Esercitazioni di statistica

Test delle ipotesi

Stefania Spina

Universitá di Napoli Federico II

stefania.spina@unina.it

17 Dicembre 2014

INTRODUZIONE

L'obiettivo di molti studi è quello di verificare se i dati raccolti concordano con determinate previsioni che, solitamente, derivano dagli aspetti teorici che hanno guidato la ricerca.

↓

Queste PREVISIONI sono dette IPOTESI SULLA POPOLAZIONE

IPOTESI

In statistica, un' IPOTESI è un' AFFERMAZIONE SULLA POPOLAZIONE. Solitamente è la previsione che un parametro, impiegato per descrivere alcune caratteristiche di una variabile, assuma un particolare valore numerico o ricada in un certo intervallo di valori.

Ipotesi statistica

E' possibile distinguere **due tipologie** di **ipotesi statistica**:

- **semplice**, quando si afferma che il **parametro** sia pari ad un dato **valore numerico**, cioè $\theta = \theta_0$;
- **composta**, quando si afferma che il **parametro** sia pari ad **uno tra più valori numerici**, che a sua volta si distingue in:
  - **unidirezionale**, quando si afferma che il **parametro** sia maggiore o minore di un dato valore numerico, cioè $\theta > \theta_0$ ***oppure*** $\theta < \theta_0$;
  - **bidirezionale** quando si afferma che il **parametro** sia diverso da un particolare valore numerico dello spazio parametrico, ovvero può assumere un qualsiasi altro valore $\theta \neq \theta_0$;

TEST DELLE IPOTESI

Il test consiste nel formulare un'ipotesi e nel verificare se con i dati campionari a disposizione è possibile **rifiutarla** o no.

L'ipotesi che viene formulata è l'ipotesi nulla (H_0) e rappresenta di solito l'affermazione su un particolare valore assunto dal parametro.

Se il campione fornisce risultati fortemente in contrasto con H_0, questa viene rifiutata a favore dell'ipotesi alternativa (H_1).

Test e regole di decisione

Da un punto di vista operativo, un **test è una statistica** che fa corrispondere ad ogni campione casuale $(X_1, \ldots, X_n)$ un valore numerico che può essere classificato secondo due diverse possibilità:

- coerente con H_0;
- non coerente con H_0.

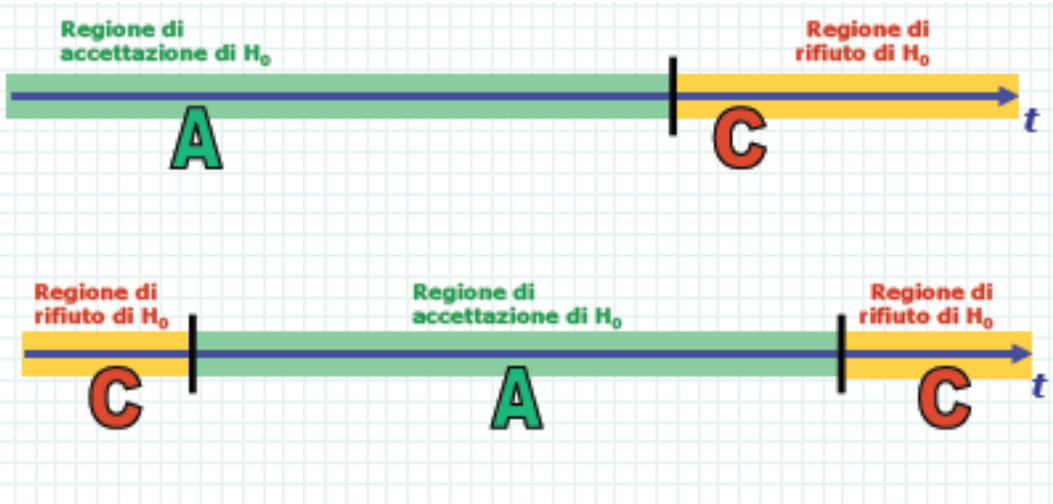
Un test statistico dà quindi luogo alla **ripartizione dello spazio campionario** in due **sottoinsiemi complementari**:

- un insieme A costituito dai valori del test che sono compatibili con l'ipotesi nulla H_0, denominato **regione di accettazione**;
- un insieme C che raggruppa i valori del test considerati incompatibili con H_0, denominato **regione critica o di rifiuto**.

Test e regole di decisione

La **regola decisionale** è definita sulla base del valore campionario della statistica $\boldsymbol{T_n = T(X_1, \ldots, X_n)}$ e della **regione di accettazione**.

Quando il valore campionario della statistica $T_n = T(X_1, \ldots, X_n)$ cade nella regione critica, l'evidenza empirica del fenomeno studiato porta a ritenere che l'ipotesi H_0 non possa essere considerata valida, e quindi che non possa essere accettata come vera.



Regione critica per un test statistico con ipotesi alternativa unidirezionale:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

Regione critica per un test statistico con ipotesi alternativa bidirezionale:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Errori e probabilità di errore di I e II tipo

Indipendentemente dalla regola adottata, il test porta sempre a dover scegliere tra due possibili decisioni, H_0 e H_1 e a poter commettere **due possibili errori**, rifiutare un'ipotesi vera oppure accettare un'ipotesi falsa.

	Non rifiuto H_0	Rifiuto H_0
H_0 vera	Nessun errore confidenza=1-α	Errore di I° tipo (α)
H_0 falsa	Errore di II° tipo (b)	Nessun errore Potenza=1-β

$\alpha = P(\text{rifiutare } H_0 \mid H_0 \text{ è vera})$; $1-\alpha = P(\text{non rifiutare } H_0 \mid H_0 \text{ è vera})$

$\beta = P(\text{non rifiutare } H_0 \mid H_0 \text{ è falsa})$; $1-\beta = P(\text{rifiutare } H_0 \mid H_0 \text{ è falsa})$

FASI DA SEGUIRE PER IL TEST DELLE IPOTESI

1. Specificare H_0, H_1 ed un livello α
2. Definire una statistica per il test (statistica di cui sia definibile la distribuzione campionaria) e la zona di rifiuto per H_0 (valori della statistica aventi probabilità $< \alpha$ quando H_0 è vera).
3. Eseguire il campionamento (o l'esperimento) e calcolare la statistica.
4. Se la statistica calcolata cade nella zona di rifiuto decido di rifiutare H_0, altrimenti decido di non rifiutare H_0.

Approccio del valore critico

Approccio del p-value (o p-valore)

In alternativa all'approccio del valore critico, si può riportare direttamente il valore della probabilità *p* di commettere l'errore di I tipo (livello di significatività osservato).

Il *p*-value è una misura di quanto i dati sono in disaccordo con H_0 che si presume sia vera. E' la probabilità che il test statistico sia pari al valore osservato o a uno più grande nella direzione prevista da H_1.

Per esempio per un test per μ (σ noto), si può procedere come segue:

1. Definire H_0: $\mu = \mu_0$; $H_1 : \mu \neq \mu_0$

2. Si calcola la media campionaria e si converte nella variabile standardizzata:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

3. Si calcola la probabilità *p* di ottenere un valore esterno all'intervallo –z, +z: $P(Z< -z) + P(Z>+z)$ ovvero $P(Z > |z|)$ (per test a 2 code)

Test parametrici

Ci sono diversi test delle ipotesi che vengono più utilizzati nelle applicazioni reali.
I primi test che vengono esaminati riguardano i parametri di una v.c. normale, cioè la media e la varianza.

Test sulla media di una popolazione X

μ incognita
σ^2 nota

Supponiamo che sia $X \sim N(\mu; \sigma^2)$, con μ incognita e quindi risulti $\theta = (\mu)$, e si voglia sottoporre a test le ipotesi:

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Il **criterio di decisione** adottato è quello di rifiutare il valore μ_0 come media μ della popolazione, se la media campionaria $\overline{X}$ calcolata su un campione casuale $(X_1, \ldots, X_n)$ è molto distante dal valore μ_0 ipotizzato sotto H_0.

La **regione critica del test** sarà formata dai valori campionari di $\overline{X}$ per i quali la:

$$P(\overline{X} < \overline{X_I}|H_0) = P(\overline{X} > \overline{X_S}|H_0) = \frac{\alpha}{2}$$

Test sulla media di una popolazione X

Poichè la statistica campionaria $\overline{X}$ segue una distribuzione normale, la v.c. standardizzata $\frac{(\overline{X}-\mu)}{\frac{\sigma}{\sqrt{n}}}$ seguirà una distribuzione N(0;1) e dalle tavole della z sarà possibile trovare i **valori critici** $+z_{\alpha/2}$ e $-z_{\alpha/2}$ che attribuiscono un'area di probabilità $\alpha/2$ alle code:

$$-z_{\alpha/2} = \frac{(\overline{X_I} - \mu_0)}{\frac{\sigma}{\sqrt{n}}} \qquad + z_{\alpha/2} = \frac{(\overline{X_S} - \mu_0)}{\frac{\sigma}{\sqrt{n}}}$$

i cui **valori discriminanti** saranno

$$\overline{X_I} = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\overline{X_S} = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Test sulla media di una popolazione X

La **regola di decisione** per H_0 sarà perciò la seguente:

- si **accetta** H_0: $\mu = \mu_0$ se la **media campionaria** cade all'**interno dell'intervallo** $\mu_0 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- si **rifiuta** H_0: $\mu = \mu_0$ in favore di $H_1 : \mu \neq \mu_0$ se la **media campionaria** cade al di **fuori dell'intervallo** $\mu_0 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

La **statistica - test** da adoperare è:

$$z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

Esercizio

Un'impresa che produce filo metallico dichiara che il punto di rottura, in media, è pari a 130 kg ed ha scarto quadratico medio pari a 5. Un cliente decide sull'opportunità di acquistare il filo solo se presenta un punto di rottura che non diverge significativamente da quello dichiarato dall'impresa; a tal fine analizza un campione di 16 matasse e ne calcola il punto di rottura medio.
Supponendo che la v.c. Punto di rottura si distribuisca in modo normale, stabilire una regola di decisione per il cliente, ad un livello di significatività del 5%.

Esercizio

$$H_0 = \mu = 130$$

$$H_1 = \mu \neq 130$$

Il test da utilizzare è a 2 code, per cui $\alpha = 0.05$ $\alpha/2 = 0.025$, il valore tabulato di

$$z_{\alpha/2} = z_{0.025} = 1.96$$

La regola di decisione è la seguente:

$$z < -1.96 \; o \; z > 1.96 \; \textit{si rifiuta} \; H_0$$

$$-1.96 \leq z \leq 1.96 \; \textit{si accetta} \; H_0$$

Esercizio

I valori discriminanti saranno:

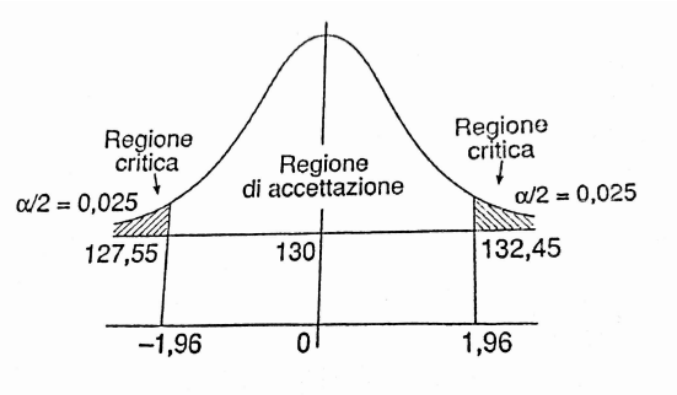
$$\overline{X_I} = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 130 - 1.96 * \frac{5}{\sqrt{16}} = 127.55$$

$$\overline{X_S} = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 130 + 1.96 * \frac{5}{\sqrt{16}} = 132.45$$

La regola di decisione è, dunque, la seguente:

$$\textit{per } \overline{X} < 127.55 \textit{ o } \overline{X} > 132.45 \textit{ si rifiuta } H_0$$
$$\textit{per } 127.55 \leq \overline{X} \leq 132.45 \textit{ si accetta } H_0$$

Graficamente



Esercizio

Un' azienda che produce scatole metalliche intende valutare se il processo produttivo opera in modo tale da garantire che la lunghezza del lato maggiore sia pari a 368 mm. Viene estratto un campione di 25 scatole. Lo scarto quadratico medio della popolazione che segue una distribuzione Normale è pari a 15 mm e la media campionaria assume il valore 372.5 mm. Utilizzare l'approccio del p- value.

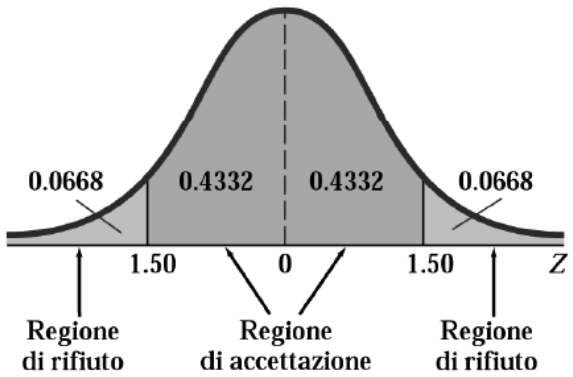
Esercizio

$$H_0 = \mu = 368$$
$$H_1 = \mu \neq 368$$

Con l'approccio del p- value si calcola la probabilità p di ottenere un valore esterno all'intervallo [-z; +z]: P(Z > z) oppure P(Z > |z|) per un test a 2 code

$$z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{(372.5 - 368)}{\frac{15}{\sqrt{25}}} = 1.5$$

Le probabilità che z assuma valori maggiori di 1.5 o minori di -1.5 (test a due code) sono pari a 0.0668 e la loro somma è perciò 0.1336.
Graficamente



Test sulla media di una popolazione X

μ incognita
σ^2 incognita

Quando la varianza della popolazione è incognita e la dimensione del campione è poco elevata la distribuzione del test media campionaria sotto H_0 è solo approssimativamente normale e più precisamente è rappresentata dalla distribuzione t_{n-1}.

Quando σ^2 è incognita, viene sostituita con il suo stimatore corretto:

$$s^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$$

e il rapporto

$$\frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$$

segue una distribuzione t di student con (n-1) gradi di libertà

Test sulla media di una popolazione X

Scelto un livello di significatività α, dalle tavole della t è possibile trovare i valori $\pm t_{\alpha/2}$ che individuano l'intervallo centrato su zero al di fuori del quale la variabile t ha una probabilità α di assumere valori. Sotto H_0 vale la relazione:

$$\overline{X_I} = \mu_0 - t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$\overline{X_S} = \mu_0 + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

La **regola di decisione** per H_0 sarà perciò la seguente:

- si **accetta** H_0: $\mu = \mu_0$ se la **media campionaria** cade all'**interno dell'intervallo** $\mu_0 \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
- si **rifiuta** H_0: $\mu = \mu_0$ in favore di $H_1 : \mu \neq \mu_0$ se la **media campionaria** cade al di **fuori dell'intervallo** $\mu_0 \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$

Test sulla media di una popolazione X

La **statistica** da usare per il test è t con n-1 gradi di libertà

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Esercizio

Un campione di 20 alberghi italiani presenta un numero di letti pari in media a 70, con scarto quadratico medio corretto pari a 25. Sapendo che gli alberghi contano in media 61 letti:
a) verificare l'ipotesi nulla che la media si possa considerare invariata, al livello di significatività del 99%;
b) verificare la stessa ipotesi rispetto ad un campione di ampiezza 100.

Esercizio

a) test sulla media bilaterale: distribuzione non nota, varianza non nota, campione piccolo

$$H_0 = 61$$
$$H_1 \neq 61$$

Dato che $\alpha = 0.01$ e il test è bilaterale e si considererà $\alpha/2 = 0.005$
I valori critici saranno:

$$t_{\alpha/2;n-1} = \pm 2.861$$

La regola di decisione sarà:

$$\textit{per } \overline{X} < -2.86 \textit{ o } \overline{X} > 2.86 \textit{ si rifiuta } H_0$$
$$\textit{per } -2.86 \leq \overline{X} \leq 2.86 \textit{ si accetta } H_0$$

La statistica test sarà:

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{70 - 61}{\frac{25}{\sqrt{20}}} = 1.61$$

Poichè:

$$-2.86 < 1.61 < 2.86$$

Esercizio

b) test sulla media bilaterale: distribuzione non nota, varianza non nota, campione grande

$$H_0 = 61$$
$$H_1 \neq 61$$

Dato che $\alpha = 0.01$ e il test è bilaterale e si considererà $\alpha/2 = 0.005$
I valori critici saranno:

$$t_{\alpha/2;n-1} = \pm 2.58$$

La regola di decisione sarà:

$$per\ \overline{X} < -2.58\ o\ \overline{X} > 2.58\ si\ rifiuta\ H_0$$
$$per\ -2.58 \leq \overline{X} \leq 2.58\ si\ accetta\ H_0$$

La statistica test sarà:

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{70 - 61}{\frac{25}{\sqrt{100}}} = 3.6$$

Poichè:

$$3.6 > 2.58$$

Si rifiuta H_0

Potenza del test

La potenza di un test statistico è la probabilità di prendere la decisione giusta, cioè la probabilità di rifiutare l'ipotesi nulla quando è falsa:

$$\pi = 1 - \beta = P(\textit{rifiutare } H_0 | H_0 \textit{ falsa})$$

La potenza di un test è la sua capacità di cogliere delle differenze, quando queste differenze esistono.

Esercizio

Si vuole studiare la durata di un processo produttivo che dal materiale grezzo porta al prodotto finito. Il venditore del meccanismo di produzione sostiene che la durata del processo si distribuisce normalmente con media pari 11 ore e scarto quadratico medio pari a 4 ore.
L'acquirente, sulla base delle valutazioni di un esperto, sospetta invece che, pur distribuendosi normalmente e con scarto quadratico medio 4, la durata media del processo sia 14 ore.
Si mettono allora in produzione 16 pezzi e si decide che il meccanismo di produzione verrà acquistato soltanto se la durata media della produzione nel campione è inferiore a 13.
Si calcolino la probabilità dell'errore del I tipo (α) e la probabilità dell'errore del II tipo (β), associati al criterio di decisione sopra riportato.

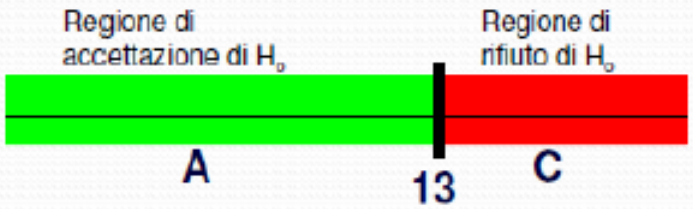
Esercizio

$$H_0 : \mu = 11 \Rightarrow X \approx N(11,4^2) \rightarrow \overline{X} \approx N\left(11, \frac{4^2}{16}\right)$$

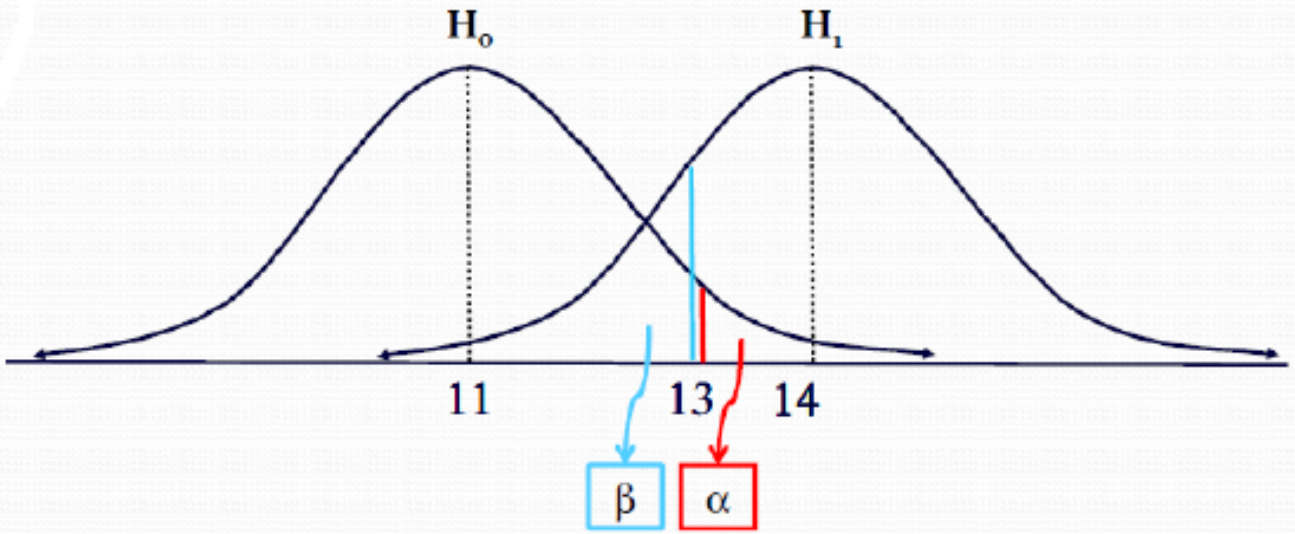
$$H_1 : \mu = 14 \Rightarrow X \approx N(14,4^2) \rightarrow \overline{X} \approx N\left(14, \frac{4^2}{16}\right)$$

Regola di decisione

rifiuto H_0 se $\overline{X} \geq 13$



Esercizio



$$\alpha = P\{\overline{X} \geq 13 \mid \mu = 11\} \Rightarrow \alpha = P\left\{Z \geq \frac{13-11}{1}\right\} = P(Z \geq 2) = 0{,}0228$$

$$\beta = P\{\overline{X} \leq 13 \mid \mu = 14\} \Rightarrow \beta = P\left\{Z \leq \frac{13-14}{1}\right\} = P(Z \leq -1) = 0{,}1587$$

Test sulla differenza tra medie

Per alcuni problemi di inferenza può essere importante stabilire se le medie di due popolazioni debbano considerarsi sostanzialmente uguali oppure no.

σ^2 nota

Si suppone di avere due popolazioni X e Y tra loro indipendenti e distribuite normalmente, rispettivamente con media μ_x eμ_y e varianza comune σ^2 si estraggono in modo casuale due campioni rappresentati da $(X_1, X_2, \ldots, X_n)$ e $(Y_1, Y_2, \ldots, Y_m)$.
Sulla base di questi due campioni si vuole verificare l'ipotesi:

$$H_0 : \mu_x = \mu_y \quad H_1 : \mu_x \neq \mu_y$$

Che possono essere riscritte come:

$$H_0 : \mu_x - \mu_y = 0 \quad H_1 : \mu_x - \mu_y \neq 0$$

Test sulla differenza tra medie

La regione critica del test H_0 che si vuole proporre non deve dipendere da μ_x $e\mu_y$ e quindi si deve trovare un test la cui distribuzione campionaria sotto H_0 non sia funzione di μ_x $e\mu_y$.

Poichè le medie campionarie

$$\overline{X} \sim N(\mu_x; \sigma^2/n)$$
$$\overline{Y} \sim N(\mu_y; \sigma^2/m)$$

sono anche indipendenti tra loro, ne consegue che la statistica campionaria

$$(\overline{X} - \overline{Y}) \sim N[(\mu_x - \mu_y); \sigma^2(1/n + 1/m)]$$

la quale non dipende da $(\mu_x - \mu_y)$ sotto $H_0 : \mu_x = \mu_y$.
Quindi si avrà:

$$H_0 : \mu_x - \mu_y = 0 \quad H_1 : \mu_x - \mu_y \neq 0$$

Test sulla differenza tra medie

La statistica test da utilizzare sarà:

$$Z_{H_0} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

Test sulla differenza tra medie

Scelto un livello di significatività per il test, dalle tavole della normale standardizzata si troveranno i valori discriminanti $\pm z_{\alpha/2}$ dai quali si potranno dedurre i valori critici del test rappresentati da:

$$(\overline{X} - \overline{Y})_I = -z_{\alpha/2} * \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

$$(\overline{X} - \overline{Y})_S = +z_{\alpha/2} * \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

Test sulla differenza tra medie

La **regola di decisione** per H_0 sarà perciò la seguente:

- si **accetta** H_0: $\mu_x - \mu_y = 0$ se la **differenza delle medie campionarie** cade all'**interno dell'intervallo** $\pm z_{\alpha/2} * \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$
- si **rifiuta** H_0

Test sulla differenza tra medie: Esercizio

Una fabbrica A produce batterie per automobili la cui durata media è distribuita normalmente con media incognita e varianza $\sigma_A^2 = 1.5$. Un'altra fabbrica B produce anch'essa batterie per auto la cui durata è distribuita normalmente con media incognita e varianza $\sigma_B^2 = 1.3$. Considerato un campione di 21 batterie della fabbrica A si è osservata una **durata media di 3.4 anni**, mentre preso un campione di 16 batterie della fabbrica B si è osservata una **durata media di 3.1 anni**. Un produttore di auto, prima di acquistare una partita di batterie, vuole verificare l'ipotesi che la durata media delle batterie delle due fabbriche A e B è uguale, fissato un livello di significatività del 5%.

Test sulla differenza tra medie: Esercizio

Le ipotesi a confronto saranno:

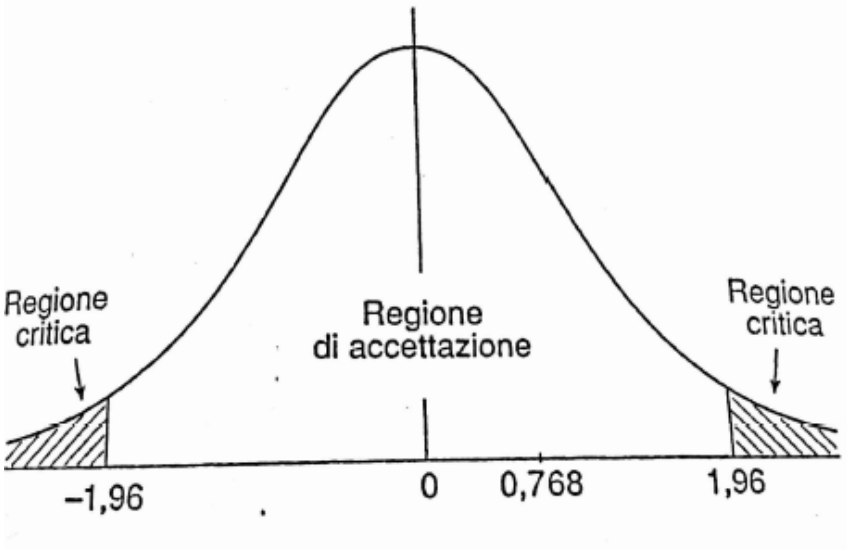
$$H_0 = \mu_A - \mu_B = 0$$

$$H_0 = \mu_A - \mu_B \neq 0$$

Tale v.c. si distribuisce come una normale standardizzata e, quindi, in corrispondenza di $\alpha/2$ (trattandosi di un test bidirezionale), si desume che i due valori critici sono -1.96 e 1.96.
La statistica test che sarà utilizzata sarà:

$$z = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{3.4 - 3.1}{\sqrt{\frac{1.5}{21} + \frac{1.3}{16}}} = 0.768$$

Test sulla differenza tra medie: Esercizio



Essendo -1.96 < 0.78 < 1.96 il valore calcolato ricade nella regione di accettazione del test, per cui si accetta l'ipotesi nulla.
Si può quindi dire che il produttore di auto, relativamente a questa verifica, è indifferente nella scelta tra A e B per la propria fornitura.

Test sulla differenza tra medie

σ^2 non è nota

Quando si suppone che le varianze delle due popolazioni siano uguali, una loro stima congiunta può essere ottenuta con la media aritmetica delle due varianze campionarie, ponderata dai rispettivi gradi di libertà n_m - 1 e n_f - 1:

$$s^2 = \frac{(n_m - 1)s_m^2 + (n_f - 1)s_f^2}{n_m + n_f - 2}$$

Usandolo al posto di σ^2 il test statistico sarà:

- **Quando le popolazioni sono indipendenti e le varianze sono incognite ma uguali**

$$\frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

- **Quando le popolazioni sono indipendenti e le varianze sono incognite ma differenti**

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \sim t_{(\upsilon)} \text{ (\textit{soluzione approssimata di Welch})}$$

Test sulla differenza tra medie

Quando le popolazioni sono indipendenti e le varianze sono incognite ma uguali

Fissato α si troveranno dalla tavole della t i valori $t_{\alpha/2}$ dai quali è possibile dedurre i valori discriminanti che individuano la regione critica sotto l'ipotesi nulla che sono:

$$(\overline{X} - \overline{Y})_I = -t_{\alpha/2} * s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

$$(\overline{X} - \overline{Y})_S = +t_{\alpha/2} * s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

e la regola di decisione per l'ipotesi nulla sarà sempre la stessa:

- si **accetta** H_0: $\mu_x - \mu_y = 0$ se la **differenza delle medie campionarie** cade all'**interno dell'intervallo** $\pm t_{\alpha/2} * s\sqrt{\frac{1}{n} + \frac{1}{m}}$
- si **rifiuta** H_0

Test sulla differenza tra medie: Esercizio

Si esegue un'indagine sul reddito medio annuo (in euro) di un campione di laureati (maschi e femmine) in discipline economiche dopo un anno dalla laurea e si osserva che:

$$n_m = 10 \quad \overline{x}_m = 14000 \quad s_m^2 = 210000$$
$$n_f = 12 \quad \overline{x}_f = 12500 \quad s_f^2 = 225000$$

Sulla base dei risultati campionari, verificare la validità dell'ipotesi che non c'è differenza tra i redditi medi delle due popolazioni di maschi e femmine, ad un livello di significatività $\alpha = 0.05$; supponendo che:

$$H_0 : \mu_m - \mu_f = 0$$
$$H_1 : \mu_m - \mu_f \neq 0$$

Test sulla differenza tra medie: Esercizio

Avendo supposto che le varianze delle due popolazioni siano uguali, una loro stima congiunta può essere ottenuta con la media aritmetica delle due varianze campionarie, ponderata dai rispettivi gradi di libertà $n_m - 1$ *e* $n_f - 1$

$$s^2 = \frac{(n_m - 1)s_m^2 + (n_f - 1)s_f^2}{n_m + n_f - 2}$$

si ha quindi:

$$s^2 = \frac{9 * 210000 + 11 * 225000}{10 + 12 - 2} = 218250$$

da cui la stima dello scarto quadratico medio s= 467.17

Test sulla differenza tra medie: Esercizio

La statistica - test da adoperare è:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

per cui il valore empirico sarà:

$$t = \frac{14000 - 12500}{467.17 * \sqrt{\frac{1}{10} + \frac{1}{12}}} = 7.499$$

Test sulla differenza tra medie: Esercizio

La regola di decisione è la seguente:

$$-t_{0.025;20} \leq t \leq t_{0.025;20} \quad \textit{non si rifiuta } H_0$$

$$t < t_{0.025;20} \quad o \quad t > t_{0.025;20} \quad \textit{si rifiuta } H_0$$

Per $\alpha/2 = 0.025$ e 20 gradi di libertà si ha $|t_{0.025;20}| = 2.086$
Essendo $7.499 > 2.086$ si rifiuta H_0 e si può affermare che il reddito medio netto annuo dei maschi è maggiore di quello delle femmine.

Test per la proporzione della popolazione

Per una variabile categoriale, il parametro di interesse è la proporzione di una categoria nella popolazione.

In questo tipo di test la dimensione campionaria deve essere sufficientemente ampia in modo che la distribuzione campionaria di $\hat{p}$ sia approssimativamente normale.

Test per la proporzione della popolazione

$$X \sim Ber(1, p)$$

$$X = \begin{cases} 0 & 1-p \\ 1 & p \end{cases}$$

Le ipotesi da verificare sono:

$$H_0 : p = p_0$$
$$vs$$
$$H_1 : p \neq p_0$$
$$H_1 : p > p_0$$
$$H_1 : p < p_0$$

Test per la proporzione della popolazione

Tale popolazione è caratterizzata dal parametro p e ne consegue che per grandi campioni è approssimativamente una distribuzione normale $\left[p; \frac{p(1-p)}{n}\right]$

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \; per \; n \to \infty$$

Test d'ipotesi sulla proporzione

Se $X \sim Ber(p)$

Se è vera H_0, (per $n>30$) $\dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1)$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\text{n° di successi}}{n}$$

Test a due code

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Si rifiuta H_0 se $|z| > z_{1-\alpha/2}$ dove $z = \dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Test a una coda

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p > p_0 \end{cases}$$

Si rifiuta H_0 se $z > z_{1-\alpha}$ dove $z = \dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases}$$

Si rifiuta H_0 se $z < z_{\alpha}$ dove $z = \dfrac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Esercizio

L'ufficio di marketing di una grande azienda sostiene che il 30% delle famiglie acquista un determinato prodotto per la pulizia della casa.
Si decide di effettuare un'indagine campionaria su 200 famiglie da cui risulta che solo 45 famiglie acquistano quel prodotto.
Verificare l'affermazione dell'ufficio marketing circa l'acquisto del prodotto ad un livello di significatività $\alpha = 0.05$ utilizzando un test bilaterale.

Esercizio

Il test in esame è quello su una proporzione avente sistema di ipotesi:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

con statistica test:

$$Z = \frac{(\hat{p} - p_0)}{\sqrt{p_0 * (1 - p_0)/n}} \sim N(0, 1)$$

e regione critica $|Z| > z_{\alpha/2}$

Esercizio

$$H_0 : p = 0.30$$
$$H_1 : p \neq 0.30$$

Il valore critico, ad un livello di $\alpha = 0.05$ risulta essere $z_{\alpha/2} = \pm 1.96$
Quindi la regola di decisione consiste nel rifiutare H_0 se la variabile campionaria normalizzata risulta compresa nell'intervallo
$-1.96 \leq z \leq 1.96$
Eseguendo i calcoli si ha:

$$\hat{p} = \frac{45}{200} = 0.225$$
$$Z = \frac{0.225 - 0.30}{\sqrt{0.30 * (1 - 0.70)/200}} = -2.31$$

Poichè il valore di Z non è compreso nell'intervallo [-1.96; 1.96] si rifiuta l'ipotesi dell'ufficio marketing circa la percentuale di famiglie che acquistano il prodotto.

Test per il confronto tra proporzioni in due popolazioni

Siano date due popolazioni indipendenti

$$X \sim Ber(p_x)$$

$$Y \sim Ber(p_y)$$

di cui non sono note le due medie, rispettivamente, p_x e p_y
Le ipotesi da formulare saranno:

$$H_0 : p_x = p_y$$

$$H_1 : p_x \neq p_y$$

Test per il confronto tra proporzioni in due popolazioni

La statistica test sarà:

$$Z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1-\hat{p}) * (\frac{1}{n_x} + \frac{1}{n_y})}} \sim N(0;1)$$

dove:

$$\hat{p}_x = \frac{n.\ successi\ in\ x}{n_x}$$

$$\hat{p}_y = \frac{n.\ successi\ in\ y}{n_y}$$

$$\hat{p} = \frac{n_x \hat{p}_x + n_y \hat{p}_y}{n_x + n_y}$$

Esercizio

In uno studio sull'incidenza dell'emicrania sulle persone dedite ad attività sportive si esamina un campione di 150 ragazzi e 200 ragazze, di età compresa tra 16 e 20 anni, e si evince che 30 ragazzi e 48 ragazze soffrono di emicrania abituale.
Sulla base di questi dati è possibile concludere, ad un livello di significatività $\alpha = 0.05$, che non c'è differenza significativa tra le proporzioni di sportivi affetti da emicrania?

Esercizio

$$H_0 = p_m - p_f = 0$$

$$H_1 = p_m - p_f \neq 0$$

Le proporzioni campionarie saranno:

$$\hat{p}_m = \frac{30}{150} = 0.2$$

$$\hat{p}_f = \frac{48}{200} = 0.24$$

La stima comune sarà:

$$\hat{p} = \frac{150 * 0.2 + 200 * 0.24}{150 + 200} = 0.22$$

Esercizio

Il valore empirico della statistica - test è:

$$Z = \frac{0.2 - 0.24}{\sqrt{0.22(1 - 0.22)\left(\frac{1}{150} + \frac{1}{200}\right)}} = -0.894$$

Il test a due code per $\alpha = 0.05$ è ± 1.96
Siccome -0.894 è compreso nell'intervallo [-1.96; 1.96] l'ipotesi per cui le due proporzioni non sono differenti non è rifiutata.

Introduzione

Consideriamo il caso del modello di regressione lineare semplice:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

e la relativa stima:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

e si vuole sottoporre a test

$$H_0 : \beta_0 = 0 \quad H_0 : \beta_0 \neq 0$$
$$H_0 : \beta_1 = 0 \quad H_0 : \beta_1 \neq 0$$

Introduzione

La statistica - test da adoperare è:

- per l'ipotesi $H_0 : \beta_0 = 0$

$$T_0 = \frac{\hat{\beta}_0}{es(\beta_0)}$$

- per l'ipotesi $H_0 : \beta_1 = 0$

$$T_1 = \frac{\hat{\beta}_1}{es(\beta_1)}$$

E' data dal rapporto tra la stima campionaria e il suo errore standard, e si distribuisce, se l'ipotesi nulla è vera, come una v.c. t di Student, con n-2 gradi di libertà.

Introduzione

Le regioni critiche saranno, rispettivamente:

$$RC : |T_0| > t_{(\alpha/2;n-2)}$$

$$RC : |T_1| > t_{(\alpha/2;n-2)}$$

Esercizio

Su 7 autovetture a gasolio, scelte a caso da un dato parco macchine, è stato verificato il consumo, per miglia, prima di un determinato intervento (variabile X) e dopo l'intervento (variabile Y) ottenendo le 7 coppie di risultati seguenti:
(17.2; 18.3) (21.6; 20.8) (19.5; 20.9) (19.1; 21.2) (22.9; 22.7) (18.7; 18.6) (20.3; 21.9)
Si vuole verificare:

- se fra X ed Y esiste il legame lineare

$$y = \beta_0 + \beta_1 x + \epsilon$$

Esercizio

x_i	y_i	x_i^2
17.20	18.30	95.840
21.60	20.80	66.560
19.50	20.90	80.250
19.10	21.20	64.810
22.00	22.70	84.000
18.70	18.60	49.690
20.30	21.90	12.090
138.40	144.40	2753.240

$$\overline{x} = \frac{138.4}{7} = 19.771429$$

$$\overline{y} = \frac{144.4}{7} = 20.62857$$

$$E(X^2) = \frac{2753.24}{7} = 393.32001$$

$$E(XY) = \frac{2868.3}{7} = 409.75715$$

Esercizio

$$\hat{\beta}_1 = \frac{cov(XY)}{var(X)} = \frac{E(XY) - \overline{x} * \overline{y}}{E(X^2) * (\overline{X})^2} = \frac{409.75715 - 19.771429 * 20.62857}{393.32001 - (19.771429)^2}$$

$$= 0.78852$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 * \overline{x} = 20.62857 - 0.78852 * 19.771429 = 5.0384$$

Esercizio

Una volta ottenute le stime dei due parametri deriviamo le stime della variabile dipendente $\hat{y}$ e quelle dei residui $\hat{e} = y_i - \hat{y}_i$

x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i$	$\hat{e}_i = y_i - \hat{y}_i$
17.20	18.30	95.840	14.760	18.60094	-0.300945
21.60	20.80	66.560	49.280	22.07043	-1.270433
19.50	20.90	80.250	07.550	20.41454	0.485460
19.10	21.20	64.810	04.920	20.09913	1.100868
22.00	22.70	84.000	99.400	22.38584	0.314161
18.70	18.60	49.690	47.820	19.78372	-1.183723
20.30	21.90	12.090	44.570	21.04536	0.854645
138.40	144.40	2753.240	2868.300	144.39996	0.000033

Esercizio

$$s^2 = \frac{1}{n-2} * \sum_{i=1}^{n}(y_i - \hat{y})^2 = \frac{1}{7-2} * 5.382467 = 1.076493$$

$$es(\beta_0) = s * \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 1.0375418 * \sqrt{\frac{1}{7} + \frac{(19.771429)^2}{16.8742857}} = 5.0091713$$

$$es(\beta_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{1.0375418}{\sqrt{16.8742857}} = 0.252576$$

Esercizio

Dato che le ipotesi da testare saranno:

$$H_0 : \beta_0 = 0 \quad H_0 : \beta_0 \neq 0$$
$$H_0 : \beta_1 = 0 \quad H_0 : \beta_1 \neq 0$$

La statistica - test da adoperare è:

- per l'ipotesi $H_0 : \beta_0 = 0$

$$T_0 = \frac{\hat{\beta}_0}{es(\beta_0)} = \frac{5.0384}{5.0091713} = 1.006$$

- per l'ipotesi $H_0 : \beta_1 = 0$

$$T_1 = \frac{\hat{\beta}_1}{es(\beta_1)} = \frac{0.78852}{0.252576} = 3.12$$

Esercizio

Al livello di significatività del 5%

$$t_{(\alpha/2;n-2)} = 2.571$$

Considerate che le regioni critiche sono rispettivamente:

$$RC : |T_0| > t_{(\alpha/2;n-2)}$$
$$RC : |T_1| > t_{(\alpha/2;n-2)}$$

Si avrà che:
$1.006 < 2.571$ quindi posso accettare l'ipotesi nulla che $H_0 : \beta_0 = 0$
$3.12 > 2.571$ quindi posso rifiutare l'ipotesi nulla che $H_0 : \beta_1 = 0$

In definitiva, il modello di regressione privo di intercetta potrebbe essere quello idoneo per descrivere il fenomeno Y in funzione di X.

Introduzione

In una tabella a doppia entrata con r righe e c colonne, in cui si rilevano due caratteri (qualitativi o quantitativi, discreti o continui) X e Y, sia n_{ij} la frequenza con cui si presenta la coppia di modalità $(x_i; y_j)$, sia dato, inoltre, un campione casuale di n unità, si voglia verificare, al livello di significatività α, l'ipotesi nulla H_0 che i due caratteri sono assolutamente indipendenti, ossia:

$$H_0 : n^*_{ij} = \frac{n_{i.} * n_{.j}}{N}$$

$$H_1 : n^*_{ij} \neq \frac{n_{i.} * n_{.j}}{N}$$

Introduzione

La statistica - test da utilizzare è:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

in cui n_{ij}^* sono le **frequenze teoriche calcolate**.
Se l'ipotesi nulla è vera, si distribuisce, al variare del campione, come una v.c. chi - quadrato con (r-1)(c-1) gradi di libertà.
Pertanto, si **rifiuta** H_0 se:

$$X^2 > \chi^2_{\alpha;(r-1)(c-1)}$$

Poichè i risultati per **distribuzioni continue** sono applicati a dati discreti si rende necessaria una correzione nel calcolo della statistica-test:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(|n_{ij} - n_{ij}^*| - 0.5)^2}{n_{ij}^*}$$

Statistica per il test di indipendenza X^2

Distanza tra frequenze osservate e frequenze teoriche

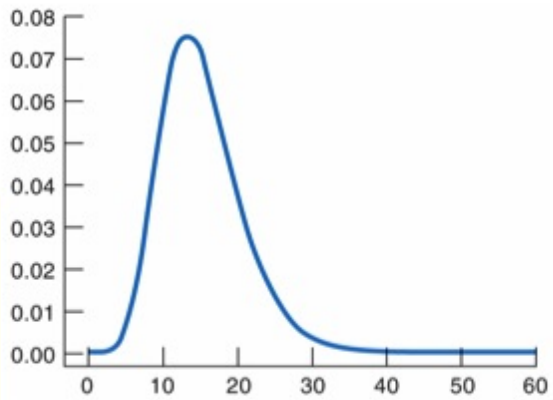
$$c_{ij} = n_{ij} - n_{ij}^*$$

$$X^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$$

$$X^2 = 0 \Leftrightarrow n_{ij} = n_{ij}^*$$

$$X^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} \sim \chi^2_{gl}$$

Una distribuzione chi-quadrato



gl= gradi di libertà

gl= (r-1)(c-1)

r= righe; c= colonne della tabella di contigenza

Test del $\chi 2$ per l'indipendenza

1. H_0: X e Y sono indipendenti

   H_1: X e Y sono dipendenti

2. Statistica per il test X^2

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-n_{ij}^{*}\right)^2}{n_{ij}^{*}} \sim \chi^2_g$$

$$\sum_{\text{tutte le celle}} \frac{(f_o - f_e)^2}{f_e}$$

3. Eseguire il campionamento (o l'esperimento) e calcolare la statistica.

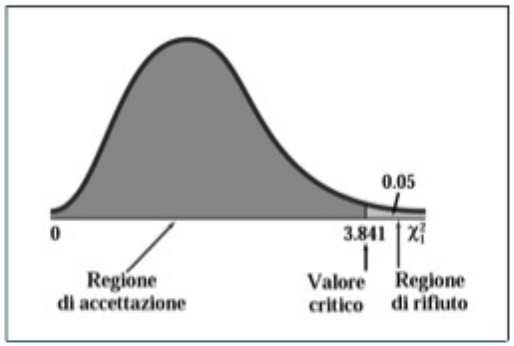
1. Se la statistica calcolata cade nella zona di rifiuto decido di rifiutare H_0, altrimenti decido di non rifiutare H_0.

Utilizzo delle tavole della distribuzione chi-quadrato

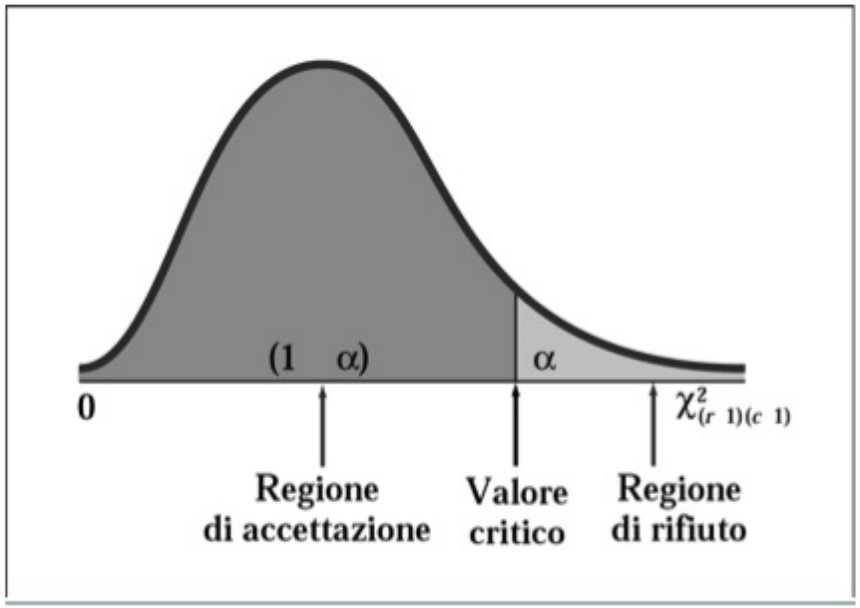
Tabella 8.5 *Determinazione del valore critico al livello di significatività $\alpha = 0.05$ di una distribuzione χ^2 con 1 grado di libertà*

Gradi di libertà	Area nella coda di destra						
	0.995	0.99	...	0.05	0.025	0.01	0.005
1			...	3.841	5.024	6.635	7.879
2	0.010	0.020	...	5.991	7.378	9.210	10.597
3	0.072	0.115	...	7.815	9.348	11.345	12.838
4	0.207	0.297	...	9.488	11.143	13.277	14.860
5	0.412	0.554	...	11.071	12.833	15.086	16.750

Fonte: Tavola E4.



Fonte: Levine, Krehbiel, Berenson, *Statistica*, 2006



Rifiuto H_o se il valore della statistica test X^2 calcolato è maggiore del valore teorico individuato sulle tavole della variabile casuale chi-quadrato con gl gradi di libertà, fissato il valore di α

$$X^2 > \chi^2_{gl,\,\alpha}$$

Esercizio

In un' indagine di mercato sulle preferenze dei consumatori per succhi di frutta ipocalorici, rispetto ai succhi di frutta tradizionali, si selezionano 50 maschi e 50 femmine all'interno di un supermercato. Al campione così estratto si chiede di esprimere una preferenza per uno dei due succhi. Dieci maschi e venti femmine dichiarano di preferire il succo di frutta ipocalorico. Esiste una differenza significativa nelle preferenze per succhi tra maschi e femmine? Si scelga $\alpha = 0.05$

	Ipocalorico	Normale	Tot.
Maschi	10	40	50
Femmine	20	30	50
Tot.	30	70	100

Esercizio

Le ipotesi saranno:

$$H_0 : n^*_{ij} = \frac{n_{i.} * n_{.j}}{N}$$
$$H_1 : n^*_{ij} \neq \frac{n_{i.} * n_{.j}}{N}$$

La statistica-test utilizzata è:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

che, sotto l'ipotesi nulla, si distribuisce come un Chi-quadrato con (r-1)(c-1) gradi di libertà.
Pertanto, si **rifiuta** H_0 se:

$$X^2 > \chi^2_{\alpha;(r-1)(c-1)}$$

Esercizio

Tabella delle frequenze osservate

	Ipocalorico	Normale	Tot.
Maschi	10	40	50
Femmine	20	30	50
Tot.	30	70	100

$$n^*_{ij} = \frac{n_{i.} * n_{.j}}{N}$$

Tabella delle frequenze teoriche

	Ipocalorico	Normale	Tot.
Maschi	15	35	50
Femmine	15	35	50
Tot.	30	70	100

Esercizio

Tabella del chi-quadrato osservato

F_ osservate	F_teoriche	(oss-teo)	(oss-teo)²	(oss-teo)²/teo
10	15	-5	25	1,66
20	15	5	25	1,66
40	35	5	25	0,71
30	35	-5	25	0,71
				4,76

Il chi- quadrato tabulato è:

$$\chi^2_{0.05;1} = 3.841$$

Poichè:

$$4.76 > 3.841$$

Si rifiuta l'ipotesi nulla di indipendenza tra il genere e la preferenza per il tipo di succo di frutta

Introduzione

I test sulla bontà di adattamento vengono introdotti per verificare l'ipotesi che i dati campionari provengano da una v.c. la cui distribuzione è nota.
A differenza dei test precedenti, la verifica ora riguarda l'intera distribuzione e non i parametri che caratterizzano una certa famiglia di v.c.

Introduzione

Tra i test presenti in letteratura ricordiamo il **test del** χ^2 che è stato introdotto per l'adattamento di distribuzioni discrete (per fenomeni qualitativi e quantitativi) ma viene utilizzato anche per distribuzioni continue, raggruppando i dati in classi di modalità.

E' una **misura di discrepanza** delle frequenze osservate rispetto ad un modello probabilistico teorico ipotizzato con un quantile della v.c. Chi-quadrato.
Tale misura è basata sul **quadrato delle differenze normalizzate tra le frequenze osservate nel campione e quelle attese se fosse vera l'ipotesi distributiva specificata in** H_0.

Se è vera H_0 **le frequenze osservate dovrebbero essere molto simili alle corrispondenti frequenze attese**

Esercizio

Il capo del personale di una grande azienda vuole verificare se le assenze dal lavoro si distribuiscano in maniera uniforme nell'arco della settimana lavorativa (di 5 giorni). Per farlo, osserva le assenze registrate nell'ultimo mese e il giorno in cui queste si sono verificate, ottenendo i risultati riportati sotto. Cosa si può concludere sulla base dei dati?

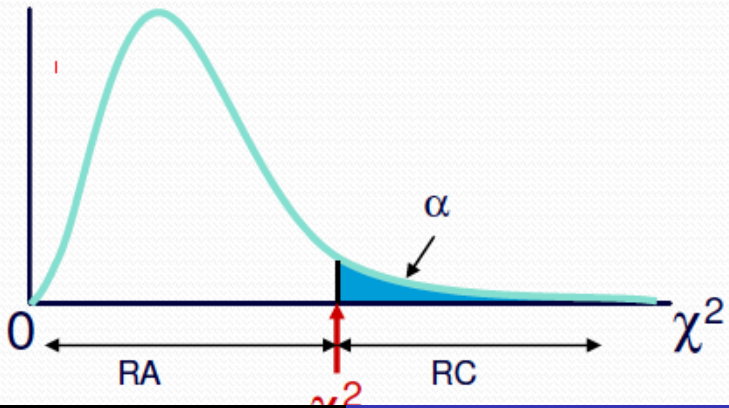
Giorno della settimana	F_osservate
1	3
2	1
3	0
4	2
5	4

Esercizio

Se è vera H_0 le frequenze osservate e teoriche dovrebbero essere molto simili, a tal fine per verificare l'ipotesi di adattamento si utilizza il seguente test:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \approx \chi^2_{(\alpha,k-1)}$$

$$RC(\alpha): \chi_c^2 > \chi^2_{(\alpha,k-1)}$$



Esercizio

Giorno della settimana	Frequenze osservate X_i	Frequenze relative teoriche p_i	Frequenze assolute teoriche np_i	$(X_i-np_i)^2/np_i$
1	3	0,2	2	0,5
2	1	0,2	2	0,5
3	0	0,2	2	2
4	2	0,2	2	0
5	4	0,2	2	2
	10	1	10	5

$$X \approx U(5) \rightarrow P(X = x_i) = p_i = \frac{1}{5} = 0.2$$

$$\chi^2_{0.05;4} = 9.5$$

$$\chi^2_{oss} < \chi^2_{0.05;1}$$

Non si rifiuta l'ipotesi nulla, quindi la distribuzione teorica si adatta alla distribuzione empirica