Intervalli di Confidenza

Corso di **Teoria dell'Inferenza Statistica**

a.a. 2009/2010 Secondo Periodo

Prof. Filippo DOMMA

Corso di Laurea in MQEGA

(Metodi Quantitativi per l'Economia e la Gestione delle Aziende)

Facoltà di Economia – UniCal

I metodi di stima puntuali anche se corredati di tutte le proprietà giudicate desiderabili e ottimali, difficilmente potranno fornire delle stime che coincidono con il parametro incognito, poiché ci si dovrà sempre attendere un certo errore di campionamento.
Nasce, quindi, l'esigenza di associare allo stimatore una misura dell'errore di stima commesso, in modo tale da valutare quanto la stima sia da considerarsi "vicina" al parametro incognito.
Tali valutazioni possono essere fatte facendo riferimento alla dispersione della distribuzione campionaria dello stimatore $T(\mathbf{X})$.
In relazione alla stima $t(\mathbf{x})$, ottenuta tramite il campione osservato, e alla precisione dello stimatore, ci saranno dei valori di θ che sulla base del campione debbono essere considerati più plausibili di altri.
Definito, quindi, il grado di plausibilità si potrà dividere lo spazio parametrico in due sottoinsiemi:

uno di valori probabili per θ secondo il grado di plausibilità fissato ed un altro di valori poco probabili per θ. Così, invece di stimare un unico valore per θ, si stimerà un insieme di valori possibili a cui verrà associato il grado di plausibilità scelto il quale deve essere interpretato come livello di confidenza per l'insieme.

Sia **X** un c.c. estratto da f(x;θ) appartenente alla famiglia di distribuzioni **P**. Diamo la seguente

Definizione. La famiglia di intervalli S(**X**) di Θ, funzione di **X** ma non di θ, è chiamato intervallo casuale.

S(**X**) è del tipo

$$\left\{\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X})\right\}$$

dove $\underline{\theta}(\mathbf{X})$ e $\overline{\theta}(\mathbf{X})$

sono, rispettivamente, il limite inferiore e superiore dell'intervallo casuale.

Definizione. Per un dato valore di α (usualmente piccolo), $0< \alpha <1$,
un intervallo di confidenza al $100(1-\alpha)\%$ per θ
è la realizzazione di un intervallo casuale tale che

$$P_r\left\{\underline{\theta}(X) \le \theta \le \overline{\theta}(X)\right\} \ge 1-\alpha \qquad \forall \theta \in \Theta$$

La quantità

$$\inf_{\theta \in \Theta} P_r\left\{\underline{\theta}(X) \le \theta \le \overline{\theta}(X)\right\}$$

in genere è uguale ad $1-\alpha$, è chiamato
coefficiente fiduciario dell'intervallo casuale.

Un metodo generale per la costruzione di intervalli di confidenza è costituito dalla ricerca di una funzione del campione e del parametro incognito da stimare che abbia una distribuzione indipendente da parametro stesso.

Metodo della Quantità Pivot

Definizione. Quantità Pivot.

Sia $X_1,\ldots,X_n$ un c.c. estratto da una fd (o fp) $f(x;\theta)$ appartenente a **P**.

Sia $Q=q(X_1,\ldots,X_n;\theta)$ una funzione del c.c. e del parametro incognito θ. Se la v.c. Q ha distribuzione indipendente da θ, allora Q è detta **quantità pivot**.

Se $Q=q(\mathbf{X};\theta)$ è una quantità pivot, allora per ogni α fissato, $0<\alpha<1$, esisteranno due valori q_1 e q_2 dipendenti da α, tali che

$$P_r\{q_1 < Q < q_2\} = 1-\alpha$$

Per ogni c.o. $x_1,\ldots,x_n$, si ha:

$$q_1 < q(x_1,\ldots,x_n;\theta) < q_2$$

Ora, se da questa doppia diseguaglianza, riusciamo a calcolare la seguente:

$$t_1(x_1,\ldots,x_n)<\theta<t_2(x_1,\ldots,x_n)$$

Per funzioni t_1 e t_2 indipendenti da θ, allora (t_1,t_2) è un intervallo fiduciario (di confidenza) al $100(1-\alpha)\%$ per θ.

In tal modo costruiamo un I.C. per θ in due fasi:

1ª - individuare la q.p.;

2ª - invertire la doppia diseguaglianza in termini di θ.

Campionamento da popolazioni Normali

Sia **X** un c.c. iid estratto da

$$P = \left\{ N(.,.) : (\mu, \sigma^2) \in \Re \times \Re^+ \setminus \{0\} \right\}$$

Costruire un I.C. per μ con il metodo della quantità pivot.

Esistono due casi distinti:

a) varianza nota;

b) varianza sconosciuta.

A) Varianza nota

$$P = \left\{ N(.,.) : (\mu, \sigma_0^2) \in \Re \times \Re^+ \setminus \{0\} \right\}$$

1) Individuazione della Quantità Pivot.

Per individuare la q.p. dobbiamo costruire una funzione del c.c. **X** e del parametro incognito (μ) con fd (o fp) indipendente dal parametro incognito.

Sia **X** un c.c. iid estratto da N(.;.). Sappiamo che, in tale contesto, la media campionaria ha la seguente distribuzione

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma_0^2}{n} \right)$$

E' evidente che la media campionaria non è una quantità pivot.

Consideriamo la standardizzazione della media campionaria, cioè:

$$Z = \frac{\overline{X} - \mu}{\sigma_0 / \sqrt{n}} \sim N(0,1)$$

Z è una quantità pivot per μ; infatti, si ha:

1) è funzione del c.c. **X** e del parametro incognito μ, cioè $Z = q(\mathbf{X}; \mu)$

2) *Z* ha distribuzione indipendente dal parametro incognito μ.

Fissato α, con $0 < \alpha < 1$, possiamo trovare due valori, q_1 e q_2, tali che

$$P_r\{q_1 < Z < q_2\} = 1 - \alpha$$

2) Inversione della doppia diseguaglianza in termini di μ;

$$1-\alpha = P_r\{q_1 < Z < q_2\} = P_r\{q_1 < q(\mathbf{X};\mu) < q_2\} =$$

$$= P_r\left\{ q_1 < \frac{\overline{X} - \mu}{\sigma_0 / \sqrt{n}} < q_2 \right\} =$$

$$= P_r\left\{ q_1 \sigma_0 / \sqrt{n} < \overline{X} - \mu < q_2 \sigma_0 / \sqrt{n} \right\} =$$

$$= P_r\left\{ -\overline{X} + q_1 \sigma_0 / \sqrt{n} < -\mu < -\overline{X} + q_2 \sigma_0 / \sqrt{n} \right\} =$$

$$= P_r\left\{ \overline{X} - q_1 \sigma_0 / \sqrt{n} > \mu > \overline{X} - q_2 \sigma_0 / \sqrt{n} \right\} =$$

$$= P_r\left\{ \overline{X} - q_2 \sigma_0 / \sqrt{n} < \mu < \overline{X} - q_1 \sigma_0 / \sqrt{n} \right\}$$

La determinazione di q_1 e q_2 dipende da α e dalla fd (o fp) della q.p.

Generalmente, α viene fissato ad un valore molto basso 0.01, 0.05.

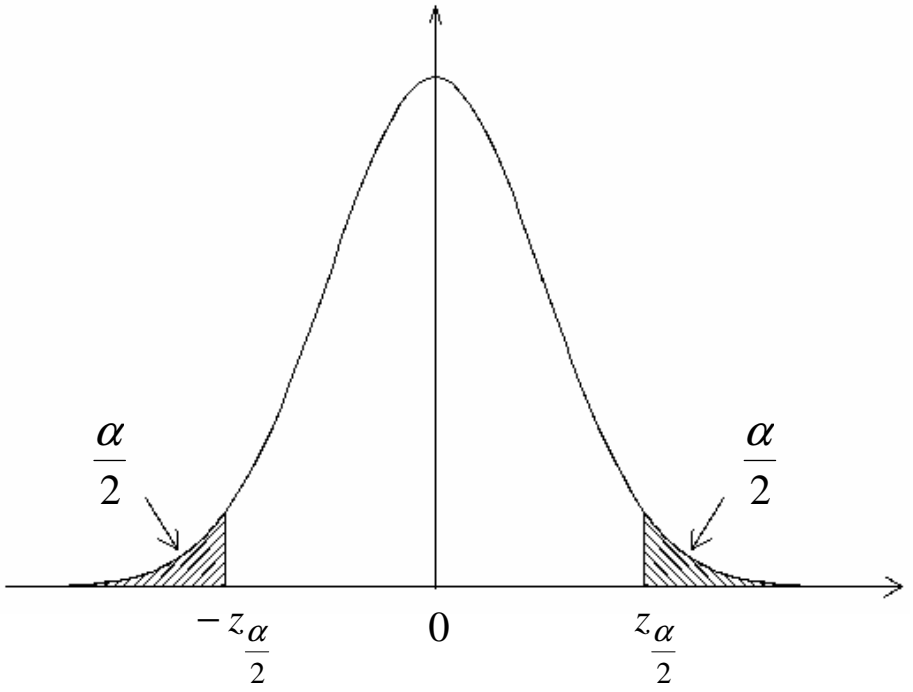
Ripartiamo in parti uguali, sulle code della normale standardizzata, α in modo tale che

$$P_r\{Z < q_1\} = \frac{\alpha}{2} \Rightarrow$$

$$\Rightarrow q_1 = -z_{\frac{\alpha}{2}}$$

$$P_r\{Z > q_2\} = \frac{\alpha}{2} \Rightarrow$$

$$\Rightarrow q_2 = z_{\frac{\alpha}{2}}$$



Sostituendo si ha:

$$P_r\left\{\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}<\mu<\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right\}=1-\alpha$$

Gli estremi dell'intervallo casuale (T_1,T_2), sono:

$$T_1=\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}} \qquad T_2=\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}$$

Osservato il c.c. $\mathbf{x}=(x_1,\ldots,x_n)$, l'Intervallo di confidenza al $100(1-\alpha)\%$ per μ è:

$$\left\{\overline{x}-z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\ ,\ \overline{x}+z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right\}$$

Il termine fiduciario nasce dalla seguente osservazione:
se estraessimo dalla popolazione ripetutamente campioni di dimensione n e se calcolassimo per ognuno di questi l'intervallo (t_1,t_2), la frequenza relativa di intervalli che contengono θ tenderebbe al $100(1-\alpha)\%$. Abbiamo, quindi, una considerevole fiducia che l'intervallo osservato contenga θ. La misura della nostra fiducia è $100(1-\alpha)\%$.

Esempio.

Sia $(X_1,....X_9)$ un c.c. estratto da una popolazione statistica che è ben adattata da una v.c. Normale con media incognita e varianza pari a 25. Utilizzando le realizzazioni finite delle variabili casuali componenti il campione, abbiamo calcolato il valore della media campionaria, pari a 27. Determinare l'intervallo di confidenza per la media della popolazione al livello di confidenza pari al 97%.

B) Varianza sconosciuta

$$P = \left\{ N(.,.) : (\mu, \sigma^2) \in \Re \times \Re^+ \setminus \{0\} \right\}$$

1) Individuazione della Quantità Pivot.

Per individuare la q.p. dobbiamo costruire una funzione del c.c. **X** e del parametro incognito (μ) con fd (o fp) indipendente dal parametro incognito.

Dire perché la quantità

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

non è utile per costruire un I.C. per μ, con σ sconosciuta.

Osservazione:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Si dimostra, inoltre, che:

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Sappiamo che:

$$T = \frac{Z}{\sqrt{\frac{V}{(n-1)}}} \sim t(n-1)$$

In definitiva, la v.c. T ha distribuzione indipendente da parametri incogniti (dipende solo dai gradi di libertà n-1).

Si osservi, ora, che

$$T = \frac{Z}{\sqrt{\frac{V}{(n-1)}}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{(n-1)}}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{S} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

In definitiva, abbiamo

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

Quest'ultima quantità è una q.p.; infatti, T è funzione di **X** e di μ, con f.d. indipendente da parametri incogniti. Individuata la q.p., fissato α, possiamo trovare q_1 e q_2 tali che

$$P_r\{q_1 < T < q_2\} = 1-\alpha$$

2) Inversione della doppia diseguaglianza in termini di μ;

$$1-\alpha = P_r\{q_1 < T < q_2\} = P_r\{q_1 < q(\mathbf{X};\mu) < q_2\} =$$

$$= P_r\left\{q_1 < \frac{\overline{X}-\mu}{S/\sqrt{n}} < q_2\right\} =$$

$$= P_r\left\{q_1 S/\sqrt{n} < \overline{X} - \mu < q_2 S/\sqrt{n}\right\} =$$

$$= P_r\left\{-\overline{X} + q_1 S/\sqrt{n} < -\mu < -\overline{X} + q_2 S/\sqrt{n}\right\} =$$

$$= P_r\left\{\overline{X} - q_1 S/\sqrt{n} > \mu > \overline{X} - q_2 S/\sqrt{n}\right\} =$$

$$= P_r\left\{\overline{X} - q_2 S/\sqrt{n} < \mu < \overline{X} - q_1 S/\sqrt{n}\right\}$$

La determinazione di q_1 e q_2 dipende da α e dalla fd (o fp) della q.p.

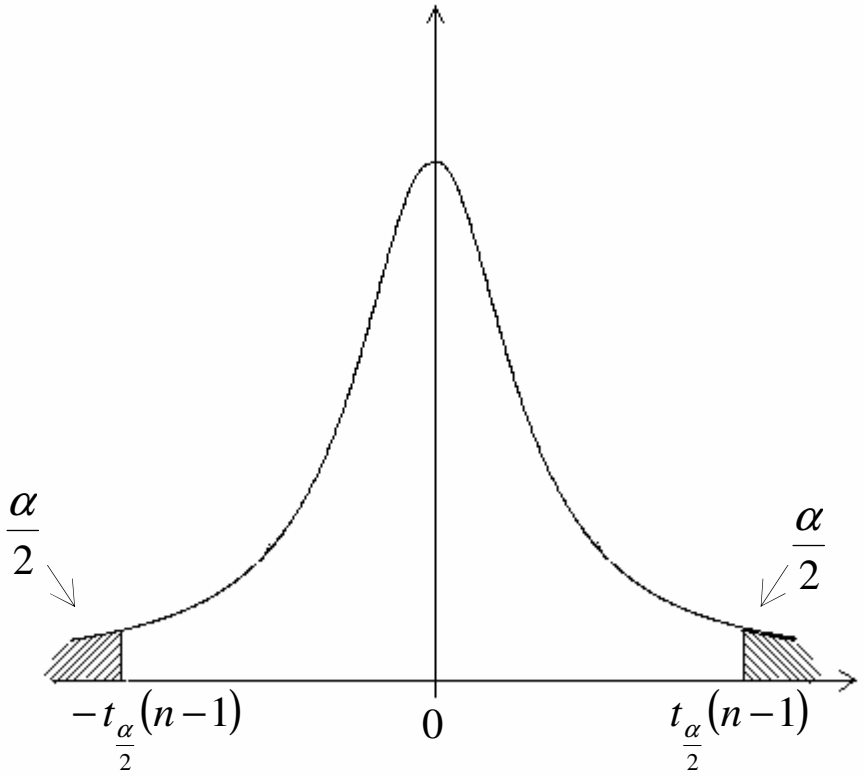
Ripartiamo in parti uguali, sulle code della t di Student, α in modo tale che

$$P_r\{T < q_1\} = \frac{\alpha}{2} \Rightarrow$$

$$\Rightarrow q_1 = -t_{\frac{\alpha}{2}}(n-1)$$

$$P_r\{T > q_2\} = \frac{\alpha}{2} \Rightarrow$$

$$\Rightarrow q_2 = t_{\frac{\alpha}{2}}(n-1)$$



Sostituendo si ha:

$$P_r\left\{\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)<\mu<\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

Gli estremi dell'intervallo casuale (T_1,T_2), sono:

$$T_1=\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1) \qquad T_2=\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)$$

Osservato il c.c. $\mathbf{x}=(x_1,\ldots,x_n)$, l'Intervallo di confidenza al $100(1-\alpha)\%$ per μ è:

$$\left\{\overline{x}-\frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\ ,\ \overline{x}+\frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right\}$$

Esempio.

In un grande comune rurale, da un'indagine campionaria su 900 famiglie, è risultato un reddito medio di 2.3 milioni ed una deviazione standard di 0.8 milioni. Si determini l'intervallo di confidenza per il reddito medio annuo di tutte le famiglie, sotto l'ipotesi che lo stesso segua una v.c. Normale, al livello di confidenza del 95%.

Osservazione - 1

Dato l'intervallo

$$\left\{ \bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} \;,\; \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} \right\}$$

La lunghezza (L) dell'intervallo di confidenza è definita come differenza tra gli estremi dell'intervallo stesso, cioè

$$L = \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} - \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} = 2 z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}$$

Si definisce errore la lunghezza dell'intervallo diviso 2, cioè

$$\varepsilon = \frac{L}{2} = z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}$$

- a parità di α:
  - L diminuisce al diminuire di σ;
  - L diminuisce all'aumentare di n.

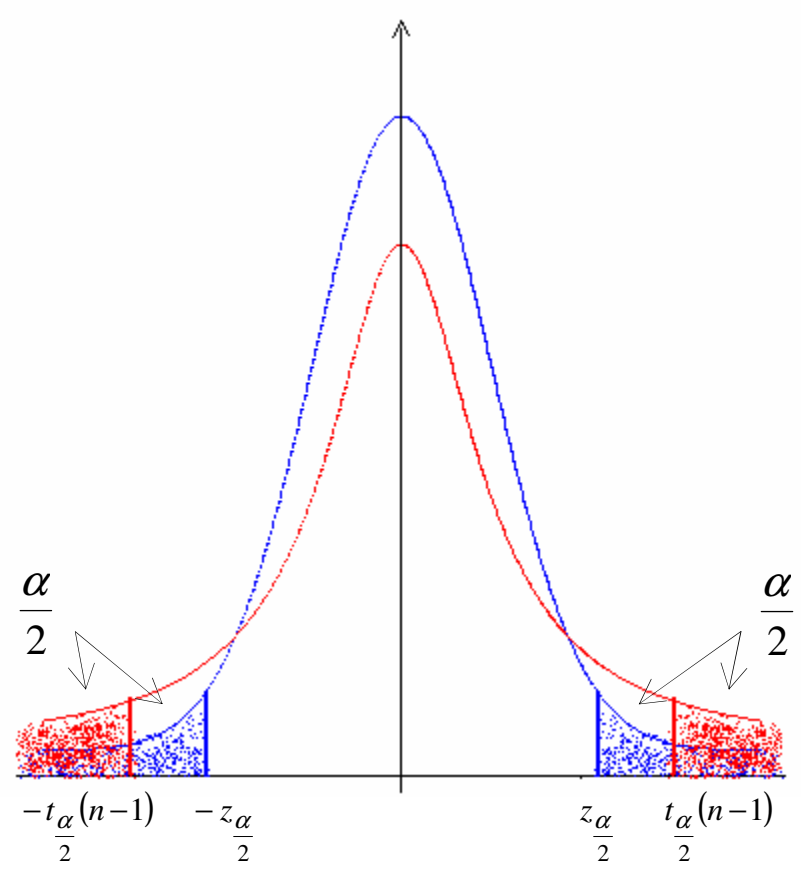
- a parità di n e σ:
  - L diminuisce all'aumentare di α

  [in tal caso diminuisce il grado di fiducia $(1-\alpha)$]

Situazione ottima: L piccolo - $(1-\alpha)$ elevato

Osservazione - 2

Fissato α, a parità di σ e s e della dimensione campionaria n, gli intervalli di confidenza per la media della popolazione costruiti con T sono più ampi.



Osservazione - 3

In alcuni casi, è necessario calcolare la dimensione campionaria minima affinché l'I.C. abbia una lunghezza prefissata.

Così,ad esempio, nel caso di I.C. per μ con σ noto, si ha:

$$n: \; L \le \ell \;\Rightarrow 2z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} \le \ell \quad \Rightarrow 2z_{\frac{\alpha}{2}} \frac{\sigma_0}{\ell} \le \sqrt{n}$$

$$\Rightarrow \; n \ge \left(2z_{\frac{\alpha}{2}} \frac{\sigma_0}{\ell} \right)^2$$

Esempio

Le uova prodotte in una azienda agricola hanno un peso che si distribuisce secondo una normale con media μ incognita e varianza paria a 49. Determinare la dimensione del campione che consente di stimare μ, mediante la media campionaria, con un errore non superiore a 4 con una probabilità di 0.95.

Campionamento da popolazioni Normali

Sia **X** un c.c. iid estratto da

$$P = \left\{ N(.,.) : (\mu, \sigma^2) \in \Re \times \Re^+ \setminus \{0\} \right\}$$

Costruire un I.C. per σ^2 con il metodo della quantità pivot.

Esistono due casi distinti:

a) μ sconosciuta;

b) μ nota.

A) Media sconosciuta

1) Individuazione della Quantità Pivot.

Per individuare la q.p. dobbiamo costruire una funzione del c.c. **X** e del parametro incognito (σ^2) con fd (o fp) indipendente dal parametro incognito.

Si è visto in precedenza che la quantità

$$V = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2(n-1)$$

E' evidente che V è una funzione del c.c. **X** e del parametro incognito σ^2 ; inoltre, si distribuisce secondo una chi-quadrato con (n-1) gradi di libertà ovvero la distribuzione non dipende da parametri incogniti. Da ciò si può concludere che V è una quantità pivot per σ^2.

2) Inversione della doppia diseguaglianza in termini di σ^2;

Fissato α, possiamo determinare q_1 e q_2 tali che

$$1-\alpha = P_r\{q_1 < V < q_2\} = P_r\{q_1 < q(\mathbf{X};\sigma^2) < q_2\} =$$

$$= P_r\left\{q_1 < \frac{(n-1)S^2}{\sigma^2} < q_2\right\} = P_r\left\{\frac{1}{q_1} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{q_2}\right\} =$$

$$= P_r\left\{\frac{(n-1)S^2}{q_1} > \sigma^2 > \frac{(n-1)S^2}{q_2}\right\} =$$

$$= P_r\left\{\frac{(n-1)S^2}{q_2} < \sigma^2 < \frac{(n-1)S^2}{q_1}\right\}$$

La determinazione di q_1 e q_2 dipende da α e dalla fd (o fp) della q.p.

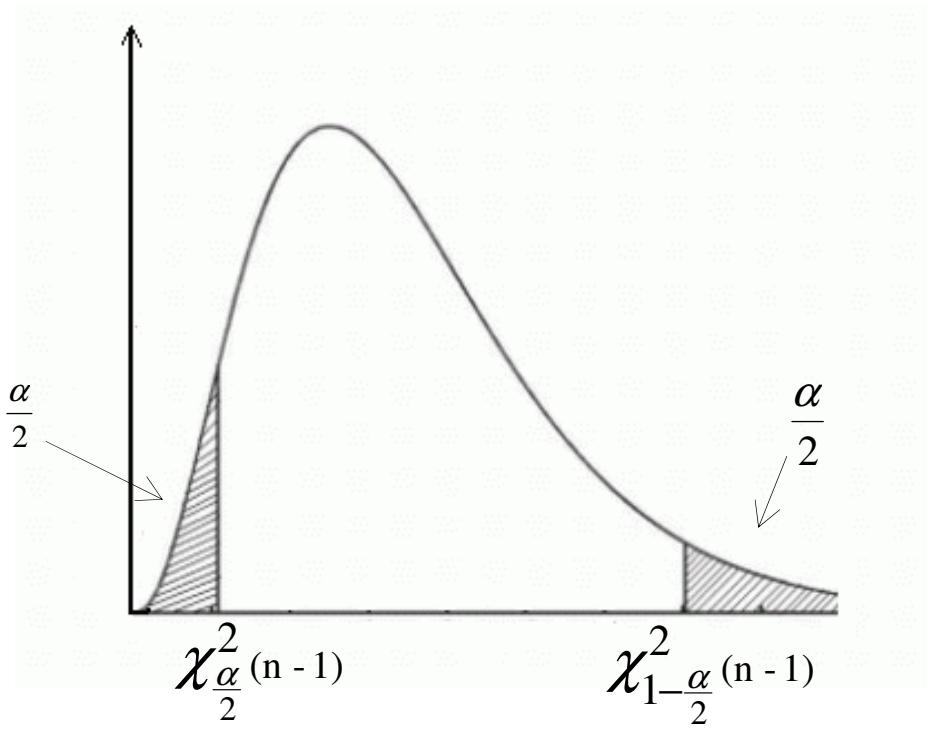
Ripartiamo in parti uguali, sulle code della Chi-quadrato, α in modo tale che

$$P_r\{V < q_1\} = \frac{\alpha}{2} \Rightarrow$$

$$\Rightarrow q_1 = \chi^2_{\frac{\alpha}{2}}(n-1)$$

$$P_r\{V > q_2\} = \frac{\alpha}{2} \Rightarrow$$

$$\Rightarrow q_2 = \chi^2_{1-\frac{\alpha}{2}}(n-1)$$



Sostituendo si ha:

$$P_r\left\{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right\} = 1-\alpha$$

Gli estremi dell'intervallo casuale (T_1,T_2), sono:

$$T_1 = \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \qquad T_2 = \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$$

Osservato il c.c. $\mathbf{x}=(x_1,\ldots,x_n)$, l'Intervallo di confidenza al $100(1-\alpha)\%$ per σ^2 è:

$$\left\{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right\}$$

B) Media nota

1) Individuazione della Quantità Pivot.

Per individuare la q.p. dobbiamo costruire una funzione del c.c. **X** e del parametro incognito (σ^2) con fd (o fp) indipendente dal parametro incognito.

$$V = \frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$$

E' evidente che V è una funzione del c.c. **X** e del parametro incognito σ^2 ; inoltre, si distribuisce secondo una chi-quadrato con (n) gradi di libertà ovvero la distribuzione non dipende da parametri incogniti. Da ciò si può concludere che V è una quantità pivot per σ^2 .

Utilizzando lo stesso procedimento del caso (A), si ottiene l'I.C. per σ^2, cioè

$$P_r\left\{\frac{n\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} < \sigma^2 < \frac{n\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right\} = 1-\alpha$$

Gli estremi dell'intervallo casuale (T_1,T_2), sono:

$$T_1 = \frac{n\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} \qquad T_2 = \frac{n\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n)}$$

Osservato il c.c. $\mathbf{x}=(x_1,\ldots,x_n)$, l'Intervallo di confidenza al $100(1-\alpha)\%$ per σ^2 è:

$$\left\{\frac{n\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}\ ,\ \frac{n\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right\}$$

Esempio.

Il diametro delle sfere di acciaio, prodotte da una determinata industria, è adattato statisticamente da una v.c. X che si distribuisce come una Normale. Si effettua un campionamento casuale di numerosità 9 e si misura il diametro delle sfere costituenti il campione. I risultati, realizzazioni delle v.c. componenti il campione, sono i seguenti :
20.1 , 19.9 , 20 , 19.8 , 19.7 , 20.2 , 20.1 , 23.1 , 22.8
Determinare l'intervallo di confidenza al livello del 90% per il valor medio della popolazione ed un altro intervallo allo stesso livello di confidenza per la varianza della popolazione.

Intervallo di Confidenza per la differenza tra le medie di due popolazioni Normali.

Siano X ed Y due v.c. indipendenti e normalmente distribuite, cioè

$$
\begin{array}{ccc}
X \sim N(\mu_x, \sigma_x^2) & \perp & Y \sim N(\mu_y, \sigma_y^2) \\
\downarrow & & \downarrow \\
\mathbf{X}_{(m\times 1)} & \perp & \mathbf{Y}_{(n\times 1)} \\
\downarrow & & \downarrow \\
\overline{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{m}\right) & \perp & \overline{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n}\right)
\end{array}
$$

Vogliamo costruire un I.C. per la differenza tra le medie μ_x e μ_y.

Primo Caso: σ_x^2 e σ_y^2 Note

Stimatore naturale della differenza tra le medie

$$D = \overline{X} - \overline{Y}$$

E' semplice verificare che

$$E[D] = \mu_x - \mu_y$$

$$V[D] = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$$

$$D \sim N\left(\mu_x - \mu_y \ , \ \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right)$$

La v.c.

$$Z = \frac{D - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} \sim N(0,1)$$

E' una quantità pivot perché è funzione del c.c. (**X**,**Y**) e del parametro incognito (μ_x-μ_y) ed ha distribuzione indipendente da parametri incogniti.

Fissato α, possiamo determinare q_1 e q_2 tali che

$$1 - \alpha = P_r\{q_1 < Z < q_2\} =$$

$$= P_r\left\{ q_1 < \frac{D - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} < q_2 \right\} =$$

$$= P_r\left\{ q_1 \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} < D - (\mu_x - \mu_y) < q_2 \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \right\} =$$

$$= P_r\left\{ -D + q_1 \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} < -(\mu_x - \mu_y) < -D + q_2 \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \right\} =$$

$$= P_r\left\{ D - q_1\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}} > (\mu_x - \mu_y) > D - q_2\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}} \right\} =$$

$$= P_r\left\{ D - q_2\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}} < (\mu_x - \mu_y) < D - q_1\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}} \right\}$$

Ricordando che

$$P_r\{Z < q_1\} = \frac{\alpha}{2} = P_r\{Z > q_2\} \quad \Rightarrow \quad q_1 = -z_{\frac{\alpha}{2}} \qquad q_2 = z_{\frac{\alpha}{2}}$$

Si ottiene:

$$P_r\left\{D-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}<(\mu_x-\mu_y)<D+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}\right\}=1-\alpha$$

Gli estremi dell'intervallo casuale (T_1,T_2), per la differenza tra le medie nel caso di varianze note, sono:

$$T_1=D-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}} \qquad T_2=D+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}$$

Osservati $\mathbf{x}=(x_1,\ldots,x_m)$ e $\mathbf{y}=(y_1,\ldots,y_n)$, l'I.C. per la differenza tra le medie $(\mu_x-\mu_y)$, nel caso di varianze note, al $100(1-\alpha)\%$ è:

$$\left\{ d - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \quad , \quad d + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \right\}$$

dove $$d = \bar{x} - \bar{y}$$

con $$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$$ e $$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Se le varianza sono sconosciute, Z non è una quantità pivot per $(\mu_x-\mu_y)$.

Secondo Caso: varianze uguali ma sconosciute

$$\sigma_x^2 = \sigma_y^2 = \sigma^2$$

Stimatore naturale della differenza tra le medie

$$D = \overline{X} - \overline{Y}$$

Da quanto detto in precedenza, si evince che

$$Z = \frac{D - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} = \frac{D - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1)$$

Si osserva, inoltre, che

$$V_1 = \frac{(m-1)S_x^2}{\sigma^2} = \frac{\sum_{i=1}^{m}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(m-1)$$

$$V_2 = \frac{(n-1)S_y^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2(n-1)$$

Poiché V_1 e V_2 sono indipendenti, dalla proprietà riproduttiva della v.c. chi-quadrato, si ha:

$$V_1 + V_1 = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right\} \sim \chi^2(n+m-2)$$

Dato che le v.c. Z e V_1+V_2 sono indipendenti, possiamo costruire la v.c. t-student, cioè

$$T = \frac{Z}{\sqrt{\frac{(V_1 + V_2)}{(m + n - 2)}}} \sim t(n + m - 2)$$

Tale rapporto si può scrive nel seguente modo:

$$T = \frac{\frac{D - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}}{\sqrt{\frac{\{(m-1)S_x^2 + (n-1)S_y^2\}}{(m+n-2)\sigma^2}}} = \frac{D - (\mu_x - \mu_y)}{S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

dove

$$S_p^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{(m+n-2)}$$

È lo stimatore non-distorto della varianza comune σ^2; infatti, si ha:

$$E\left[S_p^2\right] = \frac{(m-1)E\left[S_x^2\right] + (n-1)E\left[S_y^2\right]}{(m+n-2)} =$$

$$= \frac{(m-1)\sigma^2 + (n-1)\sigma^2}{(m+n-2)} = \sigma^2$$

In definitiva, la v.c.

$$T = \frac{D - (\mu_x - \mu_y)}{S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

è una quantità pivot perché funzione del c.c. (**X**,**Y**) e del parametro da stimare (μ_x-μ_y) con distribuzione che non dipende da parametri incogniti.

Fissato α, possiamo determinare q_1 e q_2 tali che

$$1 - \alpha = P_r\{q_1 < T < q_2\} =$$

$$= P_r\left\{ q_1 < \frac{D - (\mu_x - \mu_y)}{S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}}} < q_2 \right\} =$$

$$= P_r\left\{ q_1 S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}} < D - (\mu_x - \mu_y) < q_2 S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}} \right\} =$$

$$= P_r\left\{ -D + q_1 S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}} < -(\mu_x - \mu_y) < -D + q_2 S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}} \right\} =$$

$$= P_r\left\{ D - q_1 S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}} > (\mu_x - \mu_y) > D - q_2 S_p \times \sqrt{\frac{1}{m} + \frac{1}{n}} \right\} =$$

$$= P_r\left\{ D - q_2 S_p \times \sqrt{\frac{1}{m}+\frac{1}{n}} < (\mu_x - \mu_y) < D - q_1 S_p \times \sqrt{\frac{1}{m}+\frac{1}{n}} \right\}$$

Ricordando che

$$P_r\{T < q_1\} = \frac{\alpha}{2} = P_r\{T > q_2\} \quad \Rightarrow \quad \begin{cases} q_1 = -t_{\frac{\alpha}{2}(m+n-2)} \\ q_2 = t_{\frac{\alpha}{2}(m+n-2)} \end{cases}$$

Sostituendo, si ottiene:

$$P_r\left\{D-t_{\frac{\alpha}{2}(m+n-2)}\times S_p\sqrt{\frac{1}{m}+\frac{1}{n}}<(\mu_x-\mu_y)<D+t_{\frac{\alpha}{2}(m+n-2)}\times S_p\sqrt{\frac{1}{m}+\frac{1}{n}}\right\}=1-\alpha$$

Gli estremi dell'intervallo casuale (T_1,T_2), per la differenza tra le medie nel caso di varianze incognite ma uguali, sono:

$$T_1=D-t_{\frac{\alpha}{2}(m+n-2)}\times S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \qquad T_2=D+t_{\frac{\alpha}{2}(m+n-2)}\times S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$$

Osservati $\mathbf{x}=(x_1,\ldots,x_m)$ e $\mathbf{y}=(y_1,\ldots,y_n)$, l'I.C. per la differenza tra le medie $(\mu_x-\mu_y)$, nel caso di varianze incognite ma uguali, al 100(1-α)% è:

$$\left\{ d - t_{\frac{\alpha}{2}}(m+n-2) \times s_p \sqrt{\frac{1}{m}+\frac{1}{n}} \quad , \quad d + t_{\frac{\alpha}{2}}(m+n-2) \times s_p \sqrt{\frac{1}{m}+\frac{1}{n}} \right\}$$

dove $d = \bar{x} - \bar{y}$ e $s_p^2 = \dfrac{(m-1)s_x^2 + (n-1)s_y^2}{(m+n-2)}$

con $\bar{x} = \dfrac{1}{m}\sum_{i=1}^{m} x_i$ $\qquad \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{x})^2 \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

Esempio

Per provare l'efficacia di due nuovi semi per la produzione di grani sotto condizioni climatiche normali, un'industria di semi seleziona casualmente otto aziende agricole in una regione italiana e prova entrambi i semi su una determinata superficie coltivabile. Le produzioni per le otto aziende, secondo il seme utilizzato, sono le seguenti:

A : 86, 87, 56, 93, 84, 93, 75, 79

B : 80, 79, 58, 91, 77, 82, 74, 66

Supponendo che le due produzione siano, in ogni azienda, distribuite normalmente e che le varianze delle due popolazioni siano uguali, determinare l'intervallo di confidenza per la differenza delle produzioni medie, ad un livello di confidenza del 95%.

Intervallo di Confidenza sulla probabilità di successo

Sia X una v.c. di Bernoulli con probabilità di successo pari a θ.
Si ricorda che $E(X)=\theta$ e $V(X)= \theta(1- \theta)$.
Dato un c.c. $X_1,\ldots,X_n$ estratto da $B(\theta,1)$, si vuole costruire un intervallo di Confidenza per θ al $100(1-\alpha)\%$.
Consideriamo la proporzione di successi in n-prove indipendenti

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sappiamo che

$$E\left[\overline{X}\right]= \theta \qquad V\left[\overline{X}\right]= \frac{\theta(1-\theta)}{n}$$

Dal teorema di De Moivre-Laplace, sappiamo che

$$Z = \frac{\overline{X} - E\left[\overline{X}\right]}{\sqrt{V\left[\overline{X}\right]}} = \frac{\overline{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow[n \to \infty]{d} N(0,1)$$

Si evidenzia che la v.c. Z pur essendo una q.p. (perché è funzione del c.c. e del parametro incognito ed ha distribuzione indipendente da parametri incogniti) NON può essere utilizzata per costruire un intervallo di confidenza asintotico per la probabilità di successo. Infatti, la varianza della popolazione è sconosciuta. E' necessario quindi stimare la varianza della popolazione.

Consideriamo lo stimatore naturale della varianza della popolazione

$$S^{2'} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i - \overline{X}^2 = \overline{X}\left(1 - \overline{X}\right)$$

Perché $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i$ essendo $X_i=0$ oppure $X_i=1$.

Consideriamo, ora, la varianza campionaria non-distorta

$$S^2 = \frac{n}{n-1} S^{2'} = \frac{n\overline{X}\left(1 - \overline{X}\right)}{n-1}$$

Sostituiamo a V(X) lo stimatore non-distorto S^2

$$Z = \frac{\overline{X} - E\left[\overline{X}\right]}{\sqrt{V\left[\overline{X}\right]}} = \frac{\overline{X} - \theta}{\sqrt{\frac{S^2}{n}}} = \frac{\overline{X} - \theta}{\sqrt{\frac{n\overline{X}\left(1 - \overline{X}\right)}{n(n-1)}}} = \frac{\overline{X} - \theta}{\sqrt{\frac{\overline{X}\left(1 - \overline{X}\right)}{(n-1)}}}$$

Si dimostra che

$$Z = \frac{\overline{X} - \theta}{\sqrt{\frac{\overline{X}\left(1 - \overline{X}\right)}{(n-1)}}} \xrightarrow[n \to \infty]{d} N(0,1)$$

Quest'ultima è una quantità pivot per la probabilità di successo θ.

Fissato α, possiamo determinare q_1 e q_2 tali che

$$1-\alpha = P_r\{q_1 < Z < q_2\} =$$

$$= P_r\left\{ q_1 < \frac{\overline{X} - \theta}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}}} < q_2 \right\} =$$

$$= P_r\left\{ q_1 \sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}} < \overline{X} - \theta < q_2 \sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}} \right\} =$$

$$= P_r\left\{-\overline{X} + q_1\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}} < -\theta < -\overline{X} + q_2\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}}\right\} =$$

$$= P_r\left\{\overline{X} - q_2\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}} < \theta < \overline{X} + q_1\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}}\right\}$$

Analogamente, a quanto visto in precedenza possiamo dire che

$$q_1 = -z_{\frac{\alpha}{2}} \qquad q_2 = z_{\frac{\alpha}{2}}$$

Sostituendo abbiamo

$$P_r\left\{\overline{X}-z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}}<\theta<\overline{X}+z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}}\right\}=1-\alpha$$

Gli estremi dell'intervallo casuale (T_1,T_2), per la probabilità di successo, sono:

$$T_1=\overline{X}-z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}} \qquad T_2=\overline{X}+z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n-1}}$$

Osservato $\mathbf{x}=(x_1,\ldots,x_m)$, l'I.C. asintotico per la probabilità di successo θ, al $100(1-\alpha)\%$ è:

$$\left\{ \bar{x} - z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{x}(1-\bar{x})}{n-1}} \quad , \quad \bar{x} + z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{x}(1-\bar{x})}{n-1}} \right\}$$

dove

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Esempio.

Dal deposito di una industria che produce lampade, si estrae un c.c. di 350 unità e si osserva che il 25% sono difettose. Si determini un intervallo di confidenza per la proporzione di lampade difettose prodotte dall'industria in esame, al livello di confidenza del 95%.