Nome e cognome: _____

indirizzo email: _____

Esame di Statistica – Sample Test

1. Una dieta troppo ricca di sodio può essere nociva, soprattutto in caso di disturbi renali.
Per 18 marche di acqua minerale presenti sul mercato, suddivise nelle due tipologie liscia e frizzante, è stata misurata la concentrazione di sodio.

Liscia:	3.0	1.8	4.4	2.2	3.0	3.3	3.0	3.3	2.7	4.7
Frizzante:	4.1	3.6	3.6	3.1	4.5	4.4	3.6	4.6		

Assumendo che la concentrazione di sodio nell'acqua segua una distribuzione normale, costruire un test di ipotesi alla luce del quale si vuole stabilire se è opportuno consigliare, a persone con disturbi renali, di non bere acqua frizzante.

(a) Stabilire quale debba essere l'ipotesi nulla H_0 e quale l'ipotesi alternativa H_1.

(b) Verificare le ipotesi di cui al punto (a) con un livello di significatività $\alpha = 0.1$.

(c) Che tipo di assunzione è stata fatta sulle varianze?

(d) Delineare (senza eseguire) una procedura per stabilire se l'assunzione di cui al punto (c) è plausibile o meno.

---

argomentare tutti i passaggi in modo adeguato

(a) L'ipotesi alternativa deve essere quella che porta a consigliare di eliminare l'acqua frizzante, quindi corrispondente al fatto che l'acqua frizzante contiene mediamente più sodio.

In sostanza, se μ_F e μ_L sono le quantità medie di sodio contenute nell'acqua frizzante e liscia rispettivamente, l'ipotesi alternativa sarà $\mu_F > \mu_L$. Il sistema di ipotesi da verificare è

$$H_0 : \mu_F - \mu_L \leq 0 \qquad \textit{contro} \qquad H_1 : \mu_F - \mu_L > 0.$$

(b) Nell'ipotesi che le varianze delle due popolazioni siano uguali, sappiamo che

$$\frac{\bar{X}_F - \bar{X}_L - (\mu_F - \mu_L)}{\sqrt{S_p^2 \left(\frac{1}{n_F} + \frac{1}{n_L}\right)}} \sim t_{n_F+n_L-2}$$

dove $\bar{X}_F$ è la media campionaria relativa alle acque frizzanti, $\bar{X}_L$ quella relativa alle acque lisce e $S_p^2 = \frac{(n_F-1)S_F^2+(n_L-1)S_L^2}{n_F+n_L-2}$ è la varianza "pooled".

Rifiutiamo l'ipotesi nulla se la quantità di sinistra, in corrispondenza di $\mu_F = \mu_L$ è "grande", per cui la regione critica è

$$R = \left\{ \frac{\bar{X}_F - \bar{X}_L}{\sqrt{S_p^2 \left(\frac{1}{n_F} + \frac{1}{n_L}\right)}} > t_{n_F+n_L-2;0.10} \right\}.$$

Poiché abbiamo $t_{16;0.10} = 1.337$, e, dai dati campionari

$$\bar{x}_F = 3.94 \qquad s_F^2 = 0.29 \qquad n_F = 8$$
$$\bar{x}_L = 3.14 \qquad s_L^2 = 0.77 \qquad n_L = 10$$

per cui $s_p^2 = 0.56$. La statistica test è $\frac{0.8}{\sqrt{0.56\left(\frac{1}{8}+\frac{1}{10}\right)}} = 2.25$ e quindi il campione osservato è nella regione critica per cui rifiutiamo l'ipotesi nulla.

(c-d) Abbiamo ipotizzato l'uguaglianza tra le varianze, cosa che andrebbe verificata con un test basato sulla F di Fisher.

2. Un docente sa, dall'esperienza passata, che il punteggio (in centesimi) all'esame finale dei suoi studenti ha valore atteso 77 e deviazione standard 15. In questo semestre il docente ha due classi, una di 64 studenti e una di 25.

Quanto vale approssimativamente la probabilità che il punteggio medio della classe da 25 superi quello della classe da 64?

---

argomentare tutti i passaggi in modo adeguato

Indichiamo con $\bar{X}$ il punteggio medio della classe da 25 e con $\bar{Y}$ quello relativo alla classe da 64. Sono entrambe medie campionarie, quindi hanno valore atteso pari a quello della popolazione (77) e varianza pari a quella della popolazione divisa per la numerosità. Inoltre, il Teorema del Limite Centrale consente di approssimare le loro distribuzioni di $\bar{X}$ e $\bar{Y}$ con delle normali. Quindi

$$\bar{X} \sim N(\mu = 77, \sigma^2_{\bar{X}} = \frac{15^2}{25}) \qquad \bar{Y} \sim N(\mu = 77, \sigma^2_{\bar{Y}} = \frac{15^2}{64})$$

Per l'indipendenza tra le due classi,

$$\bar{X} - \bar{Y} \sim N(0, \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}} = 12.52)$$

per cui la probabilità (approssimata) richiesta è

$$P(\bar{X} > \bar{Y}) \;\; = \;\; P(\bar{X} - \bar{Y} > 0) = P\left(\frac{\bar{X} - \bar{Y} - 0}{\sqrt{12.52}} > 0\right) = P(Z > 0) = \frac{1}{2}.$$

3. Sia $X_1, \cdots, X_n$ un campione casuale estratto da una popolazione X con distribuzione uniforme sull'intervallo $(\theta, 2)$.

(a) Determinare lo stimatore dei momenti $\hat{\theta}_M$ (ottenuto uguagliando il valore atteso della popolazione $E(X)$ alla media campionaria $\bar{X}$).

(b) Calcolare la distorsione (bias) e l'errore quadratico medio (MSE) dello stimatore individuato al punto (a).

(c) Il campione osservato è

$$x_1 = 1.1 \qquad x_2 = 1.8 \qquad x_3 = 0.9 \qquad x_4 = 1.2 \qquad x_5 = 1.5 \qquad x_6 = 1.3$$

Calcolare la stima corrispondente allo stimatore individuato al punto (a).

---

argomentare tutti i passaggi in modo adeguato

Sappiamo che per una popolazione X con distribuzione uniforme su $(\theta, 2)$ si ha

$$E(X) = \frac{\theta + 2}{2} \qquad V(X) = \frac{(2-\theta)^2}{12}$$

(a) Lo stimatore dei momenti è soluzione dell'equazione $\bar{X} = \frac{\theta+2}{2}$, per cui

$$\hat{\theta}_M = 2\bar{X} - 2.$$

(b) Il valore atteso di $\bar{X}$ è $E(\bar{X}) = E(X) = \frac{\theta+2}{2}$, per cui

$$E(\hat{\theta}_M) = 2E(\bar{X}) - 2 = \theta$$

quindi lo stimatore è non distorto.

Ricordiamo che per uno stimatore non distorto l'MSE coincide con la varianza, per cui

$$MSE(\hat{\theta}_M) \;=\; V((\hat{\theta}_M)) = V(2\bar{X} - 2) = 4V(\bar{X}) = 4\frac{V(X)}{n} = \frac{(2-\theta)^2}{3n}.$$

(c) La media del campione osservato è $\bar{x} = 1.3$, quindi la stima dei momenti è $2 \cdot 1.3 - 2 = 0.6$.

4. Il modello di regressione

$$Y = \beta x + e \qquad \text{con } e \sim N(0, \sigma^2)$$

è detto "regressione attraverso l'origine" perché la retta corrispondente passa per l'origine degli assi. Supponendo di avere un campione di n coppie di osservazioni (x_i, Y_i),

(a) determinare lo stimatore dei minimi quadrati B di β;

(b) determinare la distribuzione dello stimatore B.

---

argomentare tutti i passaggi in modo adeguato

(a) Bisogna minimizzare $g(\beta) = \sum_{i=1}^n (Y_i - \beta x_i)^2$.
Derivando rispetto a β *otteniamo*

$$\frac{dg}{d\beta} = \sum_{i=1}^n 2(Y_i - \beta x_i)(-x_i)$$

L'equazione da risolvere quindi è

$$\frac{dg}{d\beta} = 0 \quad \Longleftrightarrow \quad \sum_{i=1}^n x_i Y_i - \beta \sum_{i=1}^n x_i^2 = 0 \quad \Longleftrightarrow \quad \beta = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

e quindi lo stimatore dei minimi quadrati è

$$B = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

(b) Lo stimatore B *è combinazione lineare di normali indipendenti (le* Y_i*), ciascuna delle quali ha valore atteso* $E(Y_i) = \beta x_i$ *e varianza* $V(Y_i) = \sigma^2$*. Quindi* B *ha distribuzione normale con parametri*

$$\begin{aligned} E(B) &= E\left(\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right) = \frac{\sum_{i=1}^n x_i E(Y_i)}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i \beta x_i}{\sum_{i=1}^n x_i^2} = \frac{\beta \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} = \beta \\ V(B) &= V\left(\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right) = \frac{\sum_{i=1}^n x_i^2 V(Y_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sum_{i=1}^n x_i^2 \sigma^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2} \end{aligned}$$