Capitolo 3
Richiami di Statistica

1. Il teorema del limite centrale suggerisce che quando la dimensione del campione (n) è grande, la distribuzione della media campionaria ($\overline{Y}$) è approssimativamente $N\left(\mu_Y, \sigma^2_{\overline{Y}}\right)$ con $\sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n}$. Data una popolazione con $\mu_Y = 100,\ \sigma^2_Y = 43.0,$ abbiamo

   (a) $n = 100,\ \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n} = \frac{43}{100} = 0.43,$ e

   $$\Pr(\overline{Y} < 101) = \Pr\left(\frac{\overline{Y} - 100}{\sqrt{0.43}} < \frac{101 - 100}{\sqrt{0.43}}\right) \approx \Phi(1.525) = 0.9364.$$

   (b) $n = 64,\ \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{64} = \frac{43}{64} = 0.6719,$ e

   $$\Pr(101 < \overline{Y} < 103) = \Pr\left(\frac{101 - 100}{\sqrt{0.6719}} < \frac{\overline{Y} - 100}{\sqrt{0.6719}} < \frac{103 - 100}{\sqrt{0.6719}}\right)$$
   $$\approx \Phi(3.6599) - \Phi(1.2200) = 0.9999 - 0.8888 = 0.1111.$$

   (c) $n = 165,\ \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n} = \frac{43}{165} = 0.2606,$ e

   $$\Pr(\overline{Y} > 98) = 1 - \Pr(\overline{Y} \le 98) = 1 - \Pr\left(\frac{\overline{Y} - 100}{\sqrt{0.2606}} \le \frac{98 - 100}{\sqrt{0.2606}}\right)$$
   $$\approx 1 - \Phi(-3.9178) = \Phi(3.9178) = 1.0000 \text{ (rounded to four decimal places).}$$

3. Definiamo la preferenza di ogni votante con Y. $Y = 1$ se il votante preferisce il candidato uscente e $Y = 0$ se preferisce lo sfidante. Y è quindi una variabile casuale di Bernoulli con probabilità $\Pr(Y = 1) = p$ e $\Pr(Y = 0) = 1 - p$. Dalle precedenti lezioni sappiamo che Y ha media p e varianza $p(1 - p)$.

   (a) $\hat{p} = \frac{215}{400} = 0.5375.$

   (b) $\text{var}(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{n} = \frac{0.5375 \times (1 - 0.5375)}{400} = 6.2148 \times 10^{-4}.$ L'errore standard è $\text{SE}(\hat{p}) = (\text{var}(\hat{p}))^{\frac{1}{2}} = 0.0249.$

   (c) La statistica t ottenuta dai dati è

   $$t^{act} = \frac{\hat{p} - \mu_{p,0}}{\text{SE}(\hat{p})} = \frac{0.5375 - 0.5}{0.0249} = 1.506.$$

   Vista la numerosità del campione $(n = 400),$ possiamo ottenere il valore-p per testare $H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$:

   $$valore - p = 2\Phi(-\left|t^{act}\right|) = 2\Phi(-1.506) = 2 * 0.066 = 0.132$$

(d) Il valore-p nel caso di ipotesi unilaterali $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$ è

$$valore - p = 1 - \Phi(t^{act}) = 1 - \Phi(1.506) = 1 - 0.934 = 0.066$$

(e) La lettera (c) richiede un test bilaterale ed il valore-p è l'area sotto le code della distribuzione normale standard al di fuori di ± (la statistica calcolata). La lettera (d) è un test unilaterale è il valore-p è l'area sotto la distribuzione normale standard a destra della statistica t calcolata.

(f) Per il test $H_0 : p = 0.5$ vs. $H_1 : p > 0.5,$ non possiamo rifiutare l'ipotesi nulla al livello di significatività del 5%. Il valore-p 0.066 è maggiore di 0.05. Allo stesso modo, la statistica t calcolata 1.506 è minore del valore critico 1.645 per un test unilaterale al livello di significatività del 5%. Il test suggerisce che l'indagine empirica non contiene prove statisticamente significanti circa un possibile vantaggio del candidato uscente nei confronti dello sfidante.

3. Dimensione campione uomini $n_1 = 100,$ media campinonaria $\overline{Y}_1 = 3100,$ deviazione standard campionaria $s_1 = 200.$ Dimensione campione donne $n_2 = 64,$ media campionaria $\overline{Y}_2 = 2900,$ deviazione standard campionaria $s_2 = 320.$ L'errore standard di $\overline{Y}_1 - \overline{Y}_2$ è

$SE(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{200^2}{100} + \frac{320^2}{64}} = 44.721.$

(a) Le ipotesi per testare la differenza nei salari medi mensili sono

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

La statistica t per testare l'ipotesi nulla è

$$t^{act} = \frac{\overline{Y}_1 - \overline{Y}_2}{SE(\overline{Y}_1 - \overline{Y}_2)} = \frac{3100 - 2900}{44.721} = 4.4722.$$

Il valore-p è dato da:

$$valore - p = 2\Phi(-\left|t^{act}\right|) = 2\Phi(-4.4722) = 2 * (3.8744 * 10^{-6}) = 7.7488 * 10^{-6}.$$

Il livello estremamente basso del valore-p implica che la differenza nei salari mensili tra uomini e donne è statisticamente significativa. Possiamo rifiutare l'ipotesi nulla con un alto livello di confidenza.

(b) Dal punto precedente, abbiamo una forte evidenza statistica circa la differenza tra i salari medi tra uomini e donne. Per verificare la presenza di discriminazione sessuale nelle politiche di compenso, possiamo usare un'ipotesi alternativa unilaterale

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 > 0.$$

Con una statistica t $t^{act} = 4.4722,$ il valore-p per il test unilaterale è:

$$valore - p = 1 - \Phi(t^{act}) = 1 - \Phi(4.4722) = 1 - 0.999996126 = 3.874 * 10^{-6}$$

Da solo questo risultato non implica discriminazione tra sessi da parte della ditta. Tale discriminazione significa che tra due lavoratori, indentici in tutto escluso il genere, percepiscono salari diversi. La descrizione dei dati suggerisce molta cura nel verificare la comparazione di mansioni similari. È altresì importante controllare per le caratteristiche dei lavoratori che possono influenzare la loro produttività (istruzionie, anni d'esperienza, etc.). Se tali caratteristiche sono sistematicamente diverse tra uomini e donne, allora potrebbero essere responsabili delle differenze salariali. Dato che tali verifiche non sono state fatte, sembra prematuro giungere a conclusin circa la discriminazione tra sessi.