 0.784 + 1.96 * 0.09 = 0.96

0.607< ln(OR) < 0.96

4) Riconvertire i valori logartimici
LI $e^{0.607} = 1.83$
LS $e^{0.967} = 2.61$

1.83<OR<2.61

L'intervallo non contiene 1, quindi l'associazione tra obesità e calcolosi urinaria è significativa