

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____

Nome: _____

Matricola: _____

Corso di studi: _____

A.A.: 2021/2022

Docente: Gianluca Orlando

Appello: gennaio 2023

Data: 26/01/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I seguenti dati indicano la relazione tra velocità di lettura (parole al minuto) e il numero di settimane trascorse in un programma di lettura veloce per 10 studenti:

settimane	2	3	8	11	4	5	9	7	5	7
velocità di lettura	21	42	102	130	52	57	105	85	62	90

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule dei coefficienti) e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Soluzione. 1. Lo scatterplot è rappresentato in Figura 1.

2. Denotiamo con $(x_1, y_1), \dots, (x_n, y_n)$, $n = 10$, i dati del campione. Cerchiamo la retta di equazione

$$y = ax + b$$

che meglio approssima i dati, utilizzando il metodo dei minimi quadrati. Vogliamo minimizzare l'errore

$$r(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2.$$

Imponiamo che il gradiente rispetto ad (a, b) sia nullo, ovvero,

$$0 = \partial_a r(a, b) = -2 \sum_{i=1}^n (y_i - ax_i - b)x_i = -2 \sum_{i=1}^n (x_i y_i - ax_i^2 - bx_i),$$

$$0 = \partial_b r(a, b) = -2 \sum_{i=1}^n (y_i - ax_i - b)$$

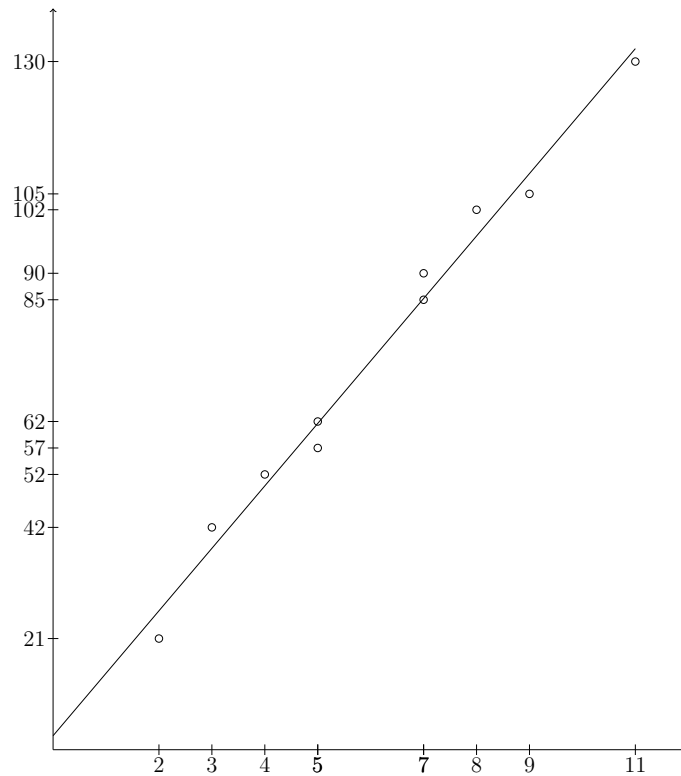


Figura 1: Scatterplot e retta di regressione lineare.

Dalla seconda equazione segue che

$$nb = \sum_{i=1}^n (y_i - ax_i) \implies b = \bar{y} - a\bar{x}.$$

Sostituendo nella prima,

$$\begin{aligned} \sum_{i=1}^n (x_i y_i - ax_i^2 - bx_i) &= 0 \implies \sum_{i=1}^n (x_i y_i - ax_i^2 - x_i \bar{y} + a\bar{x} x_i) = 0 \\ &\implies a \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^n x_i y_i - n\bar{x} \bar{y} \\ &\implies a = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}. \end{aligned}$$

Completiamo la tabella con i valori necessari a calcolare a e b :

											somma
x_i	2	3	8	11	4	5	9	7	5	7	61
y_i	21	42	102	130	52	57	105	85	62	90	746
x_i^2	4	9	64	121	16	25	81	49	25	49	443
y_i^2	441	1764	10404	16900	2704	3249	11025	7225	3844	8100	65656
$x_i y_i$	42	126	816	1430	208	285	945	595	310	630	5387

Pertanto $\bar{x} = 61/10 = 6.1$ e $\bar{y} = 746/10 = 74.6$. Segue che

$$a = \frac{5387 - 10 \cdot 6.1 \cdot 74.6}{443 - 10 \cdot 6.1^2} = \frac{836.4}{70.9} \simeq 11.80,$$

$$b = 74.6 - 11.80 \cdot 6.1 \simeq 2.64,$$

ovvero, la retta di regressione lineare ha equazione

$$y = 11.80x + 2.64.$$

3. Per calcolare il coefficiente di correlazione lineare usiamo la formula

$$\begin{aligned}\rho_{x,y} &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \sqrt{\sum_{i=1}^n y_i^2 - n \bar{y}^2}} = \frac{5387 - 10 \cdot 6.1 \cdot 74.6}{\sqrt{443 - 10 \cdot 6.1^2} \sqrt{65656 - 10 \cdot 74.6^2}} \\ &= \frac{836.4}{\sqrt{70.9} \sqrt{10004.4}} \simeq 0.9931.\end{aligned}$$

Esercizio 2. (8 punti) Si consideri un vettore aleatorio discreto (X_1, X_2) con funzione di probabilità congiunta data dalla seguente tabella:

X_1	0	1	2	3
X_2				
0	$a_{00}/8$	$a_{10}/8$	$a_{20}/8$	$a_{30}/8$
1	$a_{01}/8$	$a_{11}/8$	$a_{21}/8$	$a_{31}/8$

dove $a_{00}, a_{10}, a_{20}, a_{30}, a_{01}, a_{11}, a_{21}, a_{31} \geq 0$.

1. Trovare i valori espliciti di a_{ij} nella tabella sapendo che:

- X_1 ha legge binomiale $B(n, p)$ con parametri $n = 3$ e $p = \frac{1}{2}$.
- X_2 ha legge di Bernoulli $Be(q)$ con parametro $q = \frac{7}{8}$ ($1 = \text{successo}$).
- $\mathbb{P}(\{X_1 = 3\} | \{X_2 = 1\}) = \frac{1}{14}$.
- $\mathbb{P}(\{X_1 + X_2 = 1\}) = \frac{1}{8}$.
- $\mathbb{P}(\{X_1 = 2\}, \{X_2 = 0\}) = \frac{1}{16}$.

(Suggerimento: usare le condizioni nell'ordine in cui sono fornite.)

2. Calcolare la covarianza di X_1 e X_2 .

3. Le variabili aleatorie X_1 e X_2 sono indipendenti?

Soluzione. 1. Poiché $X_1 \sim B(n, p)$ con $n = 3$ e $p = \frac{1}{2}$, abbiamo che (scrivendo le probabilità marginali)

$$\frac{1}{8} = (1-p)^3 = \mathbb{P}(\{X_1 = 0\}) = \frac{1}{8}(a_{00} + a_{01}), \quad (1)$$

$$\frac{3}{8} = \binom{3}{1} p (1-p)^2 = \mathbb{P}(\{X_1 = 1\}) = \frac{1}{8}(a_{10} + a_{11}), \quad (2)$$

$$\frac{3}{8} = \binom{3}{2} p^2 (1-p) = \mathbb{P}(\{X_1 = 2\}) = \frac{1}{8}(a_{20} + a_{21}), \quad (3)$$

$$\frac{1}{8} = p^3 = \mathbb{P}(\{X_1 = 3\}) = \frac{1}{8}(a_{30} + a_{31}). \quad (4)$$

Poiché $X_2 \sim \text{Be}(\frac{7}{8})$ abbiamo che

$$\frac{1}{8} = \mathbb{P}(\{X_2 = 0\}) = \frac{1}{8}(a_{00} + a_{10} + a_{20} + a_{30}), \quad (5)$$

$$\frac{7}{8} = \mathbb{P}(\{X_2 = 1\}) = \frac{1}{8}(a_{01} + a_{11} + a_{21} + a_{31}). \quad (6)$$

Osservo che la condizione

$$\frac{1}{8}(a_{00} + a_{10} + a_{20} + a_{30} + a_{01} + a_{11} + a_{21} + a_{31}) = 1, \quad (7)$$

che si ottiene imponendo che la somma di tutte le probabilità sia 1, è superflua. Infatti si ottiene già sommando (1)–(4). Anche la (6) è superflua: segue automaticamente sottraendo (5) da (7). Abbiamo quindi il sistema

$$\begin{cases} a_{00} + a_{01} = 1, \\ a_{10} + a_{11} = 3, \\ a_{20} + a_{21} = 3, \\ a_{30} + a_{31} = 1, \\ a_{00} + a_{10} + a_{20} + a_{30} = 1. \end{cases} \quad (8)$$

Seguiamo il suggerimento e sfruttiamo una condizione alla volta. Da $\mathbb{P}(\{X_1 = 3|X_2 = 1\}) = \frac{1}{14}$ segue, usando la formula della probabilità condizionata e (6), che

$$\frac{1}{14} = \mathbb{P}(\{X_1 = 3|X_2 = 1\}) = \frac{\mathbb{P}(\{X_1 = 3, X_2 = 1\})}{\mathbb{P}(\{X_2 = 1\})} = \frac{a_{31}/8}{7/8} = \frac{a_{31}}{7} \implies a_{31} = \frac{1}{2}.$$

Sostituendo in (8):

$$\begin{cases} a_{00} + a_{01} = 1, \\ a_{10} + a_{11} = 3, \\ a_{20} + a_{21} = 3, \\ a_{30} = \frac{1}{2}, \\ a_{00} + a_{10} + a_{20} = \frac{1}{2}, \\ a_{31} = \frac{1}{2}. \end{cases} \quad (9)$$

Utilizziamo la condizione $\mathbb{P}(\{X_1 + X_2 = 1\}) = \frac{1}{8}$. Si ha che

$$\frac{1}{8} = \mathbb{P}(\{X_1 + X_2 = 1\}) = \mathbb{P}(\{X_1 = 0, X_2 = 1\}) + \mathbb{P}(\{X_1 = 1, X_2 = 0\}) = \frac{1}{8}(a_{01} + a_{10})$$

da cui

$$a_{01} + a_{10} = 1.$$

Sostituendo nella prima equazione nel sistema (9) otteniamo che

$$a_{01} + a_{10} = a_{00} + a_{01} \implies a_{10} = a_{00}.$$

Il sistema diventa quindi

$$\left\{ \begin{array}{l} a_{00} = a_{10} , \\ a_{10} + a_{11} = 3 , \\ a_{20} + a_{21} = 3 , \\ a_{30} = \frac{1}{2} , \\ 2a_{00} + a_{20} = \frac{1}{2} , \\ a_{31} = \frac{1}{2} , \\ a_{01} + a_{10} = 1 , \end{array} \right. \quad (10)$$

L'ultima condizione $\mathbb{P}(\{X_1 = 2\}, \{X_2 = 0\}) = \frac{1}{16}$ implica che $a_{20}/8 = \frac{1}{16}$, cioè $a_{20} = \frac{1}{2}$. Quindi

$$\left\{ \begin{array}{l} a_{00} = a_{10} , \\ a_{10} + a_{11} = 3 , \\ a_{20} + a_{21} = 3 , \\ a_{30} = \frac{1}{2} , \\ 2a_{00} + a_{20} = \frac{1}{2} , \\ a_{31} = \frac{1}{2} , \\ a_{01} + a_{10} = 1 , \\ a_{20} = \frac{1}{2} . \end{array} \right. \Rightarrow \left\{ \begin{array}{l} a_{10} = 0 , \\ a_{11} = 3 , \\ a_{21} = \frac{5}{2} , \\ a_{30} = \frac{1}{2} , \\ a_{00} = 0 , \\ a_{31} = \frac{1}{2} , \\ a_{01} = 1 , \\ a_{20} = \frac{1}{2} . \end{array} \right. \quad (11)$$

La tabella completa è quindi

X_1	0	1	2	3
X_2				
0	0	0	1/16	1/16
1	1/8	3/8	5/16	1/16

2. Per calcolare la covarianza osserviamo che $\mathbb{E}(X_1) = np = \frac{3}{2}$ mentre $\mathbb{E}(X_2) = q = \frac{7}{8}$, per come sono le leggi delle variabili aleatorie. Dobbiamo calcolare $\mathbb{E}(X_1 \cdot X_2)$. Il range di $X_1 \cdot X_2$ è dato da $\{0, 1, 2, 3\}$, quindi

$$\begin{aligned} \mathbb{E}(X_1 \cdot X_2) &= 0 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 0\}) + 1 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 1\}) + 2 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 2\}) + 3 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 3\}) \\ &= 1 \cdot \mathbb{P}(\{X_1 = 1, X_2 = 1\}) + 2 \cdot \mathbb{P}(\{X_1 = 2, X_2 = 1\}) + 3 \cdot \mathbb{P}(\{X_1 = 3, X_2 = 1\}) \\ &= \frac{3}{8} + 2 \frac{5}{16} + 3 \frac{1}{16} = \frac{19}{16} . \end{aligned}$$

Concludiamo che

$$\text{Cov}(X_1, X_2) = \mathbb{E}(X_1 \cdot X_2) - \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) = \frac{19}{16} - \frac{3}{2} \cdot \frac{7}{8} = -\frac{1}{8} .$$

3. Le variabili non sono indipendenti: la covarianza non è nulla (il fatto che la covarianza si annulli è una condizione necessaria per l'indipendenza).

Esercizio 3. (7 punti) Un'azienda produce uno smartphone con una vita media di 4 anni, dopodiché si rompe. Assumiamo che la vita dello smartphone (misurata in anni) sia una variabile aleatoria con legge esponenziale.

1. Acquisti uno smartphone. Qual è la probabilità che funzioni per più di 6 anni?
2. Acquisti uno smartphone. Passano 3 anni e funziona ancora. Qual è la probabilità che funzioni in tutto per più di 6 anni, sapendo che è successo il fatto precedente?
3. Acquisti tre smartphone (assumiamo che le vite dei tre smartphone siano indipendenti). qual è la probabilità che almeno due dei tre funzionino per più di 6 anni?
4. (**Bonus**) Acquisti due smartphone (assumiamo che le vite dei due smartphone siano indipendenti). Ne usi solo uno, finché si rompe (assumiamo che intanto lo smartphone non utilizzato non perda anni di vita). Poi inizi a usare l'altro (che ha una vita media di 4 anni). Chiamiamo vita cumulata la somma delle vite dei due smartphone. Qual è la probabilità che la vita cumulata sia più di 12 anni?

Soluzione. 1. Sia $X \sim \text{Exp}(\lambda)$ la vita (in anni) dello smartphone. Ricordiamo che $\mathbb{E}(X) = \frac{1}{\lambda}$, quindi $\frac{1}{\lambda} = 4$, cioè $\lambda = \frac{1}{4}$. La probabilità che lo smartphone funzioni più di 6 anni è

$$\mathbb{P}(\{X \geq 6\}) = \int_6^{+\infty} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x} \right]_6^{+\infty} = e^{-\lambda \cdot 6} = e^{-3/2} \simeq 22.31\%.$$

2. La legge esponenziale gode della proprietà di assenza di memoria, quindi

$$\mathbb{P}(\{X \geq 6\} | \{X \geq 3\}) = \mathbb{P}(\{X \geq 3\}) = \int_3^{+\infty} \lambda e^{-\lambda x} dx = e^{-3/4} = 47.24\%.$$

3. Introduciamo la variabile $Y \sim B(3, p)$ con $p = \mathbb{P}(\{X \geq 6\})$ (durata più di 6 anni è un successo). Ci viene chiesta la probabilità di almeno due successi in 3 prove, ovvero

$$\begin{aligned} \mathbb{P}(\{Y \geq 2\}) &= \mathbb{P}(\{Y = 2\}) + \mathbb{P}(\{Y = 3\}) = \binom{3}{2} p^2 (1-p) + \binom{3}{3} p^3 \\ &= 3(e^{-3/2})^2 (1 - e^{-3/2}) + e^{-9/4} \simeq 12.71\%. \end{aligned}$$

Possiamo anche risolvere l'esercizio direttamente senza l'uso della binomiale. Denotiamo con $X_1, X_2, X_3 \sim \text{Exp}(\frac{1}{4})$ le vite dei tre smartphone (indipendenti). Almeno due smartphone funzionano per più di 6 anni se si verificano almeno due (anche tre) delle disuguaglianze $X_1 \geq 6, X_2 \geq 6, X_3 \geq 6$, ovvero

$$\begin{aligned} &\mathbb{P}(\{X_1 \geq 6\} \cap \{X_2 \geq 6\} \cap \{X_3 < 6\}) \\ &+ \mathbb{P}(\{X_1 \geq 6\} \cap \{X_2 < 6\} \cap \{X_3 \geq 6\}) \\ &+ \mathbb{P}(\{X_1 < 6\} \cap \{X_2 \geq 6\} \cap \{X_3 \geq 6\}) \\ &+ \mathbb{P}(\{X_1 \geq 6\} \cap \{X_2 \geq 6\} \cap \{X_3 \geq 6\}). \end{aligned}$$

Usando l'indipendenza delle tre variabili si ottiene lo stesso risultato.

4. Denotiamo con $X_1, X_2 \sim \text{Exp}(\lambda)$ i tempi di vita dei due smartphone. Siamo interessati alla variabile aleatoria $X_1 + X_2$. Ricordiamo che $X_1, X_2 \sim \text{Gamma}(1, \lambda)$ e sono indipendenti,

pertanto $X_1 + X_2 \sim \text{Gamma}(2, \lambda)$. Integrando per parti e usando il fatto che $\Gamma(2) = (2-1)! = 1$, calcoliamo

$$\begin{aligned}\mathbb{P}(\{X_1 + X_2 \geq 12\}) &= \int_{12}^{+\infty} \frac{\lambda^2}{\Gamma(2)} x e^{-\lambda x} dx = \lambda \left[-x e^{-\lambda x} \right]_{12}^{+\infty} + \int_{12}^{+\infty} \lambda e^{-\lambda x} dx \\ &= 12\lambda e^{-12\lambda} + \left[-e^{-\lambda x} \right]_{12}^{+\infty} = 12\lambda e^{-12\lambda} + e^{-12\lambda} = 4e^{-3} \simeq 19.91\%.\end{aligned}$$

In alternativa, se non si ricordano le proprietà della legge Gamma, si può usare la formula per la densità della somma di due variabili indipendenti:

$$f_{X_1+X_2}(x) = f_{X_1} * f_{X_2}(x) = \int_{\mathbb{R}} f_{X_1}(y) f_{X_2}(x-y) dy = \int_0^x \lambda e^{-\lambda y} \lambda e^{-\lambda(x-y)} dy = \lambda^2 x e^{-\lambda x}$$

e continuare con il conto di sopra.

Esercizio 4. (7 punti) Un'azienda produce una margarina dietetica per cui si sa, fino a prova contraria, che il livello di acidi grassi polinsaturi (in percentuale) ha una deviazione standard di 1.2. È stata proposta una nuova tecnica di produzione del prodotto, che tuttavia comporta un costo aggiuntivo. La direzione autorizzerà un cambiamento nella tecnica di produzione se si riesce a mostrare che la deviazione standard del livello di acidi grassi polinsaturi con il nuovo processo è significativamente inferiore a 1.2. Un campione del lotto ottenuto con il nuovo metodo ha prodotto le seguenti percentuali di livello di acidi grassi polinsaturi:

16.8 17.2 17.4 16.9 16.5 17.1 18.2 16.8 15.7 16.1

Si assuma che i dati siano distribuiti con legge normale.

1. I dati sono significativi al 5% per decidere di cambiare il metodo di produzione? (N.B.: Derivare le formule)
2. Siamo interessati al più piccolo livello di significatività per cui i dati porterebbero a decidere di cambiare il metodo di produzione. In quale di questi intervalli si colloca tale valore: [0.5%, 1%), [1%, 2.5%), [2.5%, 5%), [5%, 10%)?

Soluzione. Stiamo considerando un campione casuale $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ con $n = 10$ e dove μ e σ sono incognite. Si deve impostare un test di ipotesi sulla varianza:

$$H_0 : \sigma^2 = \sigma_0^2 = 1.2^2 = 1.44, \quad H_1 : \sigma^2 < \sigma_0^2 = 1.44.$$

Infatti ci si sta chiedendo se i dati sono abbastanza significativi da rifiutare l'ipotesi che la deviazione standard sia uguale a 1.2, a favore dell'ipotesi alternativa (deviazione standard strettamente più piccola di 1.2).

Per svolgere il test di ipotesi, scegliamo la regione critica della forma

$$R_C = \{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) : s_n^2 < c\sigma_0^2\}$$

dove $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ è la varianza campionaria calcolata su valori (x_1, \dots, x_n) . La costante c nella definizione della regione critica deve essere definita in termini del livello di significatività α . Questo è la probabilità di commettere un errore del I tipo. Supponiamo allora che l'ipotesi nulla sia vera, cioè $\sigma^2 = \sigma_0^2 = 1.44$. La probabilità di commettere un errore

del I tipo (cioè di rifiutare l'ipotesi nulla a favore dell'ipotesi alternativa) è, usando la varianza campionaria $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ (stimatore corretto della varianza),

$$\alpha = \mathbb{P}(\{(X_1, \dots, X_n) \in R_C\}) = \mathbb{P}(\{S_n^2 < c\sigma_0^2\}) = \mathbb{P}\left(\left\{\frac{(n-1)S_n^2}{\sigma_0^2} < (n-1)c\right\}\right).$$

Ricordiamo che, poiché X_1, \dots, X_n sono distribuite con legge normale e $\sigma^2 = \sigma_0^2$ per l'ipotesi nulla, si ha che $\Xi_{n-1} = \frac{(n-1)S_n^2}{\sigma_0^2}$ è distribuita con una legge chi-quadro con $n-1 = 9$ gradi di libertà. Allora

$$\alpha = \mathbb{P}(\{\Xi_{n-1} < (n-1)c\}) = 1 - \mathbb{P}(\{\Xi_{n-1} \geq (n-1)c\}) \implies \mathbb{P}(\{\Xi_{n-1} \geq (n-1)c\}) = 1 - \alpha.$$

Chiamiamo $\chi_{n-1, 1-\alpha}$ il quantile della chi-quadro che verifica

$$\mathbb{P}(\{\Xi_{n-1} \geq \chi_{n-1, 1-\alpha}\}) = 1 - \alpha.$$

In questo modo, scegliendo $c = \frac{\chi_{n-1, 1-\alpha}}{n-1}$, otteniamo la condizione imposta all'inizio.

Possiamo ora prendere una decisione sul test di ipotesi calcolando i valori sui dati. Consultando le tavole della chi-quadro con $\alpha = 5\%$ otteniamo che

$$\chi_{n-1, 1-\alpha} = \chi_{9, 0.95} \simeq 3.325.$$

Quindi

$$c\sigma_0^2 = \frac{\chi_{n-1, 1-\alpha}}{n-1}\sigma_0^2 = \frac{3.325}{9}1.44 \simeq 0.53.$$

Calcoliamo la media e la varianza sul campione:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10}(16.8 + 17.2 + 17.4 + 16.9 + 16.5 + 17.1 + 18.2 + 16.8 + 15.7 + 16.1) = 16.87,$$

$$\begin{aligned} s_n^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \\ &= \frac{1}{9}(16.8^2 + 17.2^2 + 17.4^2 + 16.9^2 + 16.5^2 + 17.1^2 + 18.2^2 + 16.8^2 + 15.7^2 + 16.1^2 - 10 \cdot 16.87^2) \\ &\simeq 0.48 \end{aligned}$$

Poiché $0.48 < 0.53$, rifiutiamo l'ipotesi nulla a favore dell'ipotesi alternativa.

2. Ci basta calcolare come cambia la soglia della regione critica a seconda di α :

$$\begin{aligned} \alpha = 5\% &\implies c\sigma_0^2 = \frac{\chi_{9, 0.95}}{9}1.44 \simeq \frac{3.325}{9}1.44 \simeq 0.53, \\ \alpha = 2.5\% &\implies c\sigma_0^2 = \frac{\chi_{9, 0.975}}{9}1.44 \simeq \frac{2.700}{9}1.44 \simeq 0.432. \end{aligned}$$

Possiamo terminare qui: con il 2.5% di significatività l'ipotesi nulla non può essere rifiutata. Quindi il più piccolo livello di significatività che porta a rifiutare l'ipotesi nulla è nell'intervallo $[2.5\%, 5\%)$.