12957

**Corso di Statistica - Prof. Fabio Zucca**
**II prova in itinere - 4 luglio 2014**
**Nome e cognome:** ..... **Matricola:** .....



**Esercizio 1** Arturo Cecato ogni mattina trova qualche difficoltà ad attraversare il corridoio che separa la camera dal bagno. Non essendo mai ben sveglio, 26 mattine su 30 va a sbattere contro un mobile. Da qualche tempo ha escogitato un rimedio che sta testando: appena alzatosi dal letto infila la testa in un secchio di ghiaccio: da quando ha adottato questa nuova procedura (21 mattine) Arturo ha colpito il mobile in 14 casi.

1. Calcolare un intervallo di confidenza al 90% bilatero per la probabilità $p$ di colpire il mobile una mattina qualsiasi con la nuova procedura.

2. Valutare se il rimedio di Arturo è davvero efficace, ovvero se la probabilità $p$ di colpire il mobile dopo la cura è significativamente inferiore a quella $p_0$ prima della cura, proponendo un test e calcolandone il P-value.

**Soluzione**. Osserviamo che $\bar{p}n = 14 \geq 5$ e $(1-\bar{p})n = 7 \geq 5$ quindi si possono applicare le approssimazioni gaussiane.

1. L'intervallo di confidenza a livello $\alpha = 0.9$ (si osservi che $q_{0.95} \approx 1.645$) è determinato dai suoi estremi

$$\begin{aligned} p_1^{\pm} &:= \overline{p}_1 \pm \sqrt{\overline{p}_1(1-\overline{p}_1)/n} \cdot q_{0.95} = 2/3 \pm \sqrt{2/3(1-2/3)/21} \cdot q_{0.95} \\ &\approx 2/3 \pm 0.1029 \cdot 1.645 = 0.666 \pm 0.1693 \end{aligned}$$

   da cui $IC(0.9) = [0.4974, 0.8359]$.

2. Il test da studiare è per una popolazione:

$$H_0 : p \geq p_0, \qquad H_1 : p < p_0.$$

   Infatti se riusciamo a rifiutare $H_0$ abbiamo dimostrato che la cura è evidentemente efficace. I dati $p_0 = 13/15$, $n = 21$, $\overline{p} = 14/n = 2/3$. La statistica test è:

$$U := \frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

   e il criterio di rifiuto di $H_0$ è $u < q_\alpha$. Il valore numerico di $U$ è $u = -3\sqrt{21/26} \approx -2.6962$, il P-value $\mathbb{P}(Z < u) = \mathbb{P}(Z > -u) = 1 - 0.9965 \approx 0.0035$ cioè 0.35%, quindi la cura sembra essere realmente efficace.

   Chi avesse interpretato la prima frazione come un campionamento su un campione di ampiezza $n_1 = 30$ avrebbe $\overline{p}_1 = 26/30$, $n_2 = 21$ e $\overline{p}_2 = 14/n_2 = 2/3$. La statistica test è:

$$U := \frac{\overline{p}_2 - \overline{p}_1}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n_1} + \sqrt{\overline{p}_2(1-\overline{p}_2)/n_2}}$$

   e il criterio di rifiuto di $H_0$ è $u < q_\alpha$. Il valore numerico di $U$ è $u = \frac{-3/5}{\sqrt{13/15^3 + 2/189}} \approx -1.6647$, il P-value $\mathbb{P}(Z < u) = \mathbb{P}(Z > -u) = 1 - \mathbb{P}(Z \leq -u) = 1 - 0.95201 = 0.04799$, quindi non abbiamo un $P$-value significativo.

12957

**Nome e cognome:** ..... **Matricola:** .....

**Esercizio 2** Vogliamo capire quale tra le zone A e B sia più inquinata. Da 10 campioni della zona A e 12 campioni dalla zona B si ottiene una media campionaria della concentrazione di $NO_3$ pari a $47mg/l$ per la zona A e di $50mg/l$ per la zona B; analogamente si ottengono le varianze campionarie $25(mg/l)^2$ e $24(mg/l)^2$ per le zone $A$ e $B$ rispettivamente.

1. Si può affermare che le due varianze sono differenti?
2. Alla luce del risultato del test precedente, impostare un nuovo test per stabilire se si possa affermare che la concentrazione media di $NO_3$ nella zona B sia superiore a quella della zona A al livello $\alpha = 0.05$.
3. Stimare il $P$-value del test appena svolto.

*Sugg: Chi non riuscisse a svolgere il primo punto, svolga i due successivi sotto l'ipotesi che le due varianze vere siano uguali.*

**Soluzione**.

1. Si tratta di impostare un test di confronto di due varianze dove $H_0\colon \sigma_A^2/\sigma_B^2 = 1$ e $H_1\colon \sigma_A^2/\sigma_B^2 \neq 1$. Definiamo la stima $f = s_A^2/s_B^2 = 25/24$ e risolviamo l'uguaglianza $f = F_{\alpha_0}(n_A - 1, n_B - 1) = F_{\alpha_0}(9, 11)$. Si osservi che $F_{0.75}(9, 11) = 1.528 > 25/24 > 1/1.582 = 1/F_{0.75}(11, 9) = F_{0.25}(11, 9)$. Da cui $0.25 < \alpha_0 < 0.75$ che implica $\bar{\alpha} = 2\min(\alpha_0, 1 - \alpha_0) > 0.5$ (un calcolatore restituirebbe il valore $\alpha_0 = 0.466473361$ e $\bar{\alpha} = 0.932946723$) pertanto non ci sono sufficienti evidenze statistiche per rifiutare l'uguaglianza delle due varianze vere.

2. Pur avendo ottenuto l'uguaglianza delle due varianze come conclusione debole, impostiamo un test sotto l'ipotesi di omoschedasticità dove $H_0\colon \mu_A - \mu_B \geq 0$ e $H_1\colon \mu_A - \mu_B < 0$. Definiamo $\bar{x}_9 = 47$ e $\bar{y}_{11} = 50$ le due medie campionarie delle concentrazioni delle zone $A$ e $B$ rispettivamente. Calcoliamo la varianza pesata

$$s_p^2 := \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{9 \cdot 25 + 11 \cdot 24}{20} = 24.45.$$

La stima della statistica test è quindi

$$t = \frac{\overline{x}_9 - \overline{y}_{11}}{\sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{-3}{\sqrt{24.45\left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{-3}{\sqrt{4.574166667}} = -1.402701503.$$

La regione di rifiuto a livello $\alpha$ è $\{t < t_\alpha(n_A + n_B - 2)\}$; essendo $t_{0.05}(20) = -1.724718$, non si può rifiutare $H_0$ a livello 0.05. Quindi non ci sono evidenze statistiche per affermare che la concentrazione media di $NO_3$ nella zona B è superiore a quella della zona A.

3. Il $P$-value soddisfa l'uguaglianza $t_{\bar{\alpha}}(20) = -1.402701503$. Dalle tavole si ottiene

$$t_{0.075}(20) = -1.497035 < t_{\bar{\alpha}}(20) = -1.402701503 < 0 = -1.325341 = t_{0.1}(20)$$

da cui $0.075 < \bar{\alpha} < 0.1$.

12957

**Nome e cognome:** ..... **Matricola:** .....

**Esercizio 3** Il guadagno annuale della società ferroviaria *DustyRail*, espresso in milioni di euro, viene registrato per 10 anni consecutivi ottenendo $\sum_{i=1}^{10} x_i = -21.5$ e $\sum_{i=1}^{10} x_i^2 = 127.25$.

1. Fornire un estremo superiore (intervallo unilatero) per la varianza $\sigma^2$ del guadagno al 99%.

2. Fornire un intervallo bilatero di confidenza per il valore atteso $\mu$ del guadagno al 95%.

3. Ci sono evidenze statistiche per concludere che il guadagno annuale medio sia negativo al 5%? Stimare il $P$-value del test.

4. Ci sono evidenze statistiche per concludere che le perdite annuali medie siano superiori a 2 milioni di euro?

**Soluzione**. Si calcolano immediatamente la media e la varianza campionarie come $\overline{x}_{10} = -2.15$ mentre $s^2_{10} = (127.25 - 10 \cdot (-2.15)^2)/9 \approx 9.003$.

1. Lintervallo di confidenza cercato ha la forma

$$\left[0, \frac{(n-1)s_n^2}{\chi^2_{1-\alpha}(n-1)}\right]$$

dove $\alpha$ è il livello di confidenza.In questo caso $\chi^2_{0.01}(9) = 2.087889$ da cui

$$\left[0, \frac{9 \cdot 9.003}{2.087889}\right] = \left[0, 38.80809756\right].$$

2. L'intervallo di confidenza per la media a varianza incognita ha come estremi $\bar{x}_n \pm t_{(1+\alpha)/2}(n-1)s_n/\sqrt{n}$. Essendo $t_{0.975}(9) = 2.262159$ e $\sqrt{s^2_{10}/10} = 0.9488413988$ si ricavano gli estremi $-2.15 \pm 2.228139 \cdot \sqrt{9.003/10} = -2.15 \pm 2.14643011$. L'intervallo è quindi

$$[-4.30356989, -0.003569890132].$$

3. Per stabilire se $\mu$ è negativo tramite una conclusione forte scegliamo $H_0 : \mu \geq 0$ e $H_1 : \mu_X < 0$. Utilizziamo la seguente regione di rifiuto a livello $\alpha$

$$T = \frac{\overline{X}_{10} - \mu_0}{S_{10}/\sqrt{n}} < t_\alpha(9)$$

dove la stima di $T$ è $t = -2.266$ (essendo $\mu_0 = 0$).

Stimiamo il $P$-value del test come segue: $-2.266 = t_{\overline{\alpha}}(9) \equiv -t_{1-\overline{\alpha}}(9)$ da cui $2.266 = t_{1-\overline{\alpha}}(9)$. Essendo

$$t_{0.975}(9) = 2.262159 < 2.266 < 2.821434 = t_{0.99}(9)$$

si ha $t_{0.975}(9) < t_{1-\overline{\alpha}}(9) < t_{0.99}(9)$ da cui $0.025 > \overline{\alpha} > 0.01$ (anche se di fatto $t_{0.975}(9) \approx t_{1-\overline{\alpha}}(9)$ da cui $0.025 \approx \overline{\alpha}$). Quindi al 5% concludiamo che il guadagno medio è negativo.

4. Si deve studiare il test $H_0 : \mu_X \geq -2$ e $H_1 : \mu_X < -2$. Si utilizza pertanto la seguente regione di rifiuto a livello $\alpha$

$$T = \frac{\overline{X}_{10} - \mu_0}{S_{10}/\sqrt{n}} < t_\alpha(9)$$

dove la stima di $T$ è $t = -0.1581$ (essendo $\mu_0 = -2$).

Similmente al caso precedente, per il secondo test si ottiene la stima $\overline{\alpha} \in (0.2, 0.5)$ (ricordiamo che $t_\alpha(n) < 0$ (risp. $> 0$) se e solo se $\alpha < 1/2$ (risp. $> 1/2$)). In questo caso non ci sono evidenze statistiche per concludere che le perdite medie annuali siano superiori a 2 milioni di euro.

12957

**Nome e cognome:** ..... **Matricola:** .....

**Esercizio 4** Al casello di Milano dell'autostrada A4 si controlla il numero di auto che arrivano in un minuto (si ripete l'osservazione per 106 minuti) ottenendo la seguente tabella

| n. auto | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| occorrenze | 0 | 0 | 1 | 3 | 5 | 7 | 13 | 12 | 8 | 9 | 13 | 10 | 5 | 6 | 4 | 5 | 4 | 0 | 1 |

1. Calcolare media campionaria, mediana e quartili dei dati.
2. Si consideri ora una variabile di poisson $X \sim \mathcal{P}(\lambda)$; quanto vale il suo valore atteso?
3. Si determini, tramite un test opportuno, se una legge di Poisson sia o meno appropriata per descrivere i dati a livello del 5%. Si utilizzino le seguenti classi: $[0, 5)$, $[5, 6)$, $[6, 7)$, $[7, 8)$, $[8, 9)$, $[9, 10)$, $[10, 11)$, $[11, 12)$, $[12, 13)$, $[13, 14)$, $[14, +\infty)$.
4. Si stimi il P-value del test precedente.

**Soluzione**.

1. Facilmente si calcola la media $\bar{x}_{106} = \frac{1}{106}\sum_{i=0}^{18} i \cdot f_{ass}(i) = 9.09$. Utilizzando la definizione di quartili, ricordando che la mediana è il secondo quartile ed utilizzando la frequenza cumulativa

| n. auto | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| occorrenze | 0 | 0 | 1 | 3 | 5 | 7 | 13 | 12 | 8 | 9 | 13 | 10 | 5 | 6 | 4 | 5 | 4 | 0 | 1 |
| *freq. cumulativa* | 0 | 0 | 1 | 4 | 9 | 16 | 29 | 41 | 49 | 58 | 71 | 81 | 86 | 92 | 96 | 101 | 105 | 105 | 106 |

   otteniamo $Q_1 = 6$, $Q_2 = 9$, $Q_3 = 11$.

2. Il valore atteso di $X$ coincide con il suo parametro $\lambda$.

3. Predisponiamo un test di adattamento. Poichè il parametro della legge di Poisson è incognito, dobbiamo stimarlo tramite la media campionaria; quindi utilizziamo $\lambda = \bar{x}_{106} = 9.09$.

   Definiamo gli 11 intervalli $I_i$ come $[0,5)$, $[5,6)$, $[6,7)$, $[7,8)$, $[8,9)$, $[9,10)$, $[10,11)$, $[11,12)$, $[12,13)$, $[13,14)$, $[14,+\infty)$ e calcoliamo le probabilità teoriche $p_i = \sum_{j\in I_i} \frac{\lambda^i}{i!}e^{-\lambda} = \sum_{j\in I_i} \frac{9.09^i}{i!}e^{-9.09}$ (chiaramente $p_{11} = 1 - \sum_{i=1}^{10} p_i$). Il numero atteso di elementi del campione che cadono in $I_i$ sarà quindi $np_i$. Completiamo quindi la tabella

| | $I_1$ | $I_2$ | $I_3$ | $I_4$ | $I_5$ | $I_6$ | $I_7$ | $I_8$ | $I_9$ | $I_{10}$ | $_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_{ass}(i)$ | 9 | 7 | 13 | 12 | 8 | 9 | 13 | 10 | 5 | 6 | 14 |
| $p_i$ | 0.052 | 0.058 | 0.088 | 0.115 | 0.130 | 0.132 | 0.120 | 0.099 | 0.075 | 0.053 | 0.079 |
| $np_i$ | 5.50 | 6.17 | 9.35 | 12.15 | 13.81 | 13.96 | 12.70 | 10.50 | 7.95 | 5.56 | 8.34 |
| $f_{ass}(i)^2/np_i$ | 14.7 | 7.94 | 18.07 | 11.85 | 4.63 | 5.80 | 13.31 | 9.53 | 3.14 | 6.47 | 24.49 |

   Utilizziamo la statistica test $q = \sum_{i=1}^{11} f_{ass}(i)^2/np_i - n = 118.97 - 106 = 12.97$ e rifiutiamo a livello $\alpha$ se $q > \chi^2_{1-\alpha}(9)$ (11 classi e 1 parametro stimato). In questo caso $\chi^2_{0.95} = 16.91896$ pertanto non si può rifiutare l'adattamento ad una Poisson al 5%.

4. Dalle tavole $\chi^2_{0.9}(9) = 14.68366 > 12.97 = \chi^2_{1-\bar{\alpha}}(9) > 4,168156 = \chi^2_{0.1}(9)$ da cui $0.9 > 1 - \bar{\alpha} > 0.1$ cioè $0.1 < \bar{\alpha} < 0.9$. Un calcolatore restituirebbe un valore approssimato $\bar{\alpha} = 1 - F_{\chi^2(9)}(12.97) = 0.163783983$.