Statistica II - Esercitazione 9

Esercizio 1

Si supponga che la lunghezza dei chiodi prodotti da due ditte X e Y abbia distribuzione Normale. Per verificare l'ipotesi di uguaglianza delle lunghezze medie delle produzioni delle due ditte si effettuano $n = 10$ prove bernoulliane dalla produzione di X, ottenendo:

$$\sum_{i=1}^{n} x_i = 20 \qquad \sum_{i=2}^{n} (x_i - \overline{x})^2 = 20$$

ed $m = 6$ prove bernoulliane dalla produzione di Y, ottenendo:

$$\sum_{i=1}^{n} y_i = 13.2 \qquad \sum_{i=2}^{n} (y_i - \overline{y})^2 = 20.5$$

1. Sottoporre a verifica l'ipotesi $H_0 : \mu_x = \mu_y$, ad un livello di significatività $\alpha = 0.01$, supponendo che le varianze siano note, $\sigma_x^2 = 3$ e $\sigma_y^2 = 4.5$. Calcolare inoltre il p-value e la potenza del test per l'ipotesi alternativa semplice $H_1 : \mu_x - \mu_y = 3$.
2. Supponendo che le varianze siano ignote, identificare la regione di rifiuto e decidere se rifiutare o meno l'ipotesi nulla.

Soluzione

Siano $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ e $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ indipendenti, σ_x^2 e σ_y^2 note al punto 1, μ_x, $\mu_y \in \mathbb{R}^+$ ignote e $\mu_x - \mu_y \in \mathbb{R}$ oggetto di interesse. Siano inoltre $X_1, \dots, X_n$ i.i.d. a X e $Y_1, \dots, Y_m$ i.i.d. a Y.

1. Si vuole sottoporre a verifica di ipotesi il sistema $H_0 : \mu_x = \mu_y$ contro $H_1 : \mu_x \neq \mu_y$. Il sistema di ipotesi precedente è equivalente a $H_0 : \mu_x - \mu_y = 0$ contro $H_1 : \mu_x - \mu_y \neq 0$. Assunta vera l'ipotesi nulla, $\mu_x = \mu_y = \mu$, è possibile quindi specificare una distribuzione per la media campionaria $\overline{X}$ ed $\overline{Y}$:

$$\overline{X} \overset{H_0}{\sim} \mathcal{N}\left(\mu, \frac{\sigma_x^2}{n}\right)$$
$$\overline{Y} \overset{H_0}{\sim} \mathcal{N}\left(\mu, \frac{\sigma_y^2}{m}\right)$$

   ed essendo $X_1, \dots, X_n$ i.i.d., $Y_1, \dots, Y_n$ i.i.d. e X indipendente da Y, è possibile esplicitare una legge di distribuzione per la differenza delle medie campionarie:

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right) \overset{H_0}{\sim} \mathcal{N}\left(0, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

   Sotto ipotesi nulla è possibile quindi ricondursi alla statistica test:

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \overset{H_0}{=} \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \mathcal{N}(0, 1)$$

   Fissato un livello di significatività, è possibile individuare la regione critica associata al test. Essendo $\alpha = 0.01 \Rightarrow \frac{\alpha}{2} = 0.005$, $z_{\frac{\alpha}{2}} = -2.58$, $z_{1-\frac{\alpha}{2}} = 2.58$, la regione critica è l'insieme:

$$\begin{aligned} C &= \left\{\underline{X}, \underline{Y} : \{Z \leq z_{\frac{\alpha}{2}}\} \cup \{Z \geq z_{\frac{\alpha}{2}}\}\right\} = \left\{\underline{X}, \underline{Y} : \{Z \leq -2.58\} \cup \{Z \geq 2.58\}\right\} \\ &= \left\{\underline{X}, \underline{Y} : \left\{\overline{X} - \overline{Y} \leq -2.58 \cdot \sqrt{\frac{\sigma_n}{n} + \frac{\sigma_y}{m}}\right\} \cup \left\{\overline{X} - \overline{Y} \geq 2.58 \cdot \sqrt{\frac{\sigma_n}{n} + \frac{\sigma_y}{m}}\right\}\right\} = \\ &= \left\{\underline{X}, \underline{Y} : \{\overline{X} - \overline{Y} \leq -2.64\} \cup \{\overline{X} - \overline{Y} \geq 2.64\}\right\} \end{aligned}$$

La differenza delle medie campionare, dati i due campioni estratti, è pari a:

$$\overline{x}-\overline{y}=\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{m}\sum_{j=1}^{n}y_j=\frac{20}{10}-\frac{13.2}{6}=-0.2$$

ed essendo $-2.64<-0.2<2.64 \implies \overline{x}-\overline{y}\notin C$, l'ipotesi nulla non viene rifiutata.
Il p-value del test è pari a:

$$\begin{aligned}
p-value &= \Pr\left(|Z|>|z_{oss}|\,|H_0\right)=\Pr\left(Z>|z_{oss}||H_0\right)+\Pr\left(Z<-|z_{oss}||H_0\right)=\\
&= 2\cdot\Pr\left(Z>|z_{oss}||H_0\right)=2\left[1-\Pr\left(Z\leq|z_{oss}||H_0\right)\right]=\\
&= 2\left[1-\Pr\left(Z\leq\frac{0.2}{\sqrt{1.05}}\right)\right]=2\left[1-\Phi(0.19)\right]=\\
&= 2\left[1-0.58\right]=2\cdot0.42=0.84
\end{aligned}$$

La potenza del test, con ipotesi alternativa semplice $H_1:\mu_x-\mu_y=3$, è pari a:

$$\begin{aligned}
1-\beta &= \Pr\left(\overline{X}-\overline{Y}\in C|H_1\right)=1-\Pr\left(\overline{X}-\overline{Y}\notin\overline{C}|H_1\right)=\\
&= 1-\Pr\left(-2.64\leq\overline{X}-\overline{Y}\leq2.64|H_1\right)=\\
&= 1-\Pr\left(\frac{-2.64-3}{\sqrt{1.05}}\leq Z\leq\frac{2.64-3}{\sqrt{1.05}}|H_1\right)=\\
&= 1-\Pr\left(-5.50\leq Z\leq-0.35|H_1\right)=\\
&= 1-\left[\Phi(-0.35)-\Phi(-5.50)\right]=0.64
\end{aligned}$$

2. Supponiamo ora che le varianze siano ignote. Usando lo stimatore pooled della varianza:

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

è possibile ricondursi alla statistica test

$$T=\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}\overset{H_0}{\sim}T_{n+m-2}.$$

La statistica test precedente identifica come regione critica:

$$\begin{aligned}
C &= \left\{\underline{X},\underline{Y}:\{T\leq t_{n+m-2;\frac{\alpha}{2}}\}\cup\{T\geq t_{n+m-2;1-\frac{\alpha}{2}}\}\right\}=\\
&= \left\{\underline{X},\underline{Y}:\{T\leq-2.98\}\cup\{T\geq2.98\}\right\}
\end{aligned}$$

Essendo la varianza pooled, calcolata nei campioni estratti, pari a:

$$s_p^2=\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}=\frac{1}{n+m-2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{j=1}^{m}(y_j-\overline{y})^2\right]=2.89$$

la statistica osservata assume valore:

$$t_{oss}=\frac{\overline{X}-\overline{Y}}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}=\frac{-0.2}{\sqrt{2.89\left(\frac{1}{10}+\frac{1}{6}\right)}}=-0.23$$

$-2.98<-0.23<2.98\Rightarrow t_{oss}\notin C$, l'ipotesi nulla non viene rifiutata.

Esercizio 2

Si supponga che un laboratorio chimico abbia ricevuto in donazione un nuovo strumento atto alla misurazione dei tempi di reazione di alcune sostanze chimiche. Con lo scopo di valutare la precisione del nuovo strumento, due chimici effettuano alcune misurazioni del tempo di reazione di una sostanza; a parità di condizioni di pressione e temperatura, il chimico X effettua 11 rilevazioni con lo strumento tradizionale riscontrando una varianza campionaria corretta s_x^2 dei tempi di reazione pari a 0.9, mentre il chimico Y effettua 16 rilevazioni con il nuovo strumento riscontrando $s_y^2 = 3.3$. Supponendo che il tempo di reazione della sostanza in questione abbia distribuzione Normale, si risponda ai seguenti quesiti.

1. Si dica se sia possibile ritenere valida l'ipotesi $H_0 : \sigma_x^2 = \sigma_y^2$ contro $H_1 : \sigma_x^2 \neq \sigma_y^2$, a livello di significatività $\alpha = 0.1$;

2. In virtù di quanto ottenuto al punto precedente e senza effettuare calcoli aggiuntivi, si dica se è plausibile che il p-value abbia valore pari a 0.12, giustificando opportunamente la risposta.

Soluzione

Siano $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ e $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ indipendenti, σ_x^2 e $\sigma_y^2 \in \mathbb{R}^+$ ignote e oggetto di interesse. Siano inoltre $X_1, \ldots, X_n$ i.i.d. a X e $Y_1, \ldots, Y_m$ i.i.d. a Y.

1. Lo stimatore della varianza campionaria, con media ignota, S^2, descrive una legge di distribuzione nota, per le variabili casual X ed Y, con campioni rispettivamente di ampiezza n ed m, si ha che:

$$S_x^2 \sim \frac{\sigma_x^2}{n-1}\chi^2_{n-1} \qquad S_y^2 \sim \frac{\sigma_y^2}{m-1}\chi^2_{m-1}$$

Assumendo che l'ipotesi nulla sia vera, le varianze ignote σ_x^2 e σ_y^2 coincidono tra loro. Possiamo quindi ricondurci ad una statistica test nota, definendo il rapporto tra i due stimatori delle varianze campionarie:

$$F = \frac{S_x^2}{S_y^2} \sim \frac{\frac{\sigma_x^2}{n-1}\chi^2_{n-1}}{\frac{\sigma_y^2}{m-1}\chi^2_{m-1}} \overset{H_0}{\sim} \frac{\frac{\chi^2_{n-1}}{n-1}}{\frac{\chi^2_{m-1}}{m-1}} \sim \mathcal{F}_{n-1,m-1}$$

L'ipotesi nulla non sarà supportata dai dati quando il rapporto tra le varianze campionarie si discosta dal valore 1. Stante alla non simmetria della variabile casuale F, vogliamo identificare una regione critica ad un livello di significatività $\alpha = 0.1$, risolvendo:

$$\Pr(\mathcal{F} \leq f_{n-1,m-1;\frac{\alpha}{2}} | H_0) = \Pr(\mathcal{F} \geq f_{n-1,m-1;1-\frac{\alpha}{2}} | H_0) = \frac{\alpha}{2}$$

dove $f_{n-1,m-1;\frac{\alpha}{2}}$ e $f_{n-1,m-1;1-\frac{\alpha}{2}}$ sono i due quantili della distribuzione F, con $n-1$ e $m-1$ gradi di libertà, necessari ad identificare la regione critica. Essendo $n = 11$, $m = 16$ e $\alpha = 0.1$, dalle tavole risulta:

$$f_{n-1,m-1;1-\frac{\alpha}{2}} = f_{10,15,0.95} = 2.85$$

Ai fini della determinazione di $f_{n-1,m-1;\frac{\alpha}{2}}$, non riscontrabile dalle tavole, si può ricorrere a:

$$\frac{\alpha}{2} = \Pr(\mathcal{F}_{g,d} \leq f_{g,d;\frac{\alpha}{2}}) = \Pr\left(\frac{1}{\mathcal{F}_{g,d}} \geq \frac{1}{f_{g,d;\frac{\alpha}{2}}}\right) = \Pr\left(\mathcal{F}_{d,g} \geq \frac{1}{f_{g,d;\frac{\alpha}{2}}}\right)$$

da cui:

$$\frac{1}{f_{g,d;\frac{\alpha}{2}}} = f_{d,g;1-\frac{\alpha}{2}} \quad \Longrightarrow \quad f_{g,d;\frac{\alpha}{2}} = \frac{1}{f_{d,g;1-\frac{\alpha}{2}}}$$

ed applicando il risultato per il caso in studio otteniamo:

$$f_{n-1,m-1;\frac{\alpha}{2}} = f_{10,15;0.05} = \frac{1}{f_{15,10;0.95}} = \frac{1}{2.54} = 0.39$$

La regione critica associata al test specificato è l'insieme:

$$C = \left\{ \underline{X}, \underline{Y} : \frac{S_x^2}{S_y^2} \in \left\{ \left[0; f_{n-1,m-1;\frac{\alpha}{2}}\right] \cup \left[f_{n-1,m-1;1-\frac{\alpha}{2}}; \infty\right) \right\} \right\} =$$
$$= \left\{ \underline{X}, \underline{Y} : \frac{S_x^2}{S_y^2} \in \{[0; 0.39] \cup [2.85; \infty)\} \right\}$$

Essendo:

$$f_{oss} = \frac{s_x^2}{s_y^2} = \frac{0.9}{3.3} = 0.27$$

e $0.27 < 0.39 \;\Rightarrow\; f_{oss} \in C$, l'ipotesi nulla viene rifiutata ad un livello di singificatività $\alpha = 0.1$, le due varianze sono significativamente diverse tra loro.

2. Se il p-value relativo alla verifica dell'ipotesi in questione fosse di valore pari a 0.12, essendo

$$\alpha = 0.1 < 0.12$$

si accetterebbe H_0 a livello di significatività $\alpha = 0.1$. Ciò è in contraddizione con quanto ricavato al punto 1. Coerentemente con il rifiuto di H_0 per $\alpha = 0.1$, il valore del p-value associato deve essere minore di 0.1 e, in particolare, non può essere pari a 0.12.

Esercizio 3

Nello scorso anno accademico 80 studenti dei 104 presenti hanno superato al primo appello l'esame del corso di Statistica I istituito presso le lauree in Statistica di questo Ateneo, fenomeno descritto da una variabile casuale di Bernulli $X \sim Be(\theta_x)$, dove θ_x rappresenta la probabilità di successo. Nel primo appello dell'anno accademico precedente furono invece 66 gli studenti che superarono l'esame in questione tra i 98 presenti, fenomeno descritto da una variabile casuale di Bernulli $Y \sim Be(\theta_y)$. Si vuole verificare se, ad un livello di significatività $\alpha = 0.05$, sia sensato assumere che la probabilità di superamento dell'esame di Statistica I sia aumentata tra un anno accademico ed il successivo.

1. Individuare la regione critica.
2. Calcolare il p-value del test.

Soluzione

Siano $X_1, \dots, X_n$ i.i.d. secondo una $Be(\theta_x)$, dove $\theta_x \in (0,1)$ parametro ignoto, ed $Y_1, \dots, Y_m$ i.i.d. secondo una $Be(\theta_y)$, con $\theta_y \in (0,1)$ parametro ignoto. Sia X la variabile casuale che descrive il numero di successi ottenuti in n prove effettuate per il campione $\underline{X}$. Tale variabile si distribuisce secondo una legge di probabilità Binomiale, $X \sim Bin(n, \theta_x)$. Analogamente, sia Y la variabile casuale che descrive il numero di successi ottenuti in m prove effettuate per il campione $\underline{Y}$, dove $Y \sim Bin(m, \theta_y)$. Inoltre, è ragionevole assumere che X sia indipendente da Y.

1. Il sistema di ipotesi che si vuole sottoporre a verifica è $H_0 : \theta_x = \theta_y$ contro $H_1 : \theta_x > \theta_y$, essendo interessati a discriminare se la probabilità di superamento sia aumentata, tra un anno accademico ed il successivo. Siano $\overline{X}$ e $\overline{Y}$ gli stimatori della proporzione campionaria di successi:

   $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

   Al crescere della numerosità campionaria, è possibile utilizzare un'approssimazione asintotica per la distribuzione degli stimatori precedenti:

   $$\overline{X} \overset{n\to\infty}{\sim} \mathcal{N}\left(\theta_x, \frac{\theta_x(1-\theta_x)}{n}\right) \qquad \overline{Y} \overset{m\to\infty}{\sim} \mathcal{N}\left(\theta_y, \frac{\theta_y(1-\theta_y)}{m}\right)$$

   Pertanto vale:

   $$\overline{X} - \overline{Y} \overset{n,m\to\infty}{\sim} \mathcal{N}\left(\theta_x - \theta_y, \frac{\theta_x(1-\theta_x)}{n} + \frac{\theta_y(1-\theta_y)}{m}\right)$$

   Assumendo vera l'ipotesi nulla, $\theta_x = \theta_y = \theta$:

   $$\overline{X} - \overline{Y} \overset{H_0;n,m\to\infty}{\sim} \mathcal{N}\left(0, \theta(1-\theta)\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

   La varianza, sotto H_0, dipende solo da θ, parametro che può essere stimato con la media aritmetica ponderata delle proporzioni:

   $$\hat{\theta} = \frac{n\overline{x} + m\overline{y}}{n+m}$$

   Possiamo quindi ricorrere alla statistica test:

   $$Z = \frac{(\overline{X} - \overline{Y}) - (\theta_x - \theta_y)}{\sqrt{\frac{\theta_x(1-\theta_x)}{n} + \frac{\theta_y(1-\theta_y)}{m}}} \overset{H_0}{=} \frac{(\overline{X} - \overline{Y})}{\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n} + \frac{1}{m}\right)}} \overset{n,m\to\infty}{\sim} \mathcal{N}(0,1)$$

Essendo $H_1 : \theta_x \geq \theta_y$, ovvero $H_1 : \theta_x - \theta_y \geq 0$, la regione critica, associata a tale test, è l'insieme:

$$C = \{\underline{X}, \underline{Y} : Z \geq z_{1-\alpha}\}$$

Per il problema in analisi, $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$, quindi $z_{1-\alpha} = z_{0.95} = 1.64$, e la regione critica diventa:

$$C = \{\underline{X}, \underline{Y} : Z \geq 1.64\}$$

Per la variabile X si sono ottenuti $x = 80$ successi su $n = 104$ prove, mentre per la variabile casuale Y si sono ottenuti $y = 66$ successi su $m = 98$ prove. Quindi:

$$\begin{aligned}
\overline{x} &= \frac{80}{104} = 0.7692 \\
\overline{y} &= \frac{66}{98} = 0.6735 \\
\hat{\theta} &= \frac{80 + 66}{104 + 98} = 0.7228
\end{aligned}$$

La statistica test osservata è pari a:

$$z_{oss} = \frac{\overline{x} - \overline{y}}{\sqrt{\hat{\theta}(1 - \hat{\theta})\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{0.0957}{\sqrt{0.7228 \cdot 0.2772 \cdot 0.0198}} = 1.52$$

ed essendo $1.52 < 1.64 \Rightarrow z_{oss} \notin C$, l'ipotesi nulla non viene rifiutata, le due proporzioni non sono significativamente differenti.

2. Il p-value associato al test è pari a:

$$p - value = \Pr(Z \geq 1.52 | H_0) = 1 - \Pr(Z \leq 1.52 | H_0) = 1 - 0.9357 = 0.0643$$

che ad un livello di significatività $\alpha = 0.05$, porta a non rifiutare l'ipotesi nulla, analogamente alle conclusioni trovate utilizzando la regione critica.