Marco Minozzo e Annamaria Guolo
Laurea in Economia del Commercio Internazionale
Laurea in Economia e Amministrazione delle Imprese
Università degli Studi di Verona (sede di Vicenza)

Statistica Inferenziale
Soluzioni 1. Stima puntuale

Distribuzioni campionarie.

- Sia X una v.a. normale con valore atteso μ e varianza σ^2, $X \sim N(\mu, \sigma^2)$. Allora, $Z = (X - \mu)/\sigma$ si distribuisce come una normale standard, $Z \sim N(0, 1)$.
- Sia Z la v.a. normale standard. Allora la v.a. $T = Z^2$ segue una distribuzione chi-quadrato con un g.d.l., $T \sim \chi^2_1$. Più in generale, la somma di n v.a. normali standard al quadrato, indipendenti tra loro, dà luogo ad una χ^2_n.
- Sia T un v.a. χ^2_n. Allora $E(T) = n$ e $Var(T) = 2n$.
- Sia T_1 una v.a. distribuita come un χ^2_n e sia T_2 una v.a. distribuita come un χ^2_m, indipendente da T_1. Allora, $T_1 + T_2 \sim \chi^2_{n+m}$. La proprietà si generalizza alla somma di più v.a. con diversi g.d.l., purchè valga l'ipotesi di indipendenza.
- Sia Z una v.a. normale standard e sia T una v.a. chi-quadrato con n g.d.l. indipendente da Z. Allora, la v.a. $\frac{Z}{\sqrt{T/n}}$ si distribuisce come una t di Student con n g.d.l..
- Sia X una v.a. normale con media μ e varianza σ^2. Sia dato un campione casuale di dimensione n di osservazioni da X. La variabile media campionaria $\overline{X}_n$ e la variabile varianza campionaria corretta S^2 hanno distribuzione, rispettivamente,

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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

  Inoltre,

$$\frac{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\sigma^2}}}{\sqrt{S^2/\sigma^2}} \sim t_{n-1},$$

  sfruttando le distribuzioni di $\overline{X}_n$ e di S^2 sopra indicate e l'indipendenza tra le due v.a..

Teorema del limite centrale. Siano $X_1, \ldots, X_n$ variabili aleatorie indipendenti con valore atteso comune $E(X_i) = \mu$ e varianza comune finita $Var(X_i) = \sigma^2 < +\infty$, $i = 1, \ldots, n$. Sia $\overline{X}_n$ la variabile media campionaria. Allora, asintoticamente (per n sufficientemente grande) la variabile aleatoria

$$Y_n = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\sigma^2}}$$

si distribuisce come una normale standard.

Stima puntuale. Sia G uno *stimatore* di un parametro di interesse γ, vale a dire una statistica, funzione delle v.a. $X_1, \ldots, X_n$. Il valore assunto dallo stimatore in corrispondenza del particolare campione osservato si definisce *stima*.

Esempio 1. Sia X una v.a. $N(\mu, \sigma^2)$ e sia dato un campione casuale di numerosità n da X. Uno stimatore per μ è rappresentato dalla v.a. media campionaria, $\overline{X}_n$. Tale stimatore ha distribuzione $N(\mu, \sigma^2/n)$. La stima $\hat{\mu}$ di μ sulla base del campione osservato è data dalla media delle osservazioni del campione $\overline{x}_n$.

Esempio 2. Sia dato un campione di numerosità n da X, v.a. Bernoulliana di parametro p. Allora una stima di p è data dalla frequenza relativa dei successi, $\hat{p} = \sum_{i=1}^{n} x_i/n$. Inoltre, lo stimatore dato da $\sum_{i=1}^{n} X_i/n$ ha media p e varianza $p(1-p)/n$. Per n sufficientemente grande, sfruttando il teorema del limite centrale possiamo approssimare la distribuzione dello stimatore con una distribuzione $N(p, p(1-p)/n)$.

Sono di seguito elencate alcune proprietà desiderabili per uno stimatore.

- *Correttezza* (non distorsione). Uno stimatore è corretto (non distorto) se $E(G) = \gamma$, cioè se il suo valore atteso coincide con la quantità da stimare. In caso contrario, si dice distorto, con distorsione pari a $D = E(G) - \gamma$.

- *Efficienza*. Indichiamo con $MSE(G)$ (errore quadratico medio) la quantità $E[(G - \gamma)^2]$, che risulta essere pari anche a $D^2 + Var(G)$. Allora, dati due stimatori di γ, G_1 e G_2, diciamo che G_1 è più efficiente di G_2 se $MSE(G_1) < MSE(G_2)$.

- *Consistenza*. Uno stimatore G_n (si indica n a pedice per evidenziare la dipendenza del campione) si dice consistente (in senso debole) per γ se, $\forall \varepsilon > 0$,

$$\lim_{n \to \infty} P(|G_n - \gamma| < \varepsilon) = 1.$$

  Si tratta di una proprietà asintotica, cioè per $n \to \infty$, ovvero considerando una numerosità campionaria che aumenta indefinitamente (a livello pratico, per n sufficientemente alto).

- *Correttezza asintotica*. Uno stimatore G_n è asintoticamente non corretto per γ se

$$\lim_{n \to \infty} E(G_n) = \gamma.$$

  Anche in questo caso si tratta di una proprietà asintotica.

Esercizio A.

a) Sia C la variabile aleatoria che descrive il costo nel prossimo volo. Allora $P(C > 576,15) = P\Big(\frac{C}{6,00 \cdot 25} > \frac{576,15}{6,00 \cdot 25}\Big) = P(\chi^2_1 > 3,841)$, dove con χ^2_1 si è indicata una variabile aleatoria con distribuzione chi-quadrato con un grado di libertà. Dalle tavole si ricava che $P(\chi^2_1 > 3,841) = 0,05$.

b) Sia $C_T = \sum_{i=1}^{12} C_i$ il costo aggiuntivo complessivo di 12 voli. Essendo il costo di ogni volo indipendente da quello degli altri voli, $\frac{C_T}{6,00 \cdot 25}$ si distribuisce come una chi-quadrato con 12 g.d.l. Da cui si ricava che $P(C_T < 3.500,55) = P\Big(\frac{C_T}{6,00 \cdot 25} < \frac{3.500,55}{6,00 \cdot 25}\Big) = P(\chi^2_{12} < 23,337) = 1 - 0,025 = 0,975$.

c) Sia G_i il guadagno dell'i-esimo volo, allora $\mathrm{E}(G_i) = 1.500$ e $\mathrm{Var}(G_i) = 50^2$. Sia $D = G_i - G_{i+1}$ la differenza del guadagno in due voli successivi. Essa è distribuita come una v. a. normale con valore atteso pari a zero. Inoltre, essendo G_i e G_{i+1} indipendenti, $\mathrm{Var}(D) = \mathrm{Var}(G_i) + \mathrm{Var}(G_{i+1}) = 70,711^2$. Dalla simmetria della normale, $P(|D| > 120) = 2 \cdot P(\frac{D}{70,711} > \frac{120}{70,711}) = 2 \cdot (1 - \Phi(1,697)) = 2 \cdot (1 - 0,95513) = 0,08974$. Il valore 0,95513 è stato ottenuto per interpolazione lineare ponendo

$$\frac{\Phi(1,697) - \Phi(1,690)}{1,697 - 1,690} = \frac{\Phi(1,70) - \Phi(1,69)}{0,01}.$$

Sostituendo si ottiene $(\Phi(1,697) - 0,9633)/0,007 = (0,9554 - 0,9545)/0,01$. Da cui si ricava $\Phi(1,697) = 0,9545 + 0,01 \cdot 0,0009/0,007 = 0,95513$.

d) $G_T = \sum_{i=1}^{6} G_i$ è la somma dei guadagni di 6 voli. $\mathrm{E}(G_T) = 9.000$ e $\mathrm{Var}(G_T) = 6 \cdot \mathrm{Var}(G_i) = 122,474^2$. Da cui si ricava che $P(G_T > 9.500) = P(Z > 4,082) = 1 - \Phi(4,082) = 1 - 1 = 0$ (dove si è indicato con Z una variabile aleatoria con distribuzione normale standardizzata).

e) Ponendo $U_T = \sum_{i=1}^{6} U_i$ si ha che U_T è una v. a. normale con $\mathrm{E}(U_T) = 720$ e $\mathrm{Var}(U_T) = 216^2$. Inoltre, $C_T/(6 \cdot \mathrm{Var}(T))$ è una v. a. chi-quadrato con 6 g.d.l. Da cui si ricava che $R \cdot \sqrt{6 \cdot 6 \cdot \mathrm{Var}(T)}/\sqrt{\mathrm{Var}(U_T)}$ si distribuisce come una t di Student con 6 g.d.l., per cui $P(R > 13,99) = P(R\sqrt{6 \cdot 6 \cdot 25}/216 > 1,943)$. Dalle tavole si ricava che la probabilità richiesta è data da $P(t > 1,943) = 0,05$, dove con t si è indicata una v. a. con distribuzione t di Student.

f) Il 90-esimo percentile di una t di Student con 6 g.d.l. è pari a 1,44. Ovvero possiamo scrivere

$$P\left(R \cdot \frac{\sqrt{6 \cdot 6 \cdot \mathrm{Var}(T)}}{\sqrt{\mathrm{Var}(U_T)}} < 1,44\right) = 0,90,$$

da cui $r = 1,44\sqrt{\mathrm{Var}(U_T)}/\sqrt{6 \cdot 6 \cdot \mathrm{Var}(T)} = 10,368$.

Esercizio B.

a) Considerando $n = 30$, la probabilità richiesta è data da

$$P\Big(|\bar{X}_A - \mu_A| \geq 0,5 \cdot \sqrt{\sigma_A^2}\Big) = P\left(\frac{|\ \bar{X}_A - \mu_A\ |}{\sqrt{\sigma_A^2/n}} \geq 0,5 \cdot \sqrt{n}\right) = 2 \cdot (1 - \Phi(2,7386)) = 0,0062.$$

b) Considerando che $n = 30$, possiamo scrivere

$$0,95 = P\left(\frac{|\ \bar{X}_A - \mu_A\ |}{\sqrt{S_A^2}} \leq y\right) = P\left(\frac{\frac{|\bar{X}_A - \mu_A|}{\frac{\sigma_A}{\sqrt{n}}}}{\sqrt{\frac{S_A^2}{\sigma_A^2}}} \leq y \cdot \sqrt{n}\right) = P(|\ t_{29}\ | \leq y \cdot \sqrt{n}).$$

Dalle tavole si trova che $y \cdot \sqrt{n} = 2,045$ da cui $y = 2,045/\sqrt{30} = 0,3734$.

c) Considerando che $n = 30$, possiamo scrivere

$$0,95 = P\left(\frac{\mid \bar{X}_A - \mu_A \mid}{\sqrt{D_A^2}} \leq z\right) = P\left(\frac{\frac{\sqrt{n}|\bar{X}_A - \mu_A|}{\sigma_A}}{\sqrt{\frac{D_A^2}{\sigma_A^2}}} \leq z \cdot n\right) = P(\mid t_{30} \mid \leq z \cdot n).$$

Dalle tavole si ricava che $z \cdot n = 2,042$ da cui $z = 2,042/30 = 0,0681$.
d) Possiamo scrivere

$$P[(\bar{X}_B - \bar{X}_A) - (\mu_B - \mu_A) \leq 2] = P\left[\frac{(\bar{X}_B - \bar{X}_A) - (\mu_B - \mu_A)}{\sqrt{\frac{\sigma_A^2}{30} + \frac{\sigma_B^2}{20}}} \leq \frac{2}{\sqrt{\frac{\sigma_A^2}{30} + \frac{\sigma_B^2}{20}}}\right],$$

da cui, indicando con Z una variabile aleatoria normale standardizzata, si ha che la probabilità richiesta è data da $\Phi(Z \leq 1,92154) = 0,97267$.
e) Possiamo scrivere

$$0,05 = P(S_A^2 \geq d) = P\left(\frac{S_A^2}{\sigma_A^2} \cdot 29 \geq \frac{d}{\sigma_A^2} \cdot 29\right) = \left(\chi_{29}^2 \geq \frac{d}{10} \cdot 29\right).$$

Dalle tavole si trova che $29 \cdot d/30 = 42,557$ da cui $d = 14,6748$.

Esercizio C.
a) Sia X_A la v. a. che conta il numero di clienti attivi nel primo campione di numerosità $n_A = 50$ e sia $\hat{P}_A = \frac{X_A}{n_A}$. Per il teorema del limite centrale, $\hat{P}_A$ ha approssimativamente distribuzione normale con media p_A e varianza $\frac{p_A(1-p_A)}{n_A}$. Analogamente sia X_B la v. a. che conta il numero di clienti attivi nel secondo campione di numerosità $n_B = 50$ e sia $\hat{P}_B = \frac{X_B}{n_B}$. Per il teorema del limite centrale $\hat{P}_B$ ha approssimativamente distribuzione normale con media p_B e varianza $\frac{p_B(1-p_B)}{n_B}$. Quindi $\hat{P}_B - \hat{P}_A$ ha approssimativamente distribuzione normale con media $p_B - p_A$ e varianza pari alla somma delle due varianze asintotiche. Si vuole la probabilità di osservare uno scostamento fra le due frequenze campionarie superiore a quello osservato, cioè si vuole

$$P\left(\hat{P}_B - \hat{P}_A > \frac{27}{50} - \frac{23}{50}\right),$$

nel caso in cui $p_A = p_B = p$. In tale caso, approssimativamente, $\frac{\hat{P}_B - \hat{P}_A}{\sqrt{\frac{2p(1-p)}{50}}} \sim N(0;1)$ dove il valore di p non si conosce. Per un'estensione del teorema del limite centrale, l'approssimazione alla distribuzione normale continua a valere anche sostituendo a p nella espressione di sopra la stima $\hat{p}$. Avendo noi $\hat{p} = (\hat{p}_A 50 + \hat{p}_B 50)/100 = 50/100 = 0,5$ si ha $P(\hat{P}_B - \hat{P}_A > 0,54 - 0,46) = P\left(\frac{\hat{P}_B - \hat{P}_A}{\sqrt{2 \cdot \frac{0.5^2}{50}}} > \frac{0,54-0,46}{\sqrt{\frac{2 \cdot 0.5^2}{50}}}\right) \approx P(Z > 0,8) = 1 - \Phi(0,8) = 0,2119$, dove con Z si è indicata una v. a. con distribuzione normale standardizzata.

Esercizio D.
Entrambi gli stimatori di μ sono stimatori non distorti, essendo

$$E(T_1) = \frac{E(X_1 + X_2 + 3X_3)}{5} = \frac{\mu + \mu + 3\mu}{5} = \mu$$

e

$$E(T_2) = \frac{E(2X_1 + X_3)}{3} = \frac{2\mu + \mu}{3} = \mu.$$

Inoltre, T_1 ha varianza
$$Var(T_1) = \frac{\sigma^2 + \sigma^2 + 9\sigma^2}{25} = 11/25 \text{ e}$$
mentre T_2 ha varianza
$$Var(T_2) = \frac{4\sigma^2 + \sigma^2}{9} = 5/9.$$
Essendo i due stimatori entrambi non distorti, risulta preferibile quello con varianza minore, vale a dire T_1, essendo $11/25 < 5/9$. (Nel caso venisse meno la non distorsione degli stimatori, il confronto va fatto sulla base del MSE.)

Esercizio E.
a) Sia μ_R il valore atteso di R_i, allora $\mu_R = 5 + 0,5\mu_X$. Ora, $\bar{R}$ è corretto se $\mathrm{E}(\bar{R}) = \mu_R$. Per le proprietà di linearità del valore atteso, si ha $\mathrm{E}(\bar{R}) = \mathrm{E}(5+0,5\bar{X}) = 5+0,5\cdot\mathrm{E}(\bar{X}) = 5+0,5\mu_X = \mu_R$, per cui possiamo affermare che $\bar{R}$ è corretto per μ_R.
b) La varianza dello stimatore $\bar{R}$ è data da $\mathrm{Var}(\bar{R}) = 0,5^2 \cdot \mathrm{Var}(\bar{X}) = (0,5\cdot\sigma_X)^2/50 = 1$.
c) Lo stimatore $\bar{R}$ è consistente per $\mathrm{E}(R_i)$ se $\lim_{n\to\infty} P[|\bar{R} - \mu_R| \leq \epsilon] = 1$, per ogni $\epsilon > 0$, dove n è la numerosità campionaria. Dalla disuguaglianza di Tchebycheff si sa che $P[|\bar{R} - \mu_R| \leq k\sigma_{\bar{R}}] \geq 1 - 1/k^2$. Ponendo $\epsilon = k \cdot \sigma_{\bar{R}} = k \cdot 0,5\,\sigma_X/\sqrt{n}$, si ha che $\lim_{n\to\infty} P[|\bar{R} - \mu_R| \leq \epsilon] \geq \lim_{n\to\infty}\left(1 - \frac{0,25\sigma_X^2}{\epsilon^2 n}\right)$, ed essendo il limite a destra della disuguaglianza pari a 1 si ha la consistenza. Si noti che la probabilità non può mai superare 1, da cui l'uguaglianza.
d) Per determinare la numerosità campionaria richiesta, consideriamo che $P[|\bar{X} - \mu_X| \leq 10] = P\left[|Z| \leq \frac{10\sqrt{n}}{\sigma_X}\right] = 0,95$, dove $Z = (\bar{X} - \mu_X)/\sqrt{\sigma_X^2/n}$. Per il teorema del limite centrale si può approssimare la distribuzione di Z con una $N(0,1)$. Quindi, considerando che il percentile (della normale standardizzata) che lascia alla sua destra una probabilità di 0.975 è pari a 1,96, ponendo $1,96 = 10/\sqrt{\sigma_X^2/n}$, si ricava $\sqrt{n} = 0,196 \cdot \sqrt{200}$, ovvero $n \approx 8$. Si noti che la numerosità campionaria richiesta n è direttamente proporzionale alla varianza σ_X^2 e inversamente proporzionale all'errore di stima (che è stato posto pari a 10).

Esercizio F.
a) Si indichi con A l'affluenza alle urne, cioè il numero di elettori che andranno a votare. Allora, $A = Np$ e si vuole verificare se $N\hat{p}$ è uno stimatore corretto di A. Ora, $\mathrm{E}(N\hat{p}) = N\cdot\mathrm{E}(\hat{p}) = Np = A$, per cui $N\hat{p}$ è uno stimatore corretto per A.
b) Si vuole trovare la numerosità campionaria n tale che
$$P[|\hat{p} - p| \leq 0,05] = 0,95 = P\left[\frac{|\hat{p} - p|}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{0,05}{\sqrt{\frac{p(1-p)}{n}}}\right].$$
Essendo la varianza $p(1-p)/n$ incognita, questa si puo' porre pari al suo valore massimo $0,25/n$. Quindi, considerando che $P\left[\frac{|\hat{p}-p|\sqrt{n}}{0,5} \leq \frac{0,05\sqrt{n}}{0,5}\right] = 0,95$, usando l'approssimazione alla normale, si trova che l'evento $\frac{|\hat{p}-p|\sqrt{n}}{0,5} \leq \frac{0,05\sqrt{n}}{0,5}$ si verifica con probabilità pari a 0,95 se $0,05\sqrt{n}/0,5 = 1,96$. Da questo si ricava che la numerosità campionaria n cercata è approssimativamente pari a 384.

Esercizio G.
a) Il valore atteso dello stimatore T è dato da $E(T) = (3\mathrm{E}(X_1) + 4\mathrm{E}(X_2) + 2\mathrm{E}(X_3))/7 + a$. Essendo $\mathrm{E}(X_1) = \mathrm{E}(X_2) = \mathrm{E}(X_3) = \mu$, si ha che $\mathrm{E}(T) = (9/7)\mu + a$. Imponendo la condizione di correttezza $\mathrm{E}(T) = \mu$, si ricava che questa è verificata solo se $a = -(2/7)\mu$.
b) La distorsione di T per $a = 10$ è pari a $\mathrm{E}(T) - \mu = (2/7)\mu + 10$. Si noti che tale distorsione vale zero se $\mu = -35$.