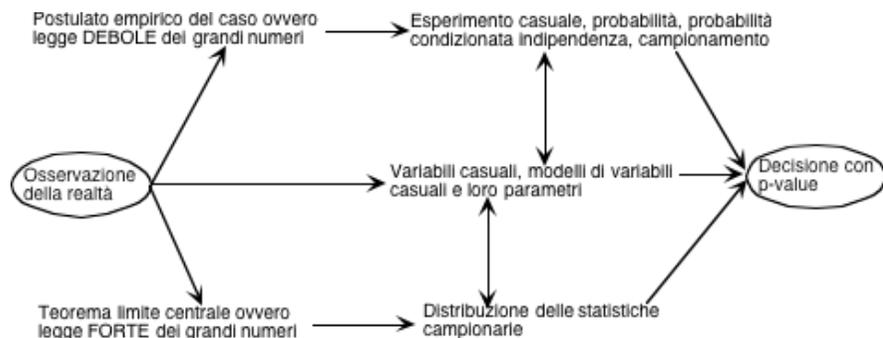


Schema dell'inferenza statistica



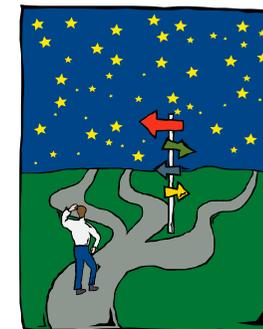
Test delle ipotesi

Si deve stabilire il grado di attendibilità di una ipotesi statistica

Se la decisione si potesse basare sulla conoscenza totale si avrebbe una conclusione definitiva: l'ipotesi è VERA o FALSA.

Come in molte scienze sperimentali non potremo dimostrare vera o falsa una affermazione.

Potremo solo affermare: è più coerente o meno coerente con i nostri dati campionari.



Formalismo dei Test

L'IPOTESI STATISTICA H_0 è una asserzione verificabile su di una variabile casuale. In genere riguarda i suoi parametri.

Supponiamo di conoscere la funzione di densità della v.c.: $f(X;\theta)$ di cui però ignoriamo il valore del parametro θ .

In genere, la H_0 ipotizza un certo valore del parametro e ne valuta la sua conformità con i dati campionari

$$H_0 : \theta - \theta_0 = 0$$

È detta "nulla" perché, di solito, niente cambia se è accettata

Formalismo dei Test/2

il valore θ_0 è scelto in base a varie considerazioni:



Dalla logica dell'esperimento



Da esperienze precedenti o indagini pilota



E' un livello critico (*baseline*) da cui allontanarsi



E' un livello desiderato

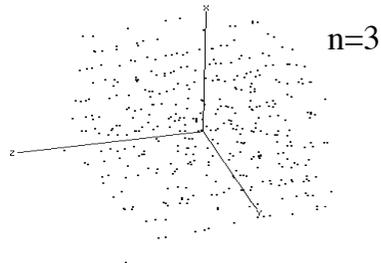


Formalismo del Test/3

Consideriamo L'UNIVERSO DEI CAMPIONI di ampiezza "n" cioè l'insieme di tutte le possibili realizzazioni della n-tupla ...

$$(X_1, X_2, \dots, X_n)$$

In tale spazio, il campione estratto è solo un punto.



L'idea del test è di assegnare ad ogni campione una probabilità detta

LIVELLO OSSERVATO DI SIGNIFICATIVITA' (p-Value)

e in base a questo decidere sull'ipotesi H_0 .

Formalismo del test/4

Le fluttuazioni campionarie fanno ritenere che ci siano differenze più o meno grandi tra campione e popolazione da cui è estratto.

Dobbiamo dotarci di uno strumento che ci aiuti a giudicare se le differenze sono significative (rifiutiamo H_0) o poco significative (non rifiutiamo H_0)

Le differenze sono da considerarsi significative se è bassa la probabilità che siano ascrivibili solo a mere variazioni campionarie.

Quanto bassa?

E' soggettivo!



La statistica test

E' l'anello di congiunzione tra universo dei campioni e valori del parametro

La distribuzione della statistica test $T(X;\theta)$ è funzione:

Del campione casuale (X_1, X_2, \dots, X_n)

Del parametro da stimare: θ



Le statistiche test più in uso sono del tipo:

$$T(X;\theta) = \frac{T - \theta}{\sigma(T)}$$

Applicazione/1

Una docente sa che -storicamente- i risultati dei suoi esami scritti hanno $\mu=25$, $\sigma=3$.

Però, nell'ultima prova, i risultati dei primi 10 compiti sono molto scadenti. Che si tratti un corso ad alta densità di ciucci?

$$\text{Ipotesi nulla } H_0 : (\mu - 25) = 0$$

$$\text{Ipotesi alternativa } H_1 : \mu < 25$$

Come statistica test sembra ovvio scegliere la media campionaria.

La docente rifiuterà H_0 se il voto medio osservato nel campione sarà molto più piccolo di 25.

Il fatto che si rifiuti H_0 non implica necessariamente che accetti H_1 . Potrebbe anche decidere di rinviare la decisione in attesa di avere più dati.

L'ipotesi alternativa è inserita soprattutto per stabilire la direzione del test



Applicazione/2

N.B. si è sostituito l'evento punto $\mu=23$ con l'evento intervallo $\mu \leq 23$

nel campione si trova: $\hat{\mu} = 23$

Se l'ipotesi H_0 fosse vera, quale sarebbe la probabilità di osservare una media campionaria inferiore o uguale a 23?

Supponendo che la distribuzione sia gaussiana, avremo

"I" significa: sotto H_0 "

$$P(\hat{\mu} \leq 23 | H_0) = P\left(Z \leq \sqrt{10} \left(\frac{23-25}{3}\right) | H_0\right) = P(Z \leq -2.11 | H_0) = 0.0174$$

sotto l'ipotesi nulla (la media è 25) è altamente improbabile (1.7%) osservare un voto medio di 23 o meno nel campione estratto da una popolazione con $\mu=25$

Ne consegue che lo scarto 23-25 non è attribuibile alle fluttuazioni campionarie, ma fa invece pensare ad una classe particolarmente ciuccia.

Applicazione/3

Dopo aver corretto altri 25 compiti risulta che, su questi, si ha $\hat{\mu} = 27$

A questo punto sorge un altro dubbio: che l'assistente del corso, imbranato come una foca, abbia mischiato i compiti con quelli della classe "advanced"



Ipotesi nulla $H_0 : (\mu - 25) = 0$

La direzione del test è ora verso i valori alti poiché i valori inferiori a 25 non fanno sorgere questo tipo di dubbio

Ipotesi alternativa $H_1 : \mu > 25$

$$P(\hat{\mu} \geq 27 | \mu = 25) \Rightarrow P\left(\frac{\hat{\mu} - 25}{1.6667} \geq \frac{27 - 25}{1.6667}\right) = P(Z \geq 1.2) = 1 - 0.8849 = 0.1151$$

Lo scarto 25-27 è poco compatibile con l'ipotesi

Un valore superiore maggiore o uguale di 27 lo si può trovare nell'11% dei campioni da popolazioni aventi media 25. L'assistente si salva.

Esempio

La dott.ssa Angelina Romano propone una procedura d'ufficio che riduce i tempi medi rispetto agli attuali $\mu=75$ minuti, pur conservando lo stesso $\sigma=9$.



Ipotesi nulla $H_0 : (\mu - 75) = 0$

Ipotesi alternativa $H_1 : \mu < 75$

Si considera un campione di $n=25$ pratiche. Quindi la media del campione -sotto H_0 - sarà approssimata dalla gaussiana con $\mu=75$ e $\sigma=9/\sqrt{25}=1.80$

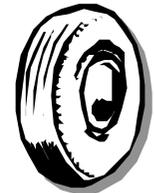
Nel campione si trova $\hat{\mu} = 69.8$

Qual'è la probabilità di ottenere un valore della statistica test inferiore o uguale del valore osservato se la media della popolazione è 75?

$$P(\hat{\mu} \leq 69.8 | \mu = 75) \Rightarrow P\left(\frac{\hat{\mu} - 75}{1.8} \leq \frac{69.8 - 75}{1.8}\right) = P\left(\frac{\hat{\mu} - 75}{1.8} \leq -2.89\right) = 0.0019$$

Esempio

L'amministratore delegato, Sig.ra Rosetta Gaudio, di una fabbrica di pneumatici sta valutando la modifica della trama del prodotto leader.



Lo studio di fattibilità segnala che la nuova trama è conveniente se la vita media dei prodotti supera le 20'000 miglia.

Ipotesi nulla $H_0 : (\mu - 20\,000) = 0$

Ipotesi alternativa $H_1 : \mu > 20\,000$

Un campione di $n=16$ prototipi viene provato dando luogo a $\hat{\mu} = 20,758$

Si sa che $\sigma=6'000$. L'amministratore Gaudio che deve fare?

$$P(\hat{\mu} \geq 20758 | \mu = 20000) \Rightarrow P\left(\frac{\hat{\mu} - 20000}{1500} \geq \frac{20758 - 20000}{1500}\right) = P\left(\frac{\hat{\mu} - 20000}{1500} \geq 0.39\right) = 1 - P\left(\frac{\hat{\mu} - 20000}{1500} < 0.39\right) = 0.3483$$

p-value

Indica la probabilità che valori della statistica test -inferiori o uguali a quello osservato- siano sopravvenuti solo per effetto della sorte.

Quindi, il P-value misura la probabilità di sbagliare, nelle condizioni date, se si rifiuta l'ipotesi nulla



Nuova procedura amministrativa

Ipotesi nulla $H_0: (\mu - 75) = 0$, $P - value = 0.0019$

La nuova procedura potrebbe essere non migliorativa rispetto alla vecchia solo in 2 casi su 1000 (circa). E' bene rifiutare H_0



Nuovo pneumatico

Ipotesi nulla $H_0: (\mu - 20\,000) = 0$, $P - value = 0.3483$

La nuova trama è migliorativa circa una volta su 3. Non è consigliabile rifiutare H_0 .

Precisazioni

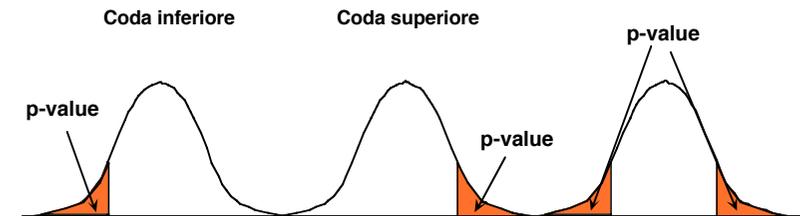
Rispetto all'ipotesi che il parametro abbia un valore prefissato ci sono tre casi:

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta < \theta_0 \end{cases}$$

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta > \theta_0 \end{cases}$$

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases}$$

Nei primi due il test è unidirezionale (o ad una coda), nel terzo è bidirezionale (o a due code).



P-Value/2

Dipende sia dalla distribuzione della statistica test che dal tipo di alternativa.

Nel caso della normale si ha:

$$\text{Se } H_1: \theta > \theta_0 \Rightarrow p\text{-value} = P(\hat{\theta} \geq \theta_c) = 1 - \phi(Z_c)$$

$$\text{Se } H_1: \theta < \theta_0 \Rightarrow p\text{-value} = P(\hat{\theta} \leq \theta_c) = \phi(Z_c)$$

$$\text{Se } H_1: \theta \neq \theta_0 \Rightarrow p\text{-value} = P(|\hat{\theta}| \geq \theta_c) = 2[1 - \phi(|Z_c|)]$$

Formule analoghe possono essere determinate per la t-Student e per le altre distribuzioni coinvolte nella verifica di ipotesi (chi-quadrato, F-Fisher, etc.)

Guida

Se $P\text{-value} \leq 1\%$.

Aldilà di ogni ragionevole dubbio si può rifiutare H_0

Se $1\% \leq P\text{-value} \leq 5\%$.

Ci sono buone ragioni per rifiutare H_0

Se $5\% \leq P\text{-value} \leq 10\%$.

Ci sono ragioni per rifiutare H_0 , ma non sono del tutto convincenti

Se $P\text{-value} > 10\%$.

E' consigliabile non rifiutare H_0



I valori sono solo apparentemente bassi.

Le condizioni di applicabilità dei test (ad esempio la distribuzione normale) sono valide solo in parte).

Di conseguenza, solo una forte evidenza può convincere a rifiutare l'ipotesi nulla (angolatura conservativa)

Esempio

Una compagnia controlla la situazione di magazzino in base alle vendite dell'anno precedente. Si supponga che gli ordini dell'anno precedente abbiano dato luogo

$$\mu=36 \text{ miliardi e } \sigma=2 \text{ miliardi}$$

Nel primo quadrimestre ($n=225$) si è riscontrato $\mu_x=35.82$. Ci sono cambiamenti?

$$H_0 : \mu = 36 \quad Z_c = \sqrt{225} \left(\frac{35.82 - 36}{2} \right) = -1.35$$

$$H_1 : \mu \neq 36$$

Poiché n è grande possiamo utilizzare la normale cioè la statistica test

Cioè circa 18 volte su 100, il campione di $n=225$ estratto dalla popolazione $N(36;2)$ disterà più di 1.35 dalla media ipotizzata. Il valore 35.82 non è allora troppo strano.

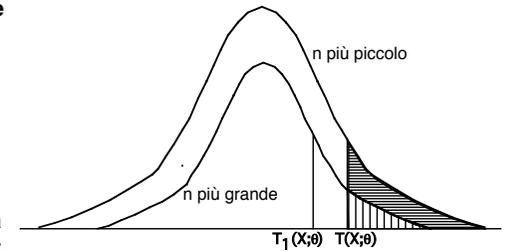
Non ci sono ragioni sufficienti per pensare ad un cambiamento

Ampiezza del campione e p-value

La statistica test è uno stimatore consistente del parametro sotto ipotesi.

Quindi, all'aumentare dell'ampiezza del campione, la sua variabilità si riduce.

Questo implica che le code della distribuzione della statistica test diventano più sottili.



A parità di p-value, la corrispondente statistica test è inferiore.

Ovvero, la stessa statistica test può avere un p-value più piccolo perché il campione è più grande.

ATTENZIONE!

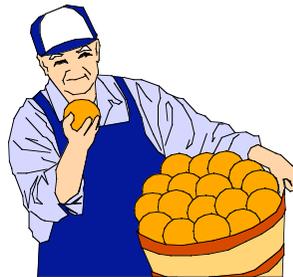
Campioni molto grandi possono rendere significative dei valori della statistica test poco rilevanti dal punto di vista pratico.

Esercizio

La produzione media per ettaro è stata di 18 Kg con $\sigma=7.12$. L'uso di un nuovo fertilizzante su $n=30$ ettari sperimentali, ha dato luogo a

$$\hat{\mu} = 19.5$$

Ipotesi nulla $H_0 : (\mu - 18) = 0$



$$P(\hat{\mu} \geq 19.5 | \mu = 18) \Rightarrow P\left(\frac{\hat{\mu} - 18}{7.12/\sqrt{30}} \geq \frac{19.5 - 18}{7.12/\sqrt{30}} \right) =$$

$$P(Z \geq 1.15) = 1 - P(Z < 1.15) = 0.1251$$

Il p-value è relativamente elevato. Non sembra ci sia sufficiente evidenza che il nuovo fertilizzante incida significativamente sulla produttività.

Excel1

| | | | |
|---------|---|--------------------------------|-----------------------|
| Importo | | | |
| 42.62 | Ampiezza | 34 | =Conta.Numeri(A2:A35) |
| 46.68 | Campionaria | | |
| 51.73 | | | |
| 62.48 | Media | 48.84 | =Media(A2:A35) |
| 50.28 | Campionaria | | |
| 43.86 | | | |
| 52.47 | Varianza nota | 25 | |
| 40.99 | Popolazione | | |
| 37.97 | | | |
| 46.00 | H ₀ : μ= | 50 | |
| 45.14 | H ₁ : μ<50 | | |
| 44.62 | | | |
| 36.49 | Statistica | -1.3501947 | =(D5-D11)/radq(D8/D2) |
| 52.35 | Test | | |
| 44.71 | | | |
| 46.21 | p-Value | 0.08847683 | =Distr.Norm.St(D14) |
| 54.02 | | | |
| 68.47 | Le ragioni per rifiutare H ₀ | non sono del tutto convincenti | |
| 51.52 | | | |
| 55.42 | | | |
| 40.73 | | | |
| 50.75 | | | |
| 41.78 | | | |
| 61.31 | | | |
| 49.74 | | | |
| 48.24 | | | |
| 42.00 | | | |
| 49.19 | | | |
| 49.81 | | | |
| 45.34 | | | |
| 57.93 | | | |
| 54.32 | | | |
| 46.90 | | | |
| 48.56 | | | |

importo medio speso in un super mercato da un campione di clienti



Excel2

Verifica di un possibile sbilanciamento di un indice aziendale

| | | | |
|--------|---|------------|-----------------------|
| Indice | | | |
| 19.36 | Ampiezza | 24 | =Conta(A2:A25) |
| 18.94 | Campionaria | | |
| 25.42 | | | |
| 18.93 | Media | 17.80 | =Media(A2:A25) |
| 10.37 | Campionaria | | |
| 11.12 | | | |
| 10.24 | Varianza nota | 36 | |
| 14.98 | Popolazione | | |
| 14.89 | | | |
| 17.80 | H ₀ : μ= | 15.0 | |
| 20.26 | H ₁ : μ > 0 | | |
| 18.74 | | | |
| 19.42 | Statistica | 2.28487638 | =(D5-D11)/radq(D8/D2) |
| 16.15 | Test | | |
| 11.02 | | | |
| 21.27 | p-Value (coda superiore) | 0.01116001 | =1-Distr.Norm.St(D14) |
| 21.11 | | | |
| 28.14 | | | |
| 29.43 | Si può rifiutare l'ipotesi che l'indice aziendale sia in equilibrio | | |
| 16.42 | Affermando che è sbilanciato verso l'alto si sbaglia l'1% | | |
| 13.54 | delle volte | | |
| 10.03 | | | |
| 16.42 | | | |
| 23.17 | | | |



Excel3

Oscillazioni di un corso azionario. Verifica della stabilità.



| | | | |
|--------|-------|--------------------------|-----------------------------------|
| Scarto | | | |
| -2.31 | -0.36 | Ampiezza | 42 =Conta(A2:A43) |
| -2.24 | -4.56 | Campionaria | |
| -3.35 | -1.97 | | |
| -0.84 | -1.37 | Media | -0.79 =Media(A2:A43) |
| -1.68 | 1.17 | Campionaria | |
| -0.75 | 1.17 | | |
| 1.36 | 0.25 | Varianza nota | 2.25 |
| -3.72 | -0.47 | Popolazione | |
| -1.88 | -0.16 | | |
| 1.31 | -2.51 | H ₀ : μ= | 0.0 |
| 0.37 | -1.22 | H ₁ : μ ≠ 0 | |
| -1.15 | -1.22 | | |
| 1.32 | -0.83 | Statistica | -3.41908683 =(D5-D11)/radq(D8/D2) |
| -1.70 | 0.54 | Test | |
| 1.11 | 0.31 | | |
| -1.29 | -1.71 | p-Value (coda inferiore) | 0.00031421 =Distr.Norm.St(D14) |
| 1.36 | -1.43 | P-value (coda superiore) | 0.99968579 |
| -1.81 | 0.76 | P-value (totale) | 0.00062842 =2*Min(D17,D18) |
| -1.67 | -1.47 | | |
| 0.64 | 0.28 | | |
| -1.05 | -0.49 | | |

Si può senzaltro rifiutare l'ipotesi che il corso azionario si stia stabilizzando intorno allo zero

Se σ è incognito...

Il calcolo del p-value ha finora ipotizzato che σ fosse noto e questo è molto irrealistico.

In genere "σ" è stimato con

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \hat{\mu})^2}{n-1}}$$

La statistica test in questo caso è basata sulla t-Student

$$t_c = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \sqrt{n} \left(\frac{\hat{\mu} - \mu}{s} \right)$$

Esempio

Una società che produce mobili ha sempre saputo che per costruire un arredo completo sono necessarie μ=20 ore/lavoro.

Di recente, si è anche assunta manodopera con contratto di formazione e si teme che questo faccia aumentare i tempi.

Dalla rilevazione dei tempi su di un campione casuale di n=9 arredi risulta:

25, 29, 23, 23, 31, 21, 27, 25, 33

$$\hat{\mu} = \frac{\sum_{i=1}^9 x_i}{9} = 26.33; \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^9 (x_i - \hat{\mu})^2}{8}} = 4$$

$$t_c = \sqrt{9} \left(\frac{26.33 - 20}{4} \right) = 4.748$$

TDIST(4.748;8;1)

$$P(t_c \geq 4.748) = 1 - P(t_c \leq 4.748) = 0.0007$$

I tempi sono realmente aumentati, chi affermasse il contrario direbbe la verità solo 7 volte su 10'000.



Esempio

L'Antitrust ha disposto un controllo sulla compagnia aerea "Facta" che propone un volo AZ tra due noti scali a 110 minuti.



In un campione di $n=49$ si trova:

$$\hat{\mu} = 108 \quad \hat{\sigma} = 7$$

Ipotesi nulla $H_0: (\mu - 110) = 0$

Ipotesi alternativa $H_1: \mu \neq 110$

$$t_c = \sqrt{49} \left(\frac{108 - 110}{7} \right) = -2,$$

$$P(t_c \leq -2) = 1 - P(t_c \geq 2) = 1 - [1 - P(t_c \leq 2)] = P(t_c \leq 2) = 0.0256$$

$$p\text{-value} = 2(0.0256) = 0.0512$$

La dichiarazione della compagnia non sembra infondata, almeno alla luce del campione analizzato

Esercizio



La biologia di molti laghi cambia in peggio a causa delle piogge acide. Un lago è considerato NON Acido se $\text{pH} \geq 6$

Ecco il livello di Ph rilevati da due studiosi italiani in $n=15$ laghi alpini

| | | | | | |
|-----|-----|-----|-----|-----|-------------------|
| 5.5 | 5.7 | 5.8 | 6.1 | 6.3 | $\hat{\mu} = 6.6$ |
| 6.3 | 6.5 | 6.6 | 6.7 | 6.9 | |
| 6.9 | 7.2 | 7.3 | 7.3 | 7.9 | |

Cosa si può dire sui laghi della zona esaminata?

Ipotesi nulla $H_0: (\mu - 6) = 0$

Ipotesi alternativa $H_1: (\mu > 6)$

$$t_c = \sqrt{15} \left(\frac{6.6 - 6}{0.6719} \right) = 3.4586$$

$$P(t_c \geq 3.4586) = 1 - P(t_c \leq 3.4586) = 1 - [1 - P(t_c \leq 3.4586)] = P(t_c \leq 3.4586) = 0.0019$$

L'ipotesi può tranquillamente essere rifiutata con un errore probabile del 2 per mille

Esercizio

Offerte su base d'asta. La cifra di riferimento 28.7 milioni di euro costituisce la media delle proposte?

| | A | B | C | D | E | F |
|----|---------|---|--------------------------------------|------------|------------------------|---|
| 1 | offerte | | TEST SULLA MEDIA -VARIANZA INCOGNITA | | | |
| 2 | 26.56 | Ampiezza | | 21 | =Conta(A2:A22) | |
| 3 | 27.25 | Campionaria | | | | |
| 4 | 28.61 | | | | | |
| 5 | 25.73 | Media | | 26.77 | =Media(A2:A43) | |
| 6 | 28.43 | Campionaria | | | | |
| 7 | 26.82 | | | | | |
| 8 | 25.81 | Varianza nota | | 2.2989 | | |
| 9 | 26.63 | Popolazione | | | | |
| 10 | 27.80 | | | | | |
| 11 | 20.29 | $H_0: \mu =$ | | 28.7 | | |
| 12 | 25.01 | $H_1: \mu \neq 28.7$ | | | | |
| 13 | 28.10 | | | | | |
| 14 | 29.98 | Statistica | | -3.840127 | =Radq(D2)*(D5-D11)/D8 | |
| 15 | 29.42 | Test | | | | |
| 16 | 29.23 | | | | | |
| 17 | 29.64 | Gradi di libertà | | 20 | =D2-1 | |
| 18 | 26.45 | | | | | |
| 19 | 26.56 | P-value (totale) | | 0.00102216 | =DISTRIB.T(-D14;D17;2) | |
| 20 | 25.36 | | | | | |
| 21 | 23.04 | | | | | |
| 22 | 25.51 | Si deve rifiutare l'ipotesi le offerte abbiano 28.7 | | | | |
| 23 | | Come base di riferimento | | | | |

Test sulla proporzione

La percentuale di clienti di cui si finanzia il credito potrebbe sfuggire al controllo e deviare rispetto al 65% dello scorso anno.

L'ufficio fidi esegue un test al 10% e trova che, in un campione di $n=315$ sono stati affidati 214 clienti cioè $H=0.6794$

Ipotesi nulla $H_0: (\pi - 0.65) = 0$

Ipotesi alternativa $H_1: \pi \neq 0.65$

$$Z_c = \frac{0.6794 - 0.65}{\sqrt{\frac{0.6794(1 - 0.6794)}{315}}} = 1.118$$

$$P(Z_c \leq -1.118) = 0.1318$$

$$p\text{-value} = 2(0.1318) = 0.2636$$



Non si può rifiutare H_0

Esercizio

In un rapporto AUDITEL è riportato che le persone che guardano i film in televisione dopo mezzanotte sono equamente suddivise tra i due sessi.

In un campione casuale di 400 persone che guardano la televisione nelle ore piccole si sono trovati 180 donne.

- a) Si effettui un test su $H_0: \pi=50\%$ contro $H_1: \pi \neq 50\%$;
 b) Si effettui un test su $H_0: \pi=50\%$ contro $H_1: \pi < 50\%$

a) $Z_c = 2, p\text{-value} = 4.6\%$

b) $Z_c = -2, p\text{-value} = 2.3\%$



Esercizio

Un'indagine sulle intenzioni di voto di un campione di universitari ha prodotto i seguenti risultati. Qual è la probabilità che il "non voto" resti sotto i 2/3?

| | A | B | C | D | E | F | G |
|----|------------------|----------|--|------------|-----------------------|---|---|
| 1 | Posizione | | TEST SULLE PROPORZIONI -Uso della Normale | | | | |
| 2 | Voto | Non voto | Ampiezza | 50 | =Conta.Valori(A2:A51) | | |
| 3 | Non voto | Non voto | Campionaria | | | | |
| 4 | Non voto | Non voto | | | | | |
| 5 | Non voto | Non voto | Risposta | Non voto | =Media(A2:A43) | | |
| 6 | Voto | Non voto | Di interessa | | | | |
| 7 | Non voto | Voto | | | | | |
| 8 | Voto | Non voto | Conteggio | 32 | =CONTA.SE(D5) | | |
| 9 | Non voto | Voto | risposte | | | | |
| 10 | Non voto | Non voto | | | | | |
| 11 | Non voto | Non voto | Proporzione | 0.6400 | =D11/D2 | | |
| 12 | Voto | Voto | campionaria | | | | |
| 13 | Voto | Non voto | | | | | |
| 14 | Non voto | Non voto | Approssimazione | 0.0679 | =RADQ(D11*(1-D11)/D2) | | |
| 15 | Voto | Non voto | Dev.Standard | | | | |
| 16 | Non voto | Voto | | | | | |
| 17 | Non voto | Non voto | $H_0: \pi=0.66$ | 0.6600 | | | |
| 18 | Voto | Non voto | $H_1: \mu < 0.66$ | | | | |
| 19 | Non voto | Voto | | | | | |
| 20 | Non voto | Voto | Statistica | -0.2946278 | =(D11-D17)/D14 | | |
| 21 | Non voto | Voto | Test | | | | |
| 22 | Voto | Voto | | | | | |
| 23 | Non voto | Non voto | p-Value | 0.38413917 | =Distr.Norm.St(D20) | | |
| 24 | Non voto | Non voto | | | | | |
| 25 | Non voto | Voto | Non c'evidenza campionaria che la percentuale del "non voto" | | | | |
| 26 | Non voto | Non voto | si mantenga al di sotto dei 2/3. | | | | |

Test sulla differenza tra medie (Z-test cioè varianze note)

Indichiamo con D la differenza ipotizzata tra le due medie (spesso è zero) abbiamo

$$Z_c = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

La stessa procedura si applica se le ampiezze dei campioni sono entrambe molto grandi ed i campioni sono indipendenti

Basterà sostituire alle varianze incognite le loro stime campionarie

Esempio

La proprietà di un ristorante vuole sapere se la nuova campagna pubblicitaria ha aumentato gli incassi. Ecco i risultati per il periodo precedente la campagna e per il periodo seguente

| | Prima | Dopo |
|----------------|-------|-------|
| \hat{n} | 50 | 30 |
| $\hat{\mu}$ | 12.55 | 13.30 |
| $\hat{\sigma}$ | 215 | 238 |

Per verificare l'ipotesi che la campagna sia stata efficace si pone

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \quad Z_c = \frac{12.55 - 13.30}{\sqrt{\frac{215^2}{50} + \frac{238^2}{30}}} = -\frac{0.75}{53.03} = -1.41$$

Poiché il p-value del test è del 7.9% non possiamo rifiutare l'ipotesi H_0 dato che un errore dell'8% è troppo alto statisticamente.

A questo punto non c'è evidenza della efficacia della campagna

Esercizio

Un'indagine sulla spesa media di trasporti tra operai ed impiegati ha dato luogo ai seguenti risultati

La differenza nelle spese è significativa all'uno per mille

| | A | B | C | D | E | F |
|----|--------|-----------|---|---|-------------|-------------|
| 1 | | | | TEST SULLA DIFFERENZA TRA MEDIE - VARIANZE NOTE | | |
| 2 | Operai | Impiegati | | SRUMENTI-ANALISI DEI DATI- | | |
| 3 | 11.14 | 13.02 | | Test z: due campioni per medie | | |
| 4 | 16.51 | 19.77 | | | | |
| 5 | 15.07 | 17.96 | | Test z: due campioni per medie | | |
| 6 | 18.41 | 22.15 | | | | |
| 7 | 10.49 | 12.20 | | | Variabile 1 | Variabile 2 |
| 8 | 19.70 | 13.77 | | Media | 15.9503 | 17.8072 |
| 9 | 14.56 | 17.32 | | Varianza nota | 3.5 | 4.1 |
| 10 | 16.12 | 19.28 | | Osservazioni | 35 | 19 |
| 11 | 17.52 | 21.03 | | Differenza ipotizzata per le medie | 0 | |
| 12 | 21.60 | 22.67 | | z | -3.3044 | |
| 13 | 16.11 | 19.27 | | P(Z<=z) una coda | 0.0005 | |
| 14 | 17.44 | 15.91 | | z critico una coda | 1.6449 | |
| 15 | 16.88 | 20.23 | | P(Z<=z) due code | 0.0010 | |
| 16 | 12.80 | 15.11 | | z critico due code | 1.9600 | |
| 17 | 12.94 | 15.28 | | | | |
| 18 | 16.98 | 20.36 | | | | |
| 19 | 15.86 | 18.96 | | | | |
| 20 | 15.95 | 14.04 | | | | |
| 21 | 16.68 | 19.99 | | | | |
| 22 | 11.14 | | | | | |
| 23 | 18.78 | | | | | |
| 24 | 12.01 | | | | | |
| 25 | 16.68 | | | | | |
| 26 | 15.41 | | | | | |
| 27 | 14.44 | | | | | |
| 28 | 16.00 | | | | | |
| 29 | 14.94 | | | | | |
| 30 | 19.10 | | | | | |
| 31 | 15.85 | | | | | |
| 32 | 18.41 | | | | | |
| 33 | 14.81 | | | | | |
| 34 | 14.84 | | | | | |
| 35 | 15.99 | | | | | |
| 36 | 19.98 | | | | | |
| 37 | 17.11 | | | | | |

Test sulla differenza tra medie (t-test, varianze eguali)

$$t_c = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{S_A \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

S_A è lo scarto quadratico medio aggregato

$$S_A = \frac{(n_1 - 1)\sigma_1 + (n_2 - 1)\sigma_2}{(n_1 + n_2 - 2)}$$

Quando i campioni sono piccoli questo test è abbastanza valido perché non è molto sensibile anche a congrue differenze tra le varianze

Esempio

Una ditta produttrice di batterie vanta una durata media $\mu_1=75$ ore con SQM di 7 ore (valori ricavati da un campione casuale semplice di $n_1=20$ batterie)

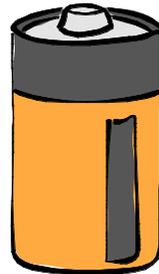
Un campione di $n_2=15$ batterie di una ditta concorrente ha dato luogo ad una media di $\mu_2=70$ ore con SQM di 5 ore.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

$$S_A = \sqrt{\frac{19 \cdot 7^2 + 14 \cdot 5^2}{20 + 15 - 2}} = \sqrt{\frac{1281}{33}} = 6.23 \quad t_c = \frac{(75 - 70) - 0}{6.23 \sqrt{\frac{1}{20} + \frac{1}{15}}} = \frac{5}{2.13} = 2.35$$

$$p\text{-value} = 1 - T_{33}(2.35) = 1.25\%$$

La ditta potrebbe avere qualche ragione a sostenere che le sue batterie durano più a lungo rispetto a quelle dei concorrenti



Esercizio

Tempi medi di completamento di un compito distinti per sesso

| | A | B | C | D | E | F | G |
|----|--------|---------|---|---|-------------|-------------|---|
| 1 | | | | TEST SULLA DIFFERENZA TRA MEDIE - PICCOLI CAMPIONI - VARIANZE INCOGNITE MA UGUALI | | | |
| 2 | | | | | | | |
| 3 | Maschi | Femmine | | | | | |
| 4 | 118.38 | 118.38 | | Test t: due campioni assumendo uguale varianza | | | |
| 5 | 118.38 | 105.74 | | | | | |
| 6 | 105.74 | 109.35 | | | Variabile 1 | Variabile 2 | |
| 7 | 109.35 | 117.19 | | Media | 119.3386 | 120.3630 | |
| 8 | 117.19 | 113.26 | | Varianza | 66.8361 | 65.5155 | |
| 9 | 113.26 | 126.59 | | Osservazioni | 15 | 19 | |
| 10 | 126.59 | 125.65 | | Varianza complessiva | 66.0933 | | |
| 11 | 125.65 | 107.41 | | Differenza ipotizzata per le gdl | 0.0000 | | |
| 12 | 107.41 | 121.63 | | Stat t | 32.0000 | | |
| 13 | 121.63 | 124.45 | | P(T<=t) una coda | -0.3648 | | |
| 14 | 124.45 | 123.42 | | t critico una coda | 0.3588 | | |
| 15 | 123.42 | 121.52 | | P(T<=t) due code | 1.6939 | | |
| 16 | 121.52 | 119.90 | | t critico due code | 0.7176 | | |
| 17 | 119.90 | 137.22 | | | 2.0369 | | |
| 18 | 137.22 | 119.27 | | | | | |
| 19 | | 120.52 | | | | | |
| 20 | | 117.90 | | | | | |
| 21 | | 135.22 | | | | | |
| 22 | | 122.27 | | | | | |
| 23 | | | | | | | |

Non ci sono evidenze convincenti che il sesso di chi esegue il compito faccia la differenza nei tempi di completamento

Test sulla differenza tra medie (t-test, varianze diverse)

Indiviamo con D la differenza ipotizzata tra le due medie (spesso è zero) abbiamo

$$t_c = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

I gradi di libertà sono approssimati dalla formula

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} \right)^2}{\left(\frac{\hat{\sigma}_1^2}{n_1} \right)^2 + \left(\frac{\hat{\sigma}_2^2}{n_2} \right)^2}$$

Esempio

Confronto di due processi produttivi;

1° tipo. Su di un campione di 13 pezzi, ha funzionato in media 202 ore con un SQM di 9 ore.

2° tipo. Su di un campione di 10 pezzi, è durato in media 195 ore con SQM di 6 ore.

Verifichiamo il p-value dell'ipotesi di uguaglianza tra le durate in un test bilaterale



Ipotesi nulla $H_0: D = 0$

Ipotesi alternativa: $H_1: D \neq 0$

$$t_c = \frac{(202 - 195) - 0}{\sqrt{\frac{81}{13} + \frac{36}{10}}} = 2.23$$

P-value: 3.7%

$$df = \frac{\left(\frac{81}{13} + \frac{36}{10} \right)^2}{\left(\frac{81}{13} \right)^2 + \left(\frac{36}{10} \right)^2} = \frac{[96.44]}{[4.675]} = [20.629] = 20$$

Si può rifiutare H_0 , ma non sarebbe sbagliato cercare più dati

Esercizio

Confronto dei guadagni iniziali tra Laureati non laureati.

Lo scarto di 300 euro non risulta significativo

| | A | B | C | D | E | F |
|----|---|--------------|---|---|-------------|-------------|
| 1 | TEST T- DUE CAMPIONI ASSUMENDO VARIANZE DIVERSE | | | | | |
| 2 | | | | | | |
| 3 | Laureati | Non laureati | | Test t: due campioni assumendo varianze diverse | | |
| 4 | 2123.77 | 1885.65 | | | | |
| 5 | 1720.96 | 1281.44 | | | | |
| 6 | 1924.34 | 1586.50 | | Variabile 1 | Variabile 2 | |
| 7 | 1670.92 | 1206.38 | | Media | 1783.5170 | 1356.0815 |
| 8 | 1559.65 | 1039.48 | | Varianza | 55948.8161 | 140122.4605 |
| 9 | 1386.49 | 779.73 | | Osservazioni | 36 | 25 |
| 10 | 1598.77 | 1098.16 | | Differenza ipotizzata per | 300 | |
| 11 | 1969.72 | 1654.58 | | gdl | 37 | |
| 12 | 1616.39 | 1124.58 | | Stat t | 1.50613 | |
| 13 | 1516.71 | 975.07 | | P(T<=t) una coda | 0.07026 | |
| 14 | 1758.07 | 1337.10 | | t critico una coda | 1.68709 | |
| 15 | 1664.51 | 1196.76 | | P(T<=t) due code | 0.14052 | |
| 16 | 1639.49 | 1159.24 | | t critico due code | 2.02619 | |
| 17 | 1878.81 | 1518.21 | | | | |
| 18 | 1798.33 | 1397.49 | | | | |
| 19 | 2076.27 | 1814.40 | | | | |
| 20 | 1695.70 | 1243.55 | | | | |
| 21 | 1777.92 | 1366.88 | | | | |
| 22 | 2443.37 | 2365.05 | | | | |
| 23 | 2002.36 | 1703.54 | | | | |
| 24 | 2155.88 | 1933.81 | | | | |
| 25 | 1481.22 | 921.83 | | | | |
| 26 | 1531.60 | 997.40 | | | | |
| 27 | 1555.76 | 1033.63 | | | | |
| 28 | 1721.02 | 1281.54 | | | | |
| 29 | 2011.78 | | | | | |
| 30 | 2120.05 | | | | | |
| 31 | 1691.47 | | | | | |
| 32 | 1815.58 | | | | | |
| 33 | 1904.83 | | | | | |
| 34 | 1517.31 | | | | | |
| 35 | 1927.16 | | | | | |
| 36 | 1913.04 | | | | | |
| 37 | 1894.05 | | | | | |
| 38 | 1744.33 | | | | | |
| 39 | 1398.98 | | | | | |

Rilevazioni abbinate

Si è già detto che è più facile incontrare il caso di due variabili osservate sulle unità di uno stesso campione (*matched pairs*) che un campione per ogni variabile. Questo però genera dipendenza e correlazione.

La statistica test per verificare la differenza tra medie è ora:

$$D = \frac{\hat{d} - d_0}{\hat{\sigma}_d / \sqrt{n}}; \quad \text{dove} \begin{cases} \hat{d} = \hat{\mu}_1 - \hat{\mu}_2 \\ d_0 = \text{differenza sotto } H_0 (\text{di solito zero}) \\ \hat{\sigma}_d^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2Cov(X_1, X_2) \end{cases}$$

usando la t-Student con (n-1) gradi di libertà

Per il calcolo della deviazione standard si possono usare direttamente le differenze tra i valori osservati delle due serie:

Esempio

La presenza di metalli nell'acqua ne influenza il sapore è -se elevata- può causare danni alla salute.

La concentrazioni di zingò (Mg/L) è stata rilevata sulla superficie e sul fondo di una sorgente in n=6 punti diversi

| punto | 1 | 2 | 3 | 4 | 5 | 6 | $\hat{\mu}$ | $\hat{\sigma}$ |
|---------|-------|-------|-------|-------|-------|-------|-------------|----------------|
| Fondo | 0.430 | 0.266 | 0.567 | 0.531 | 0.707 | 0.716 | 0.5362 | 0.00587 |
| Superf. | 0.415 | 0.238 | 0.390 | 0.410 | 0.605 | 0.609 | 0.4445 | 0.00402 |

$$D = \frac{0.0917}{\frac{0.06069}{\sqrt{6}}} = 3.701 \quad \text{P-value} = \text{TDIST}(A1,5,2) = 0.01398488$$

C'è differenza, ma non sembra poi tanto evidente

Test della differenza tra proporzioni

X_1 è il numero di unità speciali rispetto alla variabile dicotoma Y rilevata su n_1 unità della popolazione 1

X_2 è il numero di unità speciali rispetto alla variabile dicotoma W (che può essere la stessa Y) rilevata su n_2 unità della popolazione 2 (che può essere la stessa della 1).

Sia inoltre:

$n_1 \pi_1 \geq 5$ e $n_2 \pi_2 \geq 5$ (in modo che i campioni possano considerarsi "grandi")

La selezione delle unità del secondo campione avviene in modo del tutto indipendente dalla selezione delle unità del primo campione.

Test della differenza tra proporzioni ($H_0: \pi_1 - \pi_2 \neq 0$)

e la verifica di ipotesi si basa sulla statistica:

$$Z_c = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \pi_0}{\sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}} \quad \text{con} \quad \begin{cases} \hat{\pi}_1 = x_1/n_1 \\ \hat{\pi}_2 = x_2/n_2 \end{cases}$$

Ed il p-value si ricava dalla normale standardizzata

Esempio

La dott.ssa Velia Zupo sostiene che l'inserimento di specifiche immagini sulla home page dell'azienda aumenterà di almeno il 10% i contatti operativi con il sito compensando i maggiori costi. Detto fatto.

Nei tre mesi precedenti i contatti operativi furono 188 su 376 (0.50), nei tre mesi successivi 260 su 400. Verifichiamo l'ipotesi della Zupo.

$$\begin{cases} H_0: (\pi_1 - \pi_2) = 0.1 \\ H_1: (\pi_1 - \pi_2) > 0.1 \end{cases} \quad Z_c = \frac{(0.65 - 0.50) - 0.1}{\sqrt{\frac{0.65 * 0.35}{400} + \frac{0.50 * 0.50}{376}}} = 1.42$$

$$\text{P-value} = 1 - \text{NORMSDIST}(1.42) = 0.07780389$$

Non ci sono sufficienti evidenze che la Zupo abbia ragione.
Trasferita in portineria!

Test della differenza tra proporzioni ($H_0: \pi_1 - \pi_2 = 0$)

E' la situazione più diffusa. Se è vera l'ipotesi non ci sono DUE popolazioni, ma solo UNA e la stima della probabilità di successo è:

$$\hat{\pi} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1}{n_1 + n_2} \hat{\pi}_1 + \frac{n_2}{n_1 + n_2} \hat{\pi}_2$$

La verifica di ipotesi si basa sulla statistica:

$$Z_c = \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

che tende alla normalità.

Esempio

Una delle applicazioni più eclatanti della statistica avvenne nel 1954 con il vaccino di Salk per la poliomelite.

Il vaccino fu somministrato ad un campione di $n_2=200745$ bambini trovando $x_2=33$ casi di polio e ad un altro campione di $n_1=201229$ bambini venne dato un placebo nella versione DOUBLE BLIND con $X_1=110$ casi.

$$\begin{cases} H_0 : \pi_1 = \pi_2 \\ H_1 : \pi_1 > \pi_2 \end{cases} \quad Z_c = \frac{0.00038226}{\sqrt{0.00035574 * \left(\frac{1}{201229} + \frac{1}{200745}\right)}} = 6.425$$

$$P\text{-value} = 1 - \text{NORMSDIST}(6.425) = 6.62455E-11$$

La somministrazione del vaccino ha dimezzato i casi di poliomelite. Per caso questo sarebbe potuto succedere solo una volta su 50 miliardi

P-Value per la correlazione

Una volta calcolato $r(x,y)$ su di un campione di osservazioni cosa si può dire sul coefficiente di correlazione che lega le due variabili nell'intera popolazione?

$H_0 : \rho = 0$ (assenza di un legame lineare,

$H_1 : \rho \neq 0$ (presenza di un legame lineare

La statistica test che si usa è
$$t_c = \sqrt{n-2} \left(\frac{r(y,x)}{\sqrt{1 - [r(y,x)]^2}} \right)$$

La cui distribuzione è ben approssimata dalla t-Student con (n-2) gradi di libertà

p-Value per il rho di Spearman

Una volta calcolato rho su di un campione, cosa si può dire sul rho di Spearman che lega le variabili nell'intera popolazione?

$H_0 : \rho = 0$ (NON esiste una relazione monotona)

$H_1 : \rho \neq 0$ (Esiste una relazione monotona)

La statistica test che si puo' usare è la stessa del coefficiente di correlazione

$$t_c = \sqrt{n-2} \left(\frac{\rho}{\sqrt{1-\rho^2}} \right)$$

Per $n < 11$ tale formula deve essere usata con molta prudenza

La cui distribuzione è approssimata dalla t-Student con (n-2) gradi di libertà

p-Value per il tau di Kendall

$H_0: \tau = 0$ (NON esiste una relazione monotona)
 $H_1: \tau \neq 0$ (Esiste una relazione monotona)

La statistica test che si può usare è

$$t_c = \frac{3t\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

Per $n < 11$ tale formula deve essere usata con molta prudenza

La cui distribuzione è approssimata dalla Normale standardizzata

Critica all'uso dei test

L'uso dei test statistici è molto diffuso, ma essi non sono immuni da critiche:



VIOLAZIONI DELLE CONDIZIONI DI BASE

Molte condizioni (indipendenza nel campione, normalità, "n" grande) non sono verificate.

Spesso le procedure sono ROBUSTE cioè non si alterano di molto se certe condizioni vengono meno. Ma fino a che punto si può arrivare?



INTERPRETAZIONE RIGIDA DEL CRITERIO DI "ACCETTAZIONE"

Rifiutare un'ipotesi a meno di prove contrarie è chiaro. Ma "accettare" è pure chiaro? E' sicuro che l'insufficienza delle prove porti al pieno accoglimento della H_0 ? No, ovviamente. Spesso si sostituisce la locuzione NON RIFIUTO.



Piani di campionamento TROPPO SEMPLICI

Test del rapporto di verosimiglianza

Questo tipo di test poggia su un risultato valido per grandi campioni e per un numero elevato di categorie di righe e di colonne.

Il p-value è ottenuto dalla distribuzione del chi-quadrato con $(r-1)(c-1)$ gradi di libertà.

Ad esempio

$$\text{CHIDIST}(37.5617;4)=0,00000014$$

Che induce a rifiutare l'ipotesi di assenza di legame. Il legame esiste ed è abbastanza dal punto di vista statistico:

Con questa asserzione si rischia di sbagliare 7 volte su 5 milioni

Test del χ^2 (chi quadrato)

Questa metodologia è richiamata in tre distinte occasioni



Adattamento delle frequenze osservate ad un modello



Omogeneità di due o più campioni



Indipendenza in una distribuzione doppia

L'analisi di dati categoriali

Ogni osservazione di un campione di ampiezza "n" è classificata in una ed una sola di "k" categorie fisse ed invariabili.

| Categ. | fr.ass. | fr.rel. |
|----------|----------|----------|
| X_1 | n_1 | f_1 |
| X_2 | n_2 | f_2 |
| \vdots | \vdots | \vdots |
| X_k | n_k | f_k |
| | n | 1 |

Siamo convinti che dietro queste frequenze ci sia un meccanismo che determina il verificarsi di una modalità piuttosto che un'altra

L'analisi di dati categoriali/2

Supponiamo che l'acquisizione dei dati avvenga in forma di prove indipendenti svolte nelle stesse condizioni.

Ricorrono allora le condizioni del modello multinomiale e la probabilità della configurazione ottenuta è

| Categ. | $P(X = X_i)$ |
|----------|--------------|
| X_1 | π_1 |
| X_2 | π_2 |
| \vdots | \vdots |
| X_k | π_k |
| | 1 |

$$P(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \dots n_k!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_k^{n_k}$$

Note le π_i sarà possibile calcolare la probabilità di qualsiasi allocazione del campione tra le varie categorie e poi decidere sull'entità della differenza

Esempio

Classificazione di un gruppo di n=20 persone per gruppo sanguigno:

| Categ. | fr.ass. | fr.rel. |
|--------|---------|---------|
| 0 | 7 | 0.35 |
| A | 4 | 0.20 |
| B | 3 | 0.15 |
| AB | 6 | 0.30 |
| | 20 | 1.00 |

Supponiamo che i gruppi siano equiprobabili.

$$\begin{cases} H_0: \pi_i = 0.25 \text{ per ogni } i \\ H_1: \pi_i \neq 0.25 \text{ per almeno un } i \end{cases}$$

Alla luce della classificazione prodottasi nel campione, c'è sufficiente evidenza di tale meccanismo?

Sotto H_0 la probabilità di osservare la configurazione è

$$P(7, 4, 3, 6) = \frac{20!}{7! 4! 3! 6!} 0.25^{20} = 0.0042$$

Al momento non riusciamo a dire se è bassa o è alta dato che le possibili alternative sono moltissime.

il test del χ^2

il grande pregio della multinomiale è che la statistica

$$\chi_c^2 = n \left[\sum_{i=1}^k \frac{(f_i - \pi_i)^2}{\pi_i} \right] = \sum_{i=1}^k \frac{(n_i - n\pi_i)^2}{n\pi_i} = \left[\sum_{i=1}^k \frac{n_i^2}{n\pi_i} \right] - n$$

ha una distribuzione che, per "n" abbastanza grande, è ben approssimata dalla $\chi^2(k-1)$ (è richiesto che $n \cdot \pi_i \geq 5$ per ogni "i")

Poiché χ_c^2 è la somma ponderata di scarti al quadrato sarà nullo se e solo se lo sono tutti gli addendi (perfetto adattamento).

Ogni discrepanza tra teoriche ed osservate aumenta il χ_c^2 (e quindi riduce il p-value) allontanandoci da H_0 per avvicinarci ad H_1

Esempio (continua)

$$\chi_c^2 = \frac{(7-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(6-5)^2}{5} = \frac{10}{5} = 2; \quad gdl = (4-1) = 3$$

Per i gradi di libertà si tiene conto che frequenze teoriche ed osservate coincidono e che la somma la loro somma deve essere pari ad uno (quindi solo (k-1) sono libere di variare)

P-Value DISTRIB. CHI(2;3)=57.2%

Non c'è quindi evidenza di una distribuzione NON uniforme dei pazienti per gruppo sanguigno



Test sull'adattamento

Poiché la decisione verrà presa in base all'adattamento delle frequenze a modelli precostituiti si parla di test di adattamento (*Goodness-of-fit Test*).

Si vuole confermare o sconfutare l'ipotesi che le frequenze nel campione scaturiscano da una particolare distribuzione:

$$\begin{cases} H_0 : \pi_i = \pi_i^0, & i = 1, 2, \dots, k \\ H_1 : \pi_i \neq \pi_i^0, & \text{per almeno un } i \end{cases}$$

Si adopera la seguente statistica test:

$$\chi_c^2 = n \left[\sum_{i=1}^k \frac{(f_i - \pi_i^0)^2}{\pi_i^0} \right] = \left[\sum_{i=1}^n \frac{n_i^2}{n\pi_i^0} \right] - n$$

L'esperimento di Mendel

La teoria di Mendel nacque dall'incrocio di piselli gialli lisci e piselli verdi corrugati: gli ibridi presenteranno i due caratteri in proporzioni determinate.

$$H_0: G.L. = \frac{9}{16}; \quad V.L. = \frac{3}{16}; \quad G.C. = \frac{3}{16}; \quad V.C. = \frac{1}{16}$$

i dati su cui si basò Mendel furono

| Categoria | frequenze |
|------------------|-----------|
| Gialli lisci | 315 |
| Verdi lisci | 108 |
| Gialli corrugati | 101 |
| Verdi corrugati | 32 |
| | 556 |

$$\chi_c^2 = \frac{(315-312.75)^2}{312.75} + \frac{(108-106.25)^2}{106.25} + \frac{(101-104.25)^2}{312.75} + \frac{(32-34.75)^2}{34.75} = 0.47$$

il P-value del test è circa del 95% per cui non possiamo rifiutare H_0 nemmeno a questi livelli "assurdi" di significatività

C'è il sospetto che i dati siano troppo buoni per essere veri

Esercizio

Un cliente può scegliere una particolare marca di caffè fra le tre in vendita su uno scaffale sotto osservazione. I risultati su 300 acquisti sono i seguenti

| Marca | n_i |
|----------|-------|
| Old moka | 78 |
| Braz | 117 |
| Arab mix | 105 |
| | 300 |

Se i consumatori fossero **INDIFFERENTI** rispetto alle marche la probabilità di sceglierne una qualsiasi sarebbe 1/3.

L'indagine sottopone a verifica la seguente ipotesi

$$\begin{cases} H_0 : \pi_1 = \pi_2 = \pi_3 \\ H_1 : \pi_i \neq \pi_j \text{ per almeno un } i \neq j \end{cases} \quad \text{Sotto } H_0 \text{ dovremmo trovare le frequenze assolute: } 100, 100, 100.$$

RAGIONAMENTO: Fino a che punto gli scarti tra osservate e teoriche sono da attribuire alle fluttuazioni campionarie?

CALCOLO:

Poiché si opera sulle assolute riscriviamo la formula in base a queste

$$\chi_c^2 = \sum_{i=1}^k \frac{d_i^2}{n \cdot \pi_i^0}; \quad \text{con } d_i = n_i - n \cdot \pi_i^0$$

| i | n_i | $n \cdot \pi_i^0$ | d_i | d_i^2 | $\frac{d_i^2}{n \cdot \pi_i^0}$ |
|---|-------|-------------------|-------|---------|---------------------------------|
| 1 | 78 | 100 | -22 | 484 | 4.84 |
| 2 | 117 | 100 | 17 | 289 | 2.89 |
| 3 | 105 | 100 | 5 | 25 | 0.25 |
| | 300 | 300 | 0 | | $\chi_c^2 = 7.98$ |

P-value: 0.01849971

Frequenze teoriche

Se il modello è completamente specificato (ipotesi semplice) si ottengono dalle funzioni di distribuzioni o di densità

CASO DISCRETO $P(X_i = x_i) = \pi_i \quad i = 1, 2, \dots, k$

Esempio: La Poisson

H_0 : La distribuzione teorica è la Poisson con media 2 $P(X_i = x_i) = \frac{2^{x_i} e^{-2}}{X_i!}$

CASO CONTINUO $P(L_i \leq x \leq U_i) = \int_{L_i}^{U_i} f(x) dx$

Esempio: La Normale

H_0 : La distribuzione teorica è la $N(\mu = 5, \sigma = 2)$

$$P(L_i \leq x \leq U_i) = \int_{L_i}^{U_i} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \Phi\left(\frac{U_i - 5}{2}\right) - \Phi\left(\frac{L_i - 5}{2}\right) = \pi_i$$

Da notare che occorre si abbia sempre: $\pi_i \geq 0; \sum_{i=1}^k \pi_i = 1$

Esempio

Gli studenti di informatica affrontano questioni relative ai linguaggi: 1) BASIC, 2)FORTRAN, 3)C++, 4)PASCAL.

Il tutoraggio è organizzato nel presupposto che B=40%, F=25%, C=25%, P=10%

I dati raccolti hanno fatto riscontrare:

| Linguaggi | B | F | C | P |
|-----------|----|----|----|---|
| Frequenza | 52 | 38 | 21 | 9 |

Si accetta o si rifiuta l'ipotesi che l'assistenza sia ben organizzata?

$$\begin{cases} H_0: \pi_1 = 0.40, \pi_2 = 0.25, \pi_3 = 0.25, \pi_4 = 0.10 \\ H_1: \pi_1 \neq 0.40 \vee \pi_2 \neq 0.25 \vee \pi_3 \neq 0.25 \vee \pi_4 \neq 0.10 \end{cases}$$

$$\begin{aligned} \chi_c^2 &= \left(\sum_{i=1}^k \frac{n_i^2}{n\pi_i} \right) - n = \left(\frac{52^2}{120 \cdot 0.40} + \frac{38^2}{120 \cdot 0.25} + \frac{21^2}{120 \cdot 0.25} + \frac{9^2}{120 \cdot 0.1} \right) - 120 \\ &= (56.3333 + 48.1333 + 14.7000 + 6.7500) - 120 = 5.9166 \end{aligned}$$

P-value= 0.116

Non c'è evidenza sufficiente per ritenere che ci sia cattiva organizzazione

Presenza di parametri incogniti

Spesso è noto solo il tipo di modello, ma si ignora uno o più parametri

I parametri debbono essere stimati dal campione

Ci sono però conseguenze sul test χ^2 dato che l'uso duplice del campione "truca" l'adattamento facendolo sembrare migliore di quanto non sia.

Ne consegue che il p-value debba essere operato con meno gradi di libertà

Infatti, si dimostra che, se "n" è abbastanza grande, se si stimano "m" parametri il p-value adatto deriva dalla distribuzione chi-quadro con (k-1-m) gradi di libertà

Esempio

il numero di clienti in arrivo ogni ora in un ufficio postale, rilevato sulla base di un campione casuale di 100 ore, è stato il seguente:

| Clients | osservate n. | Sottoporre a verifica la seguente ipotesi | H_0 : Il modello è Poissoniano |
|----------|-----------------|--|----------------------------------|
| 0 | 38 | Usare il livello di significatività $\alpha=1\%$ | H_1 : non lo è |
| 1 | 36 | 1) Stimare la media: $\hat{\lambda} = \frac{0 \cdot 38 + 1 \cdot 36 + 2 \cdot 14 + 3 \cdot 7 + 4 \cdot 3 + 5 \cdot 2 + 6 \cdot 0}{100} = 1.07$ | |
| 2 | 14 | 2) Calcolo delle teoriche. Per comodità usiamo $\lambda=1$ | |
| 3 | 7 | | |
| 4 | 3 | | |
| 5 | 2 | | |
| ≥ 6 | 0 | | |
| | 100 | | |

$$P(X = x) = \frac{e^{-1} 1^x}{x!} = \frac{1}{e \cdot x!}$$

$$\chi_c^2 = \frac{(38 - 36.8)^2}{36.8} + \frac{(36 - 36.8)^2}{36.8} + \dots + \frac{(0 - 0.1)^2}{0.1} = 3.07$$

$$g = 7 - 1 - 1 = 5$$

| x | $\frac{100}{e \cdot x!}$ |
|----------|--------------------------|
| 0 | 36.8 |
| 1 | 36.8 |
| 2 | 18.4 |
| 3 | 6.1 |
| 4 | 1.5 |
| 5 | 0.3 |
| ≥ 6 | 0.1 |
| | 100 |

P-value=68.9%. Non si può rifiutare

Esercizio

Una lampada è venduta con accluse 4 batterie. Si esamina un campione di n=150 item riscontrando il numero di batterie difettose.

| | | | | | | |
|-----------|----|----|----|----|----|-----|
| Difetti | 0 | 1 | 2 | 3 | 4 | |
| Frequenza | 26 | 51 | 47 | 16 | 10 | 150 |

Si sottoponga a verifica l'ipotesi seguente:

$$\begin{cases} H_0: \pi_i = \binom{4}{i} p^i (1-p)^{4-i} & \text{per } i=0,1,2,\dots,4 \\ H_1: \pi_i \neq \binom{4}{i} p^i (1-p)^{4-i} \end{cases}$$

Possiamo stimare "p" come numero di difettose sul totale di batterie c:

$$p = \frac{51 + 94 + 48 + 40}{600} = 0.3883$$

| | |
|---|---------|
| 0 | 21.0013 |
| 1 | 53.3254 |
| 2 | 50.7755 |
| 3 | 21.4878 |
| 4 | 3.4101 |

$$\chi_c^2 = 15.7086$$

P-value=0.00130174. L'ipotesi è da rifiutare

Test dell'omogeneità dei campioni

il test del χ^2 può essere utilizzato anche nella seguente situazione:

Si osservano due (o più) classificazioni campionarie.

Le ampiezze possono essere diverse, ma le categorie sono identiche.

Ci si chiede se i campioni provengono dalla stessa popolazione.

Ad esempio per due campioni abbiamo

$$H_0: \pi_{1i} = \pi_{2i} \text{ per ogni "i"}$$

$$H_1: \pi_{1i} \neq \pi_{2i} \text{ per almeno un "i"}$$



La statistica test assume ora la forma

$$\chi_c^2 = \sum_{i=1}^k \frac{(\pi_{i1} - \pi_{i2})^2}{\pi_{i2}}$$

Esempio

Due campioni di misurazioni del tempo precedente la prima disfunzione in due diversi macchinari hanno dato luogo alla tabella

| Tempi | C_1 | C_2 | C_2/risc | (f1i-f'2i)/f'2i |
|---------|-----|-----|----------|-----------------|
| 0 - 20 | 29 | 56 | 32.78 | 0.436 |
| 20 - 40 | 27 | 41 | 24.00 | 0.375 |
| 40 - 60 | 14 | 32 | 18.73 | 1.195 |
| 60 - 70 | 11 | 19 | 11.12 | 0.001 |
| 70 - 80 | 8 | 13 | 7.61 | 0.020 |
| >80 | 7 | 3 | 1.76 | 15.659 |
| | 96 | 164 | 96.00 | 17.686 |

Le "teoriche" sono ottenute come: $n_{2i} = n_1 f_{2i}$

La statistica test è $\chi^2 = 17.686$ con 6-1=5 gradi di libertà

$$DISTRIB. CHI(17.686;5)=0.00336 \text{ (p-Value)}$$

L'ipotesi si deve rifiutare aldilà di ogni ragionevole dubbio

Esercizio

Numero di incidenti automobilistici lungo una strada a scorrimento veloce. Frequenza per lato di percorrenza

| Verso Sud | Settimane | Verso Nord | Settimane |
|-----------|-----------|------------|-----------|
| 0 | 39 | 0 | 82 |
| 1 | 28 | 1 | 60 |
| 2 | 19 | 2 | 40 |
| 3 | 13 | 3 | 27 |
| 4 | 8 | 4 | 18 |
| 5 | 4 | 5 | 8 |
| >5 | 2 | >5 | 4 |
| | 113 | | 239 |

Determinate il p-value dell'ipotesi di omogeneità dei due campioni

Test sull'indipendenza

Un problema ricorrente in statistica riguarda l'indipendenza o meno di due categorie di classificazione

L'ipotesi da sottoporre a verifica è

$$H_0: \pi_{ij} = \pi_i \pi_j \text{ per ogni "i e j"}$$

$$H_1: \pi_{ij} \neq \pi_i \pi_j \text{ per almeno un "i e j"}$$

Si adopera la seguente statistica test:

$$\chi_c^2 = n \left[\sum_{i=1}^k \frac{(f_{ij} - \pi_{ij})^2}{\pi_{ij}} \right] = \sum_{i=1}^n \frac{(n_{ij} - n\pi_{ij})^2}{n\pi_{ij}} = \left[\sum_{i=1}^n \frac{n_{ij}^2}{n\pi_{ij}} \right] - n$$

Esempio

Uno studio su di un campione di mariti ha prodotto i seguenti risultati

| | La moglie lavora | | | | | La moglie lavora | | | |
|-----------------------|------------------|-----------|-----------|--------|-----------------------|------------------|-----------|-----------|--------|
| | Full-time | Part-Time | Casalinga | Totale | | Full-time | Part-Time | Casalinga | Totale |
| Accetto che lavori | 148 | 51 | 49 | 248 | Accetto che lavori | 95.5055 | 68.3867 | 84.1078 | 248 |
| Non mi importa | 34 | 16 | 7 | 57 | Non mi importa | 21.9509 | 15.7179 | 19.3312 | 57 |
| Non so/Non risponde | 36 | 26 | 32 | 94 | Non so/Non risponde | 36.1997 | 25.9208 | 31.8796 | 94 |
| Non voglio che lavori | 25 | 81 | 126 | 232 | Non voglio che lavori | 89.3439 | 63.9746 | 78.6815 | 232 |
| | 243 | 174 | 214 | 631 | | 243 | 174 | 214 | 631 |

| | La moglie lavora | | | p-value= |
|-----------------------|------------------|-----------|-----------|----------|
| | Full-time | Part-Time | Casalinga | |
| Accetto che lavori | 28.8535 | 4.4204 | 14.6545 | 1.2E-29 |
| Non mi importa | 6.6139 | 0.0051 | 7.8660 | |
| Non so/Non risponde | 0.0011 | 0.0002 | 0.0005 | |
| Non voglio che lavori | 46.3393 | 4.5309 | 28.4571 | |
| | 81.8079 | 8.9566 | 50.9780 | 141.7 |

C'è fortissima evidenza **CONTRO** l'ipotesi di indipendenza come del resto era facile da immaginare guardando la tabella

Considerazioni sul χ^2

Il test del χ^2 è del tipo DISTRIBUTION FREE cioè la sua impostazione non dipende da una particolare distribuzione.

In questo senso è anche detto **NON PARAMETRICO**.

E' un test di grande generalità e può essere applicato in una gamma molto grande di situazioni.

Ma proprio questa è la sua debolezza: perchè sia veramente informativo, il rifiuto dell'ipotesi nulla deve avvenire a livelli molto elevati di significatività (ovvero valori molto bassi del p-value)

Modello lineare stocastico

La variabile casuale Y ha distribuzione gaussiana con media che dipende linearmente dal livello dell'esogena X

$$M(Y|X = x_i) = \beta_0 + \beta_1 x_i = y_i \quad i = 1, 2, \dots, n$$

La X può sia essere una variabile casuale che un blocco di costanti fisso ed invariabile

Si eseguono "n" osservazioni indipendenti di coppie di valori (y,x)

Il modello lineare stocastico parte da

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, 2, \dots, n$$

Test nel modello di regressione

Per espletare l'inferenza nel modello di regressione è necessaria una delle due ipotesi seguenti

☰ "n" è abbastanza grande da attivare il teorema limite centrale

☰ $e_i \sim N(0, \sigma^2)$, $i = 1, 2, \dots, n$ gli errori sono normali (e quindi indipendenti oltreché incorrelati)

Queste condizioni permettono di stabilire quale sia la distribuzione degli stimatori in quanto funzioni di variabili casuali.

In particolare, si dimostra che $\hat{\beta}_0, \hat{\beta}_1$ sono stimatori di massima verosimiglianza con annesse proprietà e difetti.

Normalità

L'ipotesi di normalità implica

$$y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2); \quad i = 1, 2, \dots, n$$

Ed anche

$$\hat{\beta}_0 \sim N\left[\beta_0, \frac{\sigma^2}{n} \left(1 + \frac{n\bar{x}^2}{S_{xx}}\right)\right]; \quad \hat{\beta}_1 \sim N\left[\beta_1, \frac{\sigma^2}{S_{xx}}\right]$$

Inoltre,

$$\frac{(n-2)s_e^2}{\sigma^2} \sim \chi^2(n-2)$$

L'ipotesi di normalità (diretta delle y oppure ottenuta grazie ad "n" grande ed al conseguente limite centrale) è indispensabile per la validità delle procedure inferenziali relativamente ai parametri del modello di regressione.

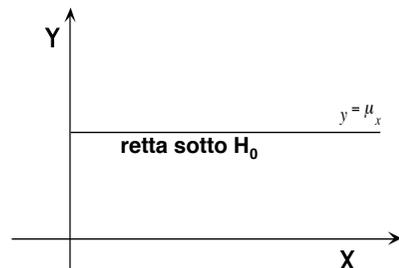
Test sul coefficiente angolare

ESISTE O NON ESISTE UNA RELAZIONE TRA ENDOGENA ED ESOGENA?

A questa domanda rispondiamo solo PARZIALMENTE: la "Y" varia o non varia LINEARMENTE con la "X"?

Questo si traduce nella verifica di ipotesi: $\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$

Se H_0 non potesse essere rifiutata la retta di regressione si presenterebbe come parallela all'asse delle X e non in grado di spiegare nulla della "Y"



Continua test su β_1

La statistica test necessaria per la verifica si ottiene come per i test sulla media. In particolare si coinvolge la media e la varianza dello stimatore

$$E(\hat{\beta}_1) = \beta_1, \quad Var(\hat{\beta}_1) = \frac{s_e^2}{S_{xx}}$$

$$t_{n-2}(\beta_1) = \frac{\hat{\beta}_1}{\frac{s_e}{\sqrt{S_{xx}}}} \quad \text{dove} \quad \begin{cases} s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} \\ S_{xx} = \sum_{i=1}^n (x_i - \mu_x)^2 \end{cases}$$

Il p-value del test si ottiene applicando la distribuzione t-Student con (n-2) gradi di libertà

Esempio

Si ritiene che il consumo di energia elettrica sia determinato da un modello lineare nella temperatura media del giorno (in gradi F)

SUMMARY OUTPUT

| Regression Statistics | |
|-----------------------|------------|
| Multiple R | 0.95084638 |
| R Square | 0.9041 |
| Adjusted R Sc | 0.8921 |
| Standard Error | 15.1954 |
| Observations | 10 |

ANOVA

| | df | SS | MS | F | Significance F |
|------------|----|-------------|-------------|---------|----------------|
| Regression | 1 | 17'416.3911 | 17'416.3911 | 75.4279 | 0.0000 |
| Residual | 8 | 1'847.2089 | 230.9011 | | |
| Total | 9 | 19'263.6000 | | | |

| | Coefficients | Standard Error | t Stat | P-value |
|-----------|--------------|----------------|---------|----------|
| β_0 | -395.5861 | 68.1479 | -5.8048 | 0.000403 |
| Temp.Med. | 6.4735 | 0.7454 | 8.6849 | 0.000024 |

La relazione è molto significativa. Si può ancora accettare l'ipotesi di $\beta_1=0$, ma rischiando di sbagliare, nelle condizioni poste, 999'986 volte su di un milione.

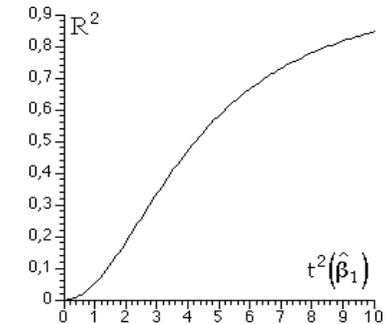
| Giorno | T.M. | C.E. |
|--------|------|------|
| 1 | 95 | 214 |
| 2 | 82 | 152 |
| 3 | 90 | 156 |
| 4 | 81 | 129 |
| 5 | 99 | 254 |
| 6 | 100 | 266 |
| 7 | 93 | 210 |
| 8 | 95 | 204 |
| 9 | 93 | 213 |
| 10 | 84 | 150 |

Adattamento e test

Se l'ipotesi $H_0 : \beta_1=0$, è respinta significa che la "Y" è, almeno in parte, spiegata da una relazione lineare nella "X". Quanta parte?

Nel caso della regressione semplice è possibile stabilire un legame diretto tra la statistica test di $\hat{\beta}_1$ ed il coefficiente di determinazione R^2 .

$$R^2 = \frac{[t(\hat{\beta}_1)]^2}{[t(\hat{\beta}_1)]^2 + (n - 2)}$$

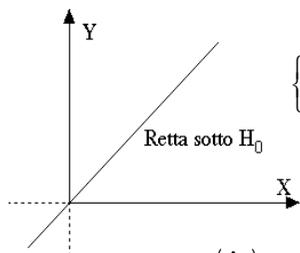


Esiste perciò una relazione biunivoca tra R^2 e la statistica test (al quadrato) per cui valori significativi di questa danno luogo a valori elevati di quello e viceversa.

Questo non è più vero nella regressione multipla

Test sull'intercetta

La verifica dell'intercetta è poco interessante dato che non ha incidenza sulla bontà di adattamento.



$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

Se si accetta l'ipotesi si dice che la retta teorica ha equazione

$$y_i = \beta_1 x_i$$

cioè passa per l'origine

$$E(\hat{\beta}_0) = \beta_0, \quad Var(\hat{\beta}_0) = \frac{s_e^2}{n} \left(1 + \frac{n\mu_x^2}{S_{xx}} \right)$$

Il p-value del test si ottiene dalla t-Student (n-2) per il valore dato da

$$t_c(\hat{\beta}_0) = \frac{\hat{\beta}_0}{\frac{s_e}{\sqrt{n}} \sqrt{1 + \frac{n\mu_x^2}{S_{xx}}}}$$

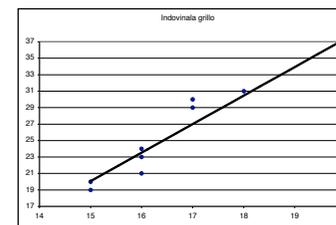
Esempio

Si ritiene che la frequenza dei trilli del grillo sia legata alla temperatura e che questa si possa indovinare dall'intensità del canto.

La relazione non è contraddetta dai (pochi) dati considerati.

L'intercetta comunque c'è. Chi dice il contrario afferma il vero 6 volte su diecimila

| Grillo | Frequenza | Temperatura |
|--------|-----------|-------------|
| 1 | 20 | 35 |
| 2 | 16 | 21 |
| 3 | 20 | 38 |
| 4 | 18 | 31 |
| 5 | 17 | 29 |
| 6 | 16 | 24 |
| 7 | 15 | 20 |
| 8 | 17 | 30 |
| 9 | 15 | 19 |
| 10 | 16 | 23 |



| Regression Statistics | |
|-----------------------|--------|
| Multiple R | 0.9640 |
| R Square | 0.9292 |
| Adjusted R Sc | 0.9204 |
| Standard Error | 1.8529 |
| Observations | 10 |

ANOVA

| | df | SS | MS | F | p-value |
|------------|----|----------|----------|----------|---------|
| Regression | 1 | 360.5333 | 360.5333 | 105.0097 | 0.0000 |
| Residual | 8 | 27.4667 | 3.4333 | | |
| Total | 9 | 388 | | | |

| | Coefficients | Standard Error | t Stat | P-value |
|--------------|--------------|----------------|---------|---------|
| Intercept | -31.9333 | 5.7808 | -5.5240 | 0.0006 |
| X Variable 1 | 3.4667 | 0.3383 | 10.2474 | 0.0000 |