X con trattino è media campionaria
S è la varianza campionaria

Politecnico di Milano – Facoltà di Ingegneria Informatica

Anno Accademico 2009/2010

Corso di Statistica (2L) per INF e TEL

Docente: Antonio Pievatolo; esercitazioni: Raffaele Argiento

Esercitazione del 29/04/09

Esercizio 1 (Dal tema d'esame del 18/07/06)

Siano $\bar{X}$ e S^2 rispettivamente la media e la varianza campionarie di un campione casuale $X_1, \ldots, X_n$ estratto da una popolazione gaussiana di media 2θ e varianza σ^2 ($X_1, \ldots, X_n$ i.i.d. $\sim \mathrm{N}(2\theta, \sigma^2)$, $\theta \in \mathbb{R}$, $\sigma^2 > 0$ entrambi incogniti).

1. Se il campione è costituito da 25 osservazioni, quanto vale $\mathbb{P}_{\theta,\sigma^2}(\bar{X} - 0.342S - 2\theta \leq 0)$?

Abbiamo misurato la pressione sistolica del sangue di 25 maschi sani e abbiamo ottenuto una media campionaria pari a 120.0mm di mercurio e un momento secondo campionario pari a 14616.00mm^2 di mercurio.

2 Sulla base di questi dati quanto siete confidenti che θ sia maggiore o uguale a 57.435?

Abbiamo raccolto ULTERIORI dati riguardanti 39 maschi sani e, per i nuovi 39 dati, abbiamo ottenuto una media campionaria pari a 110.0mm di mercurio e un momento secondo campionario pari a 12715.0mm^2

3 Aggiornate le stime puntuali di media e varianza sulla base di questi nuovi dati, usando l'intero campione di 64 misurazioni.

4 Determinate numericamente un intervallo di confidenza al 95% unilatero per il parametro θ di forma $(-\infty, c)$.

Soluzione

1. Dato che se $X_1, \ldots, X_n$ è un campione da una popolazione $\mathrm{N}(2\theta, \sigma^2)$ allora

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)} \tag{1}$$

dove $t_{(n-1)}$ indica la distribuzione $t-$Student con $n-1$ gradi di libertà. Cerchiamo di scrivere la probabilità richiesta in funzione della variabile (1):

$$\mathbb{P}_{\theta,\sigma^2}(\bar{X} - 0.3425 \cdot S - 2\theta \leq 0) = \mathbb{P}(\bar{X} - 2\theta \leq 0.3425S) = \mathbb{P}\left(\sqrt{n}\frac{\bar{X} - 2\theta}{S} \leq \sqrt{n} \cdot 0.342\right)$$
$$= \mathbb{P}(T_{n-1} \leq \sqrt{n}0.342) \quad \text{dove } T_{n-1} \sim t_{(n-1)}$$

Nel nostro caso $n = 25$, quindi $\mathbb{P}(T_{24} \leq 1.71) = 0.95$.

2 Sia $t_n(\gamma)$ il quantile di ordine γ di una distribuzione t_n. Si ha che

$$\left(\bar{X} - \frac{t_{n-1}(\gamma)}{\sqrt{n}}S, +\infty\right)$$

è un intervallo di confidenza sulla coda destra per la media di una popolazione gaussiana. La realizzazione campionaria data dall'esecizio è tale che $\bar{x} = 120$ e $M_2 = 16161$ ($M_2 = \frac{1}{n}\sum_{i=1}^n x_i^2$ è il momento campionario secondo). Il punto 2. richiede qual è il livello di confidenza dell'intervallo di confidenza sulla coda destra per θ dato da $(57.435, +\infty)$. A tal proposito risolviamo l'equazione in $t_{n-1}(\gamma)$:

$$57.435 = \frac{1}{2}\left(\bar{x} - \frac{t_{n-1}(\gamma)}{\sqrt{n}}s\right).$$

Usando i valori $n = 25$, $\bar{x} = 120$ e $s_2 = \frac{n}{n-1}(M_2 - \bar{x}^2) = 225$ si ottiene:

$$t_{24} = 1.71, \quad \text{quindi} \quad \gamma = 0.95$$

3 Sia ora $y_1, \ldots, y_{n_1}$, con $n_1 = 39$ la nuova realizzazione campionaria tale che $\bar{y} = 110$ e $M_{2,y} = 12713$. Aggiorniamo le stime ottenute dalla realizzazione $x_1, \ldots, x_n$.

$$2\bar{\theta}_{x,y} = M_{1,x,y} = \frac{x_1 + \cdots + x_n + y_1 + \cdots + y_{n_1}}{n + n_1} = 113.9062$$

$$M_{2,x,y} = \frac{x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_{n_1}^2}{n + n_1} = 13457.58$$

$$s^2_{x,y} = \frac{n + n_1}{n + n1 - 1}(M_{2,x,y} - M^2_{1,x,y}) = 490.6$$

2. Se μ è la media di una popolazione Gaussiana, allora $\mathbb{P}\big(\frac{\bar{X}-\mu}{S/\sqrt{n}} \geq -t_{n-1}(\gamma)\big) = \gamma$. Di conseguenza

$$\mathbb{P}\big(\theta \leq \frac{1}{2}\{\bar{X} + \frac{S}{\sqrt{n}}t_{n-1}(\gamma)\}\big) = \gamma$$

definisce un intervallo di confidenza sulla coda sinistra per θ, la cui realizzazione campionaria è

$$(-\infty, 59.32038)$$

▮

Esercizio 2

Una ditta produce punte da trapano. Si provano n punte dello stesso diametro producendo n fori. Si indichi con $Y_1, \ldots, Y_n$ i diametri dei fori prodotti e si supponga che $Y_1, \ldots, Y_n$ siano normali con media μ incognita e varianza σ_0^2 nota. Rispondere alle seguenti domande giustificando adeguatamente le risposte.

1. Determinare un intervallo di confidenza bilatero per la media μ di livello γ.
2. Si supponga ora che $n = 100$, $\bar{Y}_n = 5mm$, $\sigma_0^2 = 10^{-2}\text{mm}^2$, $\gamma = 95\%$ ($\bar{Y}_n$ = media campionaria). Calcolare l'intervallo di confidenza del punto precedente con questi dati.
3. Se $\sigma_0^2 = 10^{-2}\text{mm}^2$, quanto grande occorre prendere il campione per essere sicuri al 95% che la nostra stima di μ sia precisa entro 10^{-2}mm?
4. Abbiamo ora estratto un altro campione (totalmente nuovo) sempre dalla famiglia gaussiana $\text{N}(\mu, 10^{-2})$ e sappiamo che per questo nuovo campione $(4.50, 4.54)$ é un IC (simmetrico) al 95% per μ. Quale è la media campionaria della popolazione e quale la dimensione di questo nuovo campione estratto?

Soluzione

1. Sia $Y_1, \ldots, Y_n$ un campione da una popolazione $\text{N}(\mu, \sigma_0^2)$, allora $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i \sim\text{N}(\mu, \frac{\sigma_0^2}{n})$. Definiamo dunque la variabile:

$$Z := \frac{\bar{Y}_n - \mu}{\sigma_0/n} \sim \text{N}(0, 1).$$

Si osservi come la variabile Z dipenda dal parametro μ mentre la sua distribuzione è indipendente da esso.

Sia ora $z_{\frac{1+\gamma}{2}}$ il percentile di ordine $\frac{1+\gamma}{2}$ della distribuzione gaussiana standard. Si ha che

$$\mathbb{P}\left(-z_{\frac{1+\gamma}{2}i} < Z < z_{\frac{1+\gamma}{2}}\right) = \gamma \;\Rightarrow\; \mathbb{P}\left(-z_{\frac{1+\gamma}{2}} < \frac{\bar{Y}_n - \mu}{\sigma_0/n} < z_{\frac{1+\gamma}{2}}\right) = \gamma \;\Rightarrow$$
$$\Rightarrow \mathbb{P}\left(\bar{Y}_n - \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}} < \mu < \bar{Y}_n + \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}\right) = \gamma$$

Dall'ultima uguaglianza si evince che

$$\left(\bar{y}_n - \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}, \bar{y}_n + \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}\right)$$

è un intervallo di confidenza di livello γ per μ.

2. Utilizzando il valori forniti dal testo si ottiene l'intervallo

$$\left(\bar{y}_n - \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}, \bar{y}_n + \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}\right) = (4.9804, 5.0196)$$

3. La lunghezza dell'intervallo di confidenza trovato al punto 1. è la quantità non aleatoria:

$$\mathcal{L}_\gamma(n) = 2\frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}$$

La stima di μ è precisa entro 10^{-2}mm, al livello γ se $\mathcal{L}_\gamma(n)/2 < 10^{-2}$. Si ha dunque:

$$\frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}} < 10^{-2} \Rightarrow \frac{1}{\sqrt{n}} < \left(10^2 \sigma_0 z_{\frac{1+\gamma}{2}}\right)^{-1} \Rightarrow n > \left(10^2 \sigma_0 z_{\frac{1+\gamma}{2}}\right) \Rightarrow n > 384.16 \Rightarrow n > 385.$$

4. La nuova osservazione ha fornito un intervallo di confidenza $\left(\bar{y}_n - \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}, \bar{y}_n + \frac{\sigma_0}{\sqrt{n}} z_{\frac{1+\gamma}{2}}\right) = (4.5, 4.54)$ la media campionaria $\bar{y}_n$ è il centro di questo intervallo. Dunque $\bar{y}_n = \frac{4.5+4.54}{2} = 4.52$. La lunghezza dell'intervallo è 0.05, dunque

$$\mathcal{L}_\gamma(n) = 0.05 \Rightarrow n = \left(\frac{\sigma_0 z_{\frac{1+\gamma}{2}}}{0.02}\right) \Rightarrow n \simeq 96$$

∎

Esercizio 3.

Una ditta produce cinghie di trasmissione per automobili. È noto che la durata di tali cinghie, calcolata in migliaia di chilometri percorsi, ha distribuzione normale con media 50 e scarto quadratico medio 5. Denotiamo con X la durata di una cinghia. Successivamente la ditta progetta un congegno che -applicato al sistema di trasmissione- consente di allungare la "vita" di una cinghia di una quantità $Y \sim \mathrm{N}(\mu, \sigma^2)$, con parametri μ e σ^2 entrambi incogniti. In altre parole, la durata delle cinghie in presenza del congegno è $Z = X + Y$ e assumiamo che X e Y siano indipendenti. Un test su 30 automobili provviste del nuovo congegno ha permesso di raccogliere dati relativi ad un campione $Z_1, ..., Z_30$ che ha fornito $\bar{z}_{30} = 57$, $s_{30} = 6.5$, dove $\bar{Z}_n$ indica la media campionaria e S_n^2 è la varianza campionaria di $Z_1, \ldots, Z_{30}$.

1. Qual è la distribuzione di Z?

2. Sulla base del campione $Z_1, \ldots, Z_{30}$ fornire una stima per μ e σ^2.

3. Determinare un intervallo di confidenza per μ di livello $\gamma = 0.95$.

4. Determinare un intervallo di confidenza bilatero per la deviazione standard σ di livello 0.95.

Soluzione

1. X e Y sono indipendenti ed entrambe hanno distribuzione normale, dunque Z ha distribuzione normale, ovvero $Z \sim \mathrm{N}(\mu_Z, \sigma_Z^2)$. Dove:

$$\mu_Z = \mu_X + \mu_Y = 50 + \mu \quad \Rightarrow \quad \mu = \mu_Z - 50$$
$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 = 25 + \sigma^2 \quad \Rightarrow \quad \sigma^2 = \sigma_D^2 - 25$$

2. È noto che $\bar{Z}_n$ è una stima non distorta di μ_Z, quindi $\bar{Z}_n$ è una stima non distorta di μ. Allo stesso modo dato che la varianza campionaria S_Z^2 è una stima non distorta della varianza di popolazione σ_Z^2, si ha che $S_Z^2 - 25$ è non distorto per σ^2. In particolare:

$$\bar{\mu} = \bar{z}_n - 50 = 7$$
$$\bar{\sigma}^2 = s_Z^2 - 25 = 17.25$$

3. La variabile aleatoria:

$$T := \frac{\bar{Z}_n - \mu_Z}{S_Z/\sqrt{n}} \sim t_{n-1}$$

quindi

$$\mathbb{P}\left(-t_{n-1}(\frac{1+\gamma}{2}) < \frac{\bar{Z}_n - \mu_Z}{S_Z/\sqrt{n}} < t_{n-1}(\frac{1+\gamma}{2})\right) = \gamma$$

di conseguenza

$$\mathbb{P}\left(-t_{n-1}(\frac{1+\gamma}{2}) < \frac{\bar{Z}_n - (\mu + 50)}{S_Z/\sqrt{n}} < t_{n-1}(\frac{1+\gamma}{2})\right) = \gamma$$

con qualche passaggio algebrico si ricava

$$\mathbb{P}\left((\bar{Z}_n - 50) - \frac{S_Z}{\sqrt{n}}t_{n-1}(\frac{1+\gamma}{2}) < \mu < (\bar{Z}_n - 50) + \frac{S_Z}{\sqrt{n}}t_{n-1}(\frac{1+\gamma}{2})\right) = \gamma.$$

In conclusione

$$\left((\bar{z}_n - 50) - \frac{s_Z}{\sqrt{n}}t_{n-1}(\frac{1+\gamma}{2}), (\bar{z}_n - 50) + \frac{s_Z}{\sqrt{n}}t_{n-1}(\frac{1+\gamma}{2})\right) = (4.57, 9.43)$$

è l'intervallo cercato.

4. Si definisce la variabile

$$\chi := (n-1)\frac{S_Z^2}{\sigma_Z^2} \sim \chi_{n-1}^2.$$

Siano $c_1 = \chi_{n-1}^2\left(\frac{1-\gamma}{2}\right)$ e $c_2 = \chi_{n-1}^2\left(\frac{1+\gamma}{2}\right)$ i percentili di ordine $\frac{1-\gamma}{2}$ e $\frac{1+\gamma}{2}$ di una chi-quadro con $n-1$ gradi di libertà. Si ottiene:

$$\mathbb{P}\left(c_1 < (n-1)\frac{S_Z^2}{\sigma_Z^2} < c_2\right) = \gamma$$

sostituendo $\sigma_Z^2 = 25 + \sigma^2$ con qualche passaggio algebrico si ottiene

$$\mathbb{P}\left((n-1)\frac{S_Z^2}{c_2} - 25 < \sigma^2 < (n-1)\frac{S_Z^2}{c_1} - 25\right) = \gamma.$$

Quindi

$$\mathbb{P}\left(\sqrt{(n-1)\frac{S_Z^2}{c_2} - 25} < \sigma < \sqrt{(n-1)\frac{S_Z^2}{c_1} - 25}\right) = \gamma.$$

L'intervallo di confidenza richiesto è dunque:

$$\left(\sqrt{(n-1)\frac{s_Z^2}{c_2} - 25}, \sqrt{(n-1)\frac{S_Z^2}{c_1} - 25}\right) = (1.341, 7.167)$$

∎