

# Esercitazioni di statistica

## Intervalli di confidenza

Stefania Spina

Università di Napoli Federico II

*stefania.spina@unina.it*

10 Dicembre 2014

## Intervallo di confidenza (Stima intervallare)

L'intervallo di confidenza o stima intervallare è un intervallo di valori con associata una probabilità o livello di confidenza  $1 - \alpha$ . La probabilità quantifica la possibilità che l'intervallo contenga il vero valore del parametro della popolazione.

L'intervallo di confidenza è uno stimatore del parametro  $\theta$  determinato in base ad un campione casuale  $(X_1, X_2, \dots, X_n)$  che produce un insieme plausibile di valori per il parametro  $\theta$

**Intervallo Casuale**

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha$$

$$P(T_n - k \leq \theta \leq T_n + k) = 1 - \alpha$$

dove

Variabili casuali  $\longrightarrow$   $L_1 = L_1(X_1, X_2, \dots, X_n)$

$$L_2 = L_2(X_1, X_2, \dots, X_n)$$

sono funzione del campione

## Il valore di $\alpha$

Nello scegliere  $\alpha$  si deve tener presente che quanto **minore** è, tanto più **ampio** risulta l'**intervallo di confidenza** ad esso associato e quindi **minore** è la **precisione** di valori appartenenti all'intervallo di stima per  $\theta$ .

Livello di confidenza	$\alpha$	$\alpha/2$	$Z_{\alpha/2}$
<b>0.90</b>	0.10	0.05	1.645
<b>0.95</b>	0.05	0.025	1.96
<b>0.98</b>	0.02	0.01	2.33
<b>0.99</b>	0.01	0.005	2.575

## Intervallo di confidenza per il parametro $\mu$ (quando $\sigma$ è noto)

$$X \sim N(\mu; \sigma^2)$$

### Prima dell'estrazione

#### Intervallo Casuale

$$P(L_1 \leq \mu \leq L_2) = 1 - \alpha$$

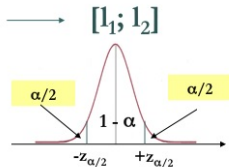
$$P\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

### Dopo l'estrazione

$$l_1 = \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$l_2 = \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

dove  $z$  è quel valore tale che l'area sottesa alla curva normale standardizzata tra  $-z$  e  $z$  è pari a  $(1 - \alpha)$ . Notiamo inoltre che  $z$  si lascia alla destra un'area pari a  $\alpha/2$  e che l'area sottesa alla curva normale standardizzata tra  $0$  e  $z$  è pari a  $(1-\alpha)/2$ .



## Esercizio

L'altezza delle matricole universitarie di sesso maschile può essere considerata una variabile con distribuzione Normale, con media incognita e varianza pari a 10.66.

Per stimare l'altezza media si estrae un campione casuale di 58 matricole e si misura l'altezza media, che risulta pari a 175.4 cm. Si definisca l'intervallo che, ad un livello di fiducia del 90, del 95 e del 99 per cento contenga il parametro incognito della popolazione.

# Esercizio

$$X \sim N(\mu; 10.66)$$

$$n = 58$$

$$\bar{x} = 175.4$$

$$1 - \alpha = \begin{cases} 0.90 \\ 0.95 \\ 0.99 \end{cases} \quad (1)$$

# Esercizio

1 caso :  $1 - \alpha = 0,90$

$$P\left(175.4 - 1.645 * \frac{3.265}{\sqrt{58}} \leq \mu \leq 175.4 + 1.645 * \frac{3.265}{\sqrt{58}}\right) = 0.90$$

$$P\left(175.4 - 0.705 \leq \mu \leq 175.4 + 0.705\right) = 0.90$$

$$P(174.70 \leq \mu \leq 176.11) = 0.90$$

## Esercizio

2 caso :  $1 - \alpha = 0.95$

$$P\left(175.4 - 1.96 * \frac{3.265}{\sqrt{58}} \leq \mu \leq 175.4 + 1.96 * \frac{3.265}{\sqrt{58}}\right) = 0.95$$

$$P\left(175.4 - 0.840 \leq \mu \leq 175.4 + 0.840\right) = 0.95$$

$$P(174.56 \leq \mu \leq 176.24) = 0.95$$

# Esercizio

3 caso :  $1 - \alpha = 0.99$

$$P\left(175.4 - 2.58 * \frac{3.265}{\sqrt{58}} \leq \mu \leq 175.4 + 2.58 * \frac{3.265}{\sqrt{58}}\right) = 0.99$$

$$P\left(175.4 - 1.106 \leq \mu \leq 175.4 + 1.106\right) = 0.99$$

$$P(174.294 \leq \mu \leq 176.506) = 0.99$$

# Esercizio

Un'azienda ha commissionato ad una società di software la riprogettazione del proprio sito web. Dopo che quest'ultimo è stato messo on-line vuole studiare gli effetti del rinnovo esaminando il numero di contatti giornaliero al sito. Sotto l'ipotesi che tali contatti hanno distribuzione Gaussiana,  $N(\mu, \sigma^2)$ , seleziona un campione casuale di 10 giorni in cui il numero di contatti è stato il seguente:

giorni	1	2	3	4	5	6	7	8	9	10
n. contatti	226	803	625	871	326	288	724	149	637	807

- 1 Costruire un intervallo di confidenza per la media dei contatti giornalieri della popolazione, fissando  $\alpha = 0.05$ .
- 2 Costruire un intervallo di confidenza per la media dei contatti giornalieri della popolazione, fissando  $\alpha = 0.02$ .

## Esercizio

L'intervallo di confidenza per la media della popolazione è dato da:

$$\bar{X} - t_{(n-1; \alpha/2)} * \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(n-1; \alpha/2)} * \frac{s}{\sqrt{n}}$$

$$P\left(\bar{X} - t_{(n-1; \alpha/2)} * \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(n-1; \alpha/2)} * \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

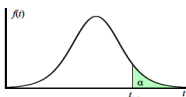
# Esercizio

Dai dati campionari emerge che:

$$\bar{x} = 545.6$$

$$s = 271$$

$$t_{(n-1; \alpha/2)} = t_{(9; 0.025)} = 2.262$$



n	$\alpha$				
	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.103	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.355	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.385	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845

## Esercizio

Avvalendosi di tali risultati l'intervallo di confidenza, con  $\alpha = 0.05$ , è:

$$\left[ 545.6 - 2.262 \frac{271}{\sqrt{10}}; 545.6 + 2.262 \frac{271}{\sqrt{10}} \right]$$

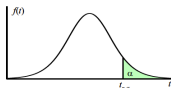
$$\left[ 351.75; 739.45 \right]$$

# Esercizio

Ad un livello di confidenza, con  $\alpha = 0.02$ , l'intervallo risulterà di ampiezza maggiore, infatti in questo caso  $t_{(9;0.01)} = 2.821$  e l'intervallo di confidenza è:

$$\left[ 545.6 - 2.821 \frac{271}{\sqrt{10}}; 545.6 + 2.821 \frac{271}{\sqrt{10}} \right]$$

$$\left[ 303.85; 787.35 \right]$$

Tavola 3 – Percentili della variabile casuale  $t$  di Student

$n$	$\alpha$				
	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.441	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.387	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012

# Introduzione

Per stimare la differenza tra le medie  $\mu_1$  e  $\mu_2$  di due popolazioni normali con varianze  $\sigma_1^2$  e  $\sigma_2^2$  note, si estrae da ognuna di esse un campione di dimensione, rispettivamente,  $n_1$  e  $n_2$ .

La differenza tra le medie campionarie  $\bar{X}_1 - \bar{X}_2$  costituisce un buon stimatore della differenza delle medie delle due popolazioni e, per campioni di elevata numerosità, si distribuisce in modo normale.

La formula:

$$P\left[(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

sta ad indicare che nel  $100(1 - \alpha)\%$  dell'universo dei campioni, la differenza tra le medie delle due popolazioni è compresa tra gli estremi:

$$\left[(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

## Esercizio 3

L'azienda Luci&lampi produce due tipi di lampadine, un tipo X, economico, e uno Y, più caro ma che dura di più.

La durata di entrambi i tipi di lampadine può essere considerata normodistribuita, con medie e varianze diverse.

Un'associazione di consumatori vuole stimare la differenza tra la durata media dei due tipi di lampadine per stabilire se questa giustifichi la differenza tra i prezzi.

L'associazione acquista dunque 50 lampadine del tipo X e 35 del tipo Y e le sottopone al test di durata, con i seguenti risultati (la media è espressa in ore):

$n_X=50$	$n_Y=35$
$\text{media}(X)=2850$	$\text{media}(Y)=4150$
$\text{varianza}(X)=32400$	$\text{varianza}(Y)=14000$

Si definisca l'intervallo per la differenza tra le medie dei due modelli di lampadina ad un livello di fiducia del 90%.

## Esercizio 3

$nX=50$	$nY=35$
media(X)=2850	media(Y)=4150
varianza(X)=32400	varianza(Y)=14000

$$P\left((\bar{X} - \bar{Y}) - z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \leq (\mu_x - \mu_y) \leq (\bar{X} - \bar{Y}) + z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right) = 1 - \alpha$$

$$P\left((2850 - 4150) - z_{\alpha/2} \sqrt{\frac{32400}{50} + \frac{14000}{35}} \leq (\mu_x - \mu_y) \leq (2850 - 4150) + z_{\alpha/2} \sqrt{\frac{32400}{50} + \frac{14000}{35}}\right) = 0.90$$

$$P\left((2850 - 4150) - 1.65 \sqrt{\frac{32400}{50} + \frac{14000}{35}} \leq (\mu_x - \mu_y) \leq (2850 - 4150) + 1.65 \sqrt{\frac{32400}{50} + \frac{14000}{35}}\right) = 0.90$$

La differenza tra la durata media dei due tipi di lampadine è compresa tra 1246 e 1353 ad un livello di fiducia del 90%

# Introduzione

Per stimare la differenza tra le medie  $\mu_1$  e  $\mu_2$  di due popolazioni normali con varianze ignote, si estrae da ognuna di esse un campione di dimensione, rispettivamente,  $n_1$  e  $n_2$ .

La differenza tra le medie campionarie  $\bar{X}_1 - \bar{X}_2$  costituisce un buon stimatore della differenza delle medie delle due popolazioni e, per campioni di elevata numerosità, si distribuisce in modo normale.

La stima della varianza delle due popolazioni si ottiene applicando la formula:

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

dove  $s_1^2$   $s_2^2$  sono le varianze corrette dei due campioni.

La formula:

$$P\left[(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2; (n_1+n_2-2)} * s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2; (n_1+n_2-2)} * s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha$$

sta ad indicare che nel  $100(1 - \alpha)\%$  dell'universo dei campioni, la differenza tra le medie delle due popolazioni è compresa tra gli estremi:

$$\left[(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2; (n_1+n_2-2)} * s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; (\bar{X}_1 + \bar{X}_2) + t_{\alpha/2; (n_1+n_2-2)} * s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

## Esercizio

Costruire un intervallo di confidenza al 90% per la differenza tra le medie di due popolazioni, servendosi dei seguenti dati tratti da due campioni:

$$\bar{X}_1 = 50 \quad n_1 = 14 \quad s_1 = 6$$

$$\bar{X}_2 = 35 \quad n_2 = 18 \quad s_2 = 9$$

## Esercizio

Per costruire l'intervallo di confidenza è necessario stimare la varianza della popolazione applicando la formula:

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

si ha:

$$s^2 = \frac{(14 - 1) * 6^2 + (18 - 1) * 9^2}{14 + 18 - 2} = 61.5$$

Inoltre, per  $\alpha = 0.1$  si ha che  $\alpha/2 = 0.05$  si desume che:

$$t_{0.05;30} = 1.697$$

# Esercizio

Dalla formula:

$$P\left[(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2; (n_1 + n_2 - 2)} * s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2; (n_1 + n_2 - 2)} * s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha$$

$$P\left[(50 - 35) - 1.697 * \sqrt{61.5} \sqrt{\frac{1}{14} + \frac{1}{18}} \leq (\mu_1 - \mu_2) \leq (50 - 35) + 1.697 * \sqrt{61.5} \sqrt{\frac{1}{14} + \frac{1}{18}}\right] = 0.90$$

Per cui, l'intervallo di confidenza è:

$$IC = [10.258; 19.742]$$

con un coefficiente di confidenza del 90%.

# Introduzione

Per stimare la proporzione  $p$  di una popolazione distribuita in modo normale, si estrae da essa un campione di numerosità  $n$  e si calcola la frequenza relativa campionaria

$\hat{p} = \frac{X}{n}$  che fornisce un buon stimatore del parametro della popolazione.

La scrittura:

$$P \left[ \hat{p} - z_{\alpha/2} * \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

sta ad indicare che nel  $100(1 - \alpha)\%$  dell'universo dei campioni, la proporzione  $p$  della popolazione è compresa tra gli estremi:

$$\left[ \hat{p} - z_{\alpha/2} * \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}; \hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

Volendo stimare la numerosità campionaria che tolleri un errore  $\epsilon$  si applica la formula:

$$n = \frac{z_{\alpha/2}^2 p(1 - p)}{\epsilon^2}$$

Stima puntuale del parametro  $p_{si}$ 

$$\hat{p}_{si} = \frac{35}{100} = 0,35$$

**X**= Guardare un programma  
televisivo (si, no)

$$n = 100$$

$$n_{si} = 35$$

Intervallo di confidenza per il parametro  $p_{si}$ 

$$\hat{p}_{si} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_{si} \cdot (1 - \hat{p}_{si})}{n}} = 0,35 \pm 1,96 \cdot \sqrt{\frac{0,35 \cdot (1 - 0,35)}{100}} =$$

$$= 0,35 \pm 1,96 \cdot 0,05$$

$$= 0,35 \pm 0,098$$

↓

$$[0,252; 0,448]$$

# Esercizio 1

Secondo un exit poll su persone che hanno votato all'elezione per sindaco, il 40% ha votato per il candidato Bianchi e il 60% per il candidato Rossi.

Ipotizzando che questo sia un campione casuale di tutti gli elettori, si definisca l'intervallo di confidenza che, ad un livello di fiducia del 99%, contenga la proporzione dei voti della popolazione per il candidato dei Bianchi nel caso in cui la dimensione campionaria sia:

- $n = 400$
- $n = 40$

Per ciascuna ampiezza indicare se sareste disposti a prevedere il vincitore.

# Esercizio 1

$$\hat{p} = 0.40$$

$$1 - \hat{p} = 0.60$$

$$n = 400$$

$$E(p) = p$$

$$\text{Var}(p) = \frac{p(1-p)}{n}$$

$$\hat{p} \approx N\left(p; \frac{p(1-p)}{n}\right) \text{ per grandi campioni}$$

$$P\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

$$P\left(0.40 - z_{\alpha/2} \sqrt{\frac{0.40 * 0.60}{400}} \leq p \leq 0.40 + z_{\alpha/2} \sqrt{\frac{0.40 * 0.60}{400}}\right) = 0.99$$

$$P\left(0.40 - 2.58 \sqrt{\frac{0.40 * 0.60}{400}} \leq p \leq 0.40 + 2.58 \sqrt{\frac{0.40 * 0.60}{400}}\right) = 0.99$$

$$P(0.34 \leq p \leq 0.46) = 0.99$$

# Esercizio 1

$$\hat{p} = 0.40$$

$$1 - \hat{p} = 0.60$$

$$n = 40$$

$$E(p) = p$$

$$\text{Var}(p) = \frac{p(1-p)}{n}$$

$$\hat{p} \approx N\left(p; \frac{p(1-p)}{n}\right) \text{ per grandi campioni}$$

$$P\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

$$P\left(0.40 - z_{\alpha/2} \sqrt{\frac{0.40 * 0.60}{40}} \leq p \leq 0.40 + z_{\alpha/2} \sqrt{\frac{0.40 * 0.60}{40}}\right) = 0.99$$

$$P\left(0.40 - 2.58 \sqrt{\frac{0.40 * 0.60}{40}} \leq p \leq 0.40 + 2.58 \sqrt{\frac{0.40 * 0.60}{40}}\right) = 0.99$$

$$P(0.20 \leq p \leq 0.60) = 0.99$$

# Esercizio 1

$$n = 40$$

$$P(0.20 \leq p \leq 0.60) = 0.99$$

Potremmo prevedere un vincitore in quanto l'intervallo contiene valori al di sotto e al di sopra di 0.50

$$n = 400$$

$$P(0.34 \leq p \leq 0.46) = 0.99$$

Si può prevedere la sconfitta di Bianchi in quanto l'intervallo contiene solo numeri al di sotto di 0.50, quelli cioè che indicano che Bianchi riceve una minoranza dei voti.

## Esercizio 2

Un programma televisivo è stato seguito dal 23% di un campione di 300 spettatori selezionati a caso.

Costruire un intervallo di confidenza al 98% per la proporzione della popolazione del pubblico televisivo in quella data fascia oraria.

## Esercizio 2

Per  $\alpha = 0.02$  si ha che  $\alpha/2 = 0.01$  e si desume che  $z_{0.01} = 2.33$

Quindi si avrà:

$$P\left[0.23 - 2.33 * \sqrt{\frac{0.23 * (1 - 0.23)}{300}} \leq p \leq 0.23 + 2.33 * \sqrt{\frac{0.23 * (1 - 0.23)}{300}}\right] = 0.98$$

Per cui l'intervallo di confidenza è:

$$IC = [0.173; 0.287]$$

con un livello di confidenza del 98%.

# Introduzione

Per stimare la differenza tra le proporzioni  $p_1$  e  $p_2$  di due popolazioni normali, si estrae da ognuna di esse un campione di dimensione, rispettivamente,  $n_1$  e  $n_2$ .

La differenza tra le proporzioni campionarie  $\hat{p}_1 - \hat{p}_2$  costituisce un buon stimatore della differenza tra le medie delle due popolazioni e, per campioni di elevata numerosità, si distribuisce in modo normale.

$$P \left[ (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \leq (p_1 - p_2) \leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right] = 1 - \alpha$$

sta ad indicare che nel  $100(1 - \alpha)\%$  dell'universo dei campioni, la differenza tra le proporzioni delle due popolazioni è compresa tra gli estremi:

$$\left[ (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}; \left[ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right] \right]$$

## Esercizio

Alcuni ricercatori hanno riscontrato che su 100 uomini trattati con farmaco 65 sono guariti da una data malattia, mentre su 120 donne trattate con lo stesso farmaco, 45 sono guarite dalla malattia. Costruire un intervallo di confidenza al 90% per la differenza tra le proporzioni delle due popolazioni ipotizzate normali.

# Esercizio

$$\hat{p}_m = \frac{65}{100} = 0.65$$

$$\hat{p}_f = \frac{45}{120} = 0.375$$

$$n_m = 100$$

$$n_f = 120$$

$$P \left[ (\hat{p}_m - \hat{p}_f) - z_{\alpha/2} \sqrt{\frac{\hat{p}_m(1 - \hat{p}_m)}{n_m} + \frac{\hat{p}_f(1 - \hat{p}_f)}{n_f}} \leq (p_m - p_f) \leq (\hat{p}_m - \hat{p}_f) + z_{\alpha/2} \sqrt{\frac{\hat{p}_m(1 - \hat{p}_m)}{n_m} + \frac{\hat{p}_f(1 - \hat{p}_f)}{n_f}} \right] = 1 - \alpha$$

$$\left[ (0.65 - 0.375) - 1.65 \sqrt{\frac{0.65(1 - 0.65)}{100} + \frac{0.375(1 - 0.375)}{120}} \leq (p_m - p_f) \leq (0.65 - 0.375) \right.$$

$$\left. + 1.65 \sqrt{\frac{0.65(1 - 0.65)}{100} + \frac{0.375(1 - 0.375)}{120}} \right]$$

$$IC[0.168; 0.382]$$

# Introduzione

Data l'equazione della retta stimata di regressione

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

dove  $\hat{\beta}_0$  è l'intercetta e  $\hat{\beta}_1$  è il coefficiente angolare, effettuiamo le stime puntuali dei parametri utilizzando il metodo dei minimi quadrati:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\text{Cov}(XY)}{\text{Var}(X)}$$

dove  $\bar{x}$  e  $\bar{y}$  sono i valori medi, rispettivamente, del campione X e Y.

# Introduzione

Per la costruzione dell'intervallo di confidenza si utilizza uno stimatore che si distribuisce come una v.c t di Student con  $n-2$  gradi di libertà. Gli intervalli di confidenza per  $\beta_0$  e  $\beta_1$  sono rispettivamente:

$$\hat{\beta}_0 - t_{\alpha/2; n-2} * es(\beta_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2; n-2} * es(\beta_0)$$

$$\hat{\beta}_1 - t_{\alpha/2; n-2} * es(\beta_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2; n-2} * es(\beta_1)$$

# Introduzione

In cui:

$$es(\beta_0) = s * \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

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

e, in cui, l'errore standard della regressione é dato da:

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

# Esercizio

Volendo costruire un modello che spieghi il Peso (espresso in kg) in funzione dell'Altezza (espressa in cm) si è osservato un campione di  $n=10$  studenti della facoltà di Economia; i dati ottenuti sono riportati nella tabella seguente:

Altezza	Peso
165	71
172	75
159	81
168	76
166	88
158	72
157	98
177	89
164	83
172	81

## Esercizio

- Stimare, con il metodo dei minimi quadrati, i coefficienti  $\beta_0$  e  $\beta_1$  del modello di regressione.

Supponendo che la variabile peso (Y) si distribuisca normalmente e soddisfi le ipotesi di indipendenza, omoschedasticità (omogeneità delle varianze) e che le medie siano sulla retta di regressione, Costruire intervalli di confidenza, ad un livello di significatività  $\alpha = 0.05$ , per:

- l'intercetta  $\beta_0$
- il coefficiente angolare  $\beta_1$

## Esercizio

Le stime dei minimi quadrati per i parametri  $\beta_0$  e  $\beta_1$  si ottengono considerando questi calcoli:

	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
	165	71	27.225	5.041	11.715
	172	75	29.584	5.625	12.900
	159	81	25.281	6.561	12.879
	168	76	28.224	5.776	12.768
	166	88	27.556	7.744	14.608
	158	72	24.964	5.184	11.376
	157	98	24.649	9.604	15.386
	177	89	31.329	7.921	15.753
	164	83	26.896	6.889	13.612
	172	81	29.584	6.561	13.932
<b>Totale</b>	<b>1.658</b>	<b>814</b>	<b>275.292</b>	<b>66.906</b>	<b>134.929</b>

# Esercizio

$$\bar{x} = \frac{1658}{10} = 165.8$$

$$\bar{y} = \frac{814}{10} = 81.4$$

$$\hat{\beta}_1 = \frac{\text{cov}(XY)}{\text{var}(X)} = \frac{E(XY) - \bar{x} * \bar{y}}{E(X^2) * (\bar{X})^2} = \frac{\frac{134929}{10} - 165.8 * 81.4}{\frac{275292}{10} - (165.8)^2} = -0.081395$$

$$\hat{\beta}_0 = 81.4 - (-0.081395) * 165.8 = 94.895291$$

La retta di regressione è, quindi:

$$\hat{y} = 94.895291 - 0.081395x$$

## Esercizio

Per la determinazione degli intervalli di confidenza i calcoli sono ottenuti nello schema seguente:

	$x_i$	$(x_i - \bar{x})^2$	$y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
	165	0,64	71	81,46512	109,5187
	172	38,44	75	80,89535	34,75516
	159	46,24	81	81,95349	0,909136
	168	4,84	76	81,22093	27,25812
	166	0,04	88	81,38372	43,77515
	158	60,84	72	82,03488	100,6988
	157	77,44	98	82,11628	252,2927
	177	125,44	89	80,48838	72,44774
	164	3,24	83	81,54651	2,11263
	172	38,44	81	80,89535	0,010951
<b>Totale</b>	<b>1.658</b>	<b>395,6</b>	<b>814</b>	<b>814</b>	<b>643,779</b>

# Esercizio

L'errore standard della regressione é dato da:

$$s = \sqrt{\frac{1}{n-2} * \sum_{i=1}^n (y_i - \hat{y})^2} = \sqrt{\frac{1}{n-2} * 643.779} = 8.97064$$

## Esercizio

Per l'intervallo di confidenza per  $\beta_0$

$$\hat{\beta}_0 - t_{\alpha/2; n-2} * es(\beta_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2; n-2} * es(\beta_0)$$

è necessario calcolare:

$$es(\beta_0) = s * \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = 8.97064 * \sqrt{\frac{1}{10} + \frac{(165.8)^2}{395.6}} = 74.832815$$

Per  $\alpha/2 = 0.025$  e  $n-2 = 8$  gradi di libertà è:

$$t_{0.025; 8} = 2.306$$

Pertanto, l'intervallo di confidenza per  $\beta_0$  è:

$$94.895291 - 2.306 * 74.832815 \leq \beta_0 \leq 94.895291 + 2.306 * 74.832815$$

ossia:

$$IC = [-77.66918; 267.45976]$$

## Esercizio

Per l'intervallo di confidenza per  $\beta_1$

$$\hat{\beta}_1 - t_{\alpha/2; n-2} * es(\beta_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2; n-2} * es(\beta_1)$$

L'errore standard:

$$es(\beta_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{8.97064}{\sqrt{395.6}} = 0.451019$$

Pertanto, l'intervallo di confidenza per  $\beta_1$  è:

$$-0.081395 - 2.306 * 0.451019 \leq \beta_1 \leq -0.081395 + 2.306 * 0.451019$$

ossia:

$$IC = [-1.12144; 0.95865]$$

con una probabilità dello 0.95