Cognome : **Nome :** **Matricola:**

Modulo: Laboratorio di Metodi Matematici e Statistici, M-Z

Appello del 23 Giugno 2017

Esercizio 1.
Per uno studio su una patologia cardio–vascolare viene misurato l'indice di massa corporea (in kg/m^2) in un campione casuale di 15 individui, ottenendo i seguenti valori:

31.6, 34.9, 43.1, 28.6, 49.2, 31.1, 27.5, 24.5, 19.6, 34.6, 42.7, 41.5, 38.6, 21.2, 28.0

a) Calcolare la media e la varianza campionarie dell'indice di massa corporea. Con riferimento al grafico della Figura 1, indicare cosa rappresenta e quali indicazioni sui dati se ne possono trarre.
b) Dopo aver determinato i quartili, disegnare il *box–plot*. Sono presenti degli *outlier*? Giustificare la risposta.
c) I dati del campione sono confrontati con dati raccolti in uno studio europeo, in cui la media è risultata pari a 28.7 kg/m^2 con una deviazione standard di 7.462 kg/m^2. In quale dei due campioni la variabilità dell'indice è maggiore? Giustificare la risposta.

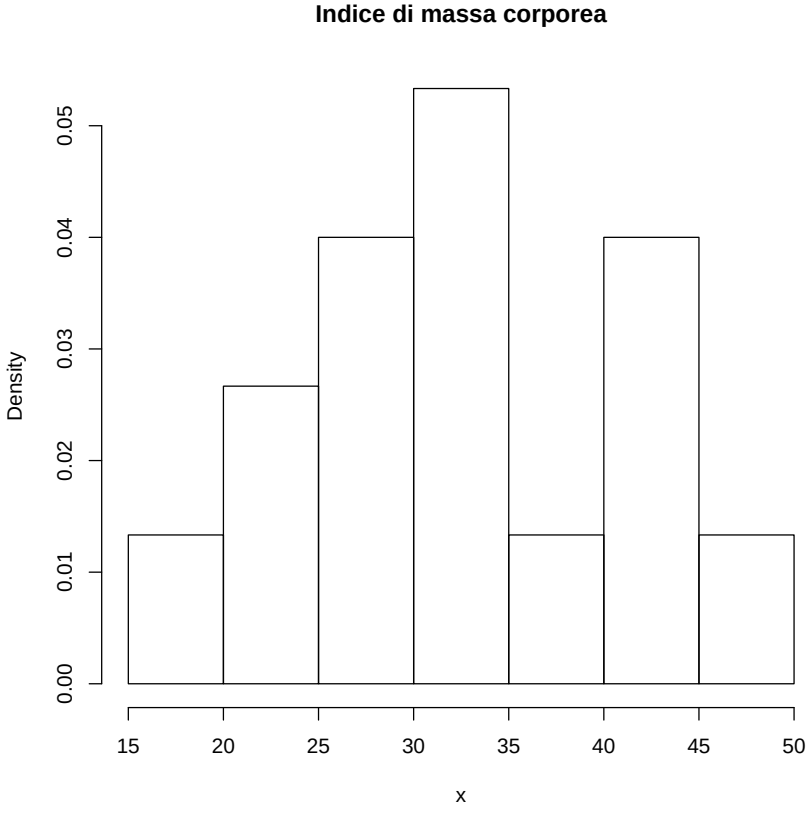

Figura 1:

Soluzione.
a) Si ha $\bar{x} = 33.113$ e $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 74.240$. Il grafico in Figura 1 rappresenta l'istogramma dei dati, cioè la rappresentazione più opportuna della distribuzione di frequenza di dati da una variabile quantitativa continua. In ascissa sono indicate le classi di valori, in questo caso tutte di uguale ampiezza, mentre l'altezza delle colonne è definita, in questo grafico, dalla densità, cioè dal rapporto tra la frequenza relativa dei dati nella classe e l'ampiezza della classe stessa. Trattandosi di classi di uguale ampiezza, l'altezza delle colonne avrebbe potuto essere data, correttamente, anche dalla frequenza relativa (o assoluta) dei dati nella classe. L'istogramma suggerisce una certa asimmetria nella frequenza

dei dati, con una indicazione di bimodalità (classe 30–35 e classe 40–45).
b) Si riordinino i dati in senso crescente:

$$19.6,\ 21.2,\ 24.5,\ 27.5,\ 28.0,\ 28.6,\ 31.1,\ 31.6,\ 34.6,\ 34.9,\ 38.6,\ 41.5,\ 42.7,\ 43.1,\ 49.2.$$

La mediana è data dal valore nella posizione $\frac{n+1}{2} = 8$: 31.6. Il primo quartile è il valore nella posizione $\frac{n+1}{4} = 4$: 27.5, mentre il terzo quartile è il valore nella posizione $3 \times \frac{n+1}{4} = 12$: 41.5. La differenza interquartile (IQR) vale, pertanto, $IQR = 41.5 - 27.5 = 14$. La lunghezza massima dei baffi è $1.5 \times 14 = 21$. Siccome $41.5 + 21 = 61.5 > 49.2$ e $27.5 - 21 = 6.5 < 19.6$, nel *box–plot* non saranno presenti valori anomali (*outlier*). Il *box–plot* è rappresentato in Figura 2.

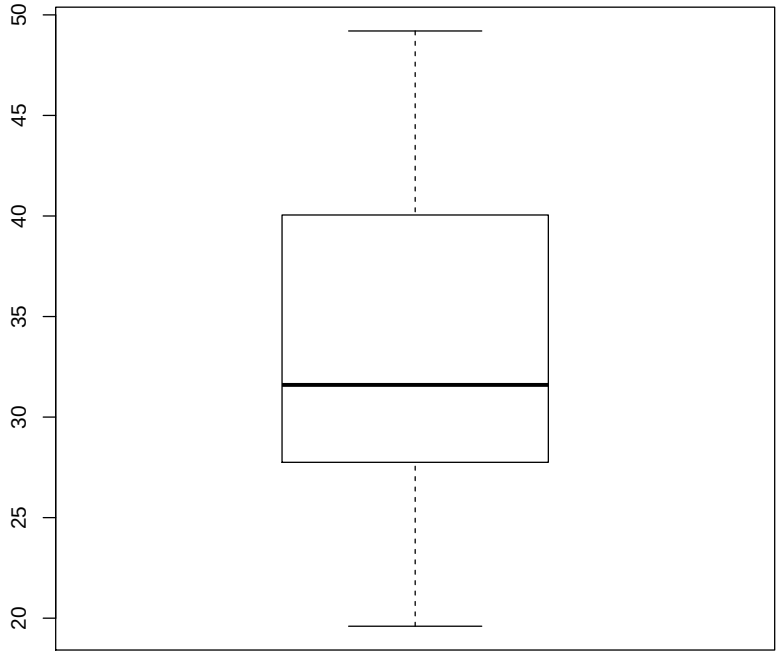

Figura 2: *Box–plot* relativo all'esercizio 1b.

c) Per confrontare la variabilità dell'indice di massa corporea nei due campioni si usa il coefficiente di variazione: $\sqrt{s^2}/\bar{x}$. Da a) si ricava che il CV per il campione dell'Esercizio vale $CV_1 = \frac{\sqrt{74.24}}{33.113} = 0.26$ mentre per il campione europeo vale $CV_E = \frac{7.462}{28.7} = 0.26$. Siccome $CV_1 = CV_E$, nei due campioni si ha la stessa variabilità.

Esercizio 2
La società di trasporto passeggeri ABC dichiara che solo il 10% degli abbonati non trova posto a sedere.
a) Calcolare la probabilità che in un campione casuale di 10 passeggeri della ABC più di uno non trovi posto a sedere.
b) Calcolare la probabilità che in un campione casuale di 100 passeggeri della ABC più di 10 non trovino posto a sedere.
c) Da un'indagine condotta per conto di una associazione di pendolari della ABC su un campione di 150 abbonati risulta che 18 degli intervistati dichiarano di non trovare posto a sedere. C'è abbastanza evidenza nel campione per dire che la ABC ha mentito?
d) Da un'indagine analoga condotta su un campione di 230 pendolari della società concorrente XYZ risulta che 35 di loro dichiarano di non trovare posto a sedere. Sulla base dei due campioni è possibile

stabilire quale delle due aziende fornisce maggiori garanzie ai rispettivi abbonati di trovare posto a sedere? Giustificare la risposta.

Soluzione.
a) Sia X la variabile casuale che indica quanti degli abbonati di ABC nel campione non trovano posto a sedere. Allora, sulla base dell'affermazione dell'azienda, $X \sim Binom(10, 0.10)$ e la probabilità richiesta si calcola agevolmente passando al complementare di "più di uno":
$P(X > 1) = 1 - P(X = 0) - P(X = 1) = 1 - \binom{10}{0}0.10^0(1 - 0.10)^{10} - \binom{10}{1}0.10^1(1 - 0.10)^9 =$
$= 1 - 0.349 - 0.387 = 0.264$
b) Con lo stesso significato di cui al punto a), ora $X \sim Binom(100, 0.10)$, e la probabilità richiesta, sebbene calcolabile (con pazienza) come $P(X > 10) = 1 - P(X = 0) - P(X = 1) - \cdots - P(X = 10)$ si può approssimare rapidamente ricorrendo all'approssimazione della binomiale con la densità normale di media $np = 100 \times 0.10 = 10$ e varianza $np(1 - p) = 10 \times 0.10 \times 0.90 = 9$, e quindi, osservando che 10 è proprio la media della variabile normale, la probabilità richiesta vale approssimativamente 0.5. (Se qualcuno volesse usare la correzione di continuità, non discussa a lezione, otterrebbe $P(X > 10) = P(X > 10.5) = P(\frac{X-10}{3} > \frac{10.5-10}{3}) \approx P(Z > 0.17) = 1 - \Phi(0.17) = 1 - 0.5675 = 0.43$, ove Z indica come al solito la gaussiana standard, che è un'approssimazione migliore del valore esatto 0.417).
c) Possiamo rispondere alla domanda sottoponendo a verifica l'ipotesi nulla $H_0 : p = 0.10$ contro l'alternativa $H_1 : p > 0.10$ (*commento: questa è l'alternativa di maggiore interesse per l'associazione di pendolari*). La proporzione campionaria di abbonati che non trovano posto a sedere è pari a $\hat{p} = \frac{18}{150} = 0.12$, superiore alla percentuale dichiarata da ABC, e la statistica test vale:

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.12 - 0.10}{\sqrt{\frac{0.10 \times 0.90}{150}}} = 0.82$$

corrispondente ad un $p - valore : P(Z > 0.82) = 1 - \Phi(0.82) = 1 - 0.791 = 0.209 > 0.05$ e pertanto non c'è abbastanza evidenza nel campione per sostenere che ABC abbia sottostimato il vero numero di abbonati che non trovano posto a sedere.
d) Per rispondere al quesito possiamo confrontare la stima campionaria della percentuale di abbonati che non trova posto a sedere in ciascuna azienda. Se le due proporzioni risulteranno statisticamente diverse, allora preferiremo la compagnia cui corrisponde la probabilità di non trovare posto a sedere minore. Sottoporremo, quindi, a verifica l'ipotesi nulla $H_0 : p_{ABC} = p_{XYZ}$ contro l'alternativa $H_1 : p_{ABC} \neq p_{XYZ}$. Si ha che $\hat{p}_{XYZ} = \frac{35}{230} = 0.152$ mentre dal punto c) si ha $\hat{p}_{ABC} = 0.12$. Inoltre, la stima della proporzione totale di abbonati che non riescono a trovare posto a sedere è data da: $p_0 = \frac{18+35}{150+230} = 0.139$. La statistica test vale:

$$\frac{|\hat{p}_{ABC} - \hat{p}_{XYZ}|}{\sqrt{p_0(1 - p_0)\left(\frac{1}{150} + \frac{1}{230}\right)}} = \frac{|0.12 - 0.152|}{\sqrt{0.139(1 - 0.139)\left(\frac{1}{150} + \frac{1}{230}\right)}} = 0.88$$

corrispondente al $p - valore = 2 \times P(Z > 0.88) = 2 \times \Phi(-0.88) = 2 \times 0.1894 = 0.3788 > 0.05$ e pertanto, confrontando i risultati dei due campioni indipendenti, non si può rifiutare l'ipotesi nulla che entrambi i servizi abbiano la stessa percentuale di abbonati che non riescono a sedersi.
Una soluzione alternativa accettata nel contesto dell'esame consiste nello stabilire le ipotesi dopo aver guardato alle proporzioni campionarie e determinare la significatività della disuguaglianza tra loro osservata. In questo caso, viene accettata come corretta la verifica dell'ipotesi nulla $H_0 : p_{ABC} = p_{XYZ}$ contro l'alternativa $H_1 : p_{ABC} < p_{XYZ}$, che porta allo stesso esito.

Esercizio 3
Per un certo studio su una patologia respiratoria sono stati rilevati $n = 50$ dati, relativi a due indici che possono indicare la presenza della patologia, AHI e ODI, ottenendo i seguenti risultati:

AHI: $\sum_{i=1}^n x_i = 1338.7, \qquad \sum_{i=1}^n x_i^2 = 68362.09$

ODI: $\sum_{i=1}^n y_i = 1381.0, \qquad \sum_{i=1}^n y_i^2 = 72110.48$

$\sum_{i=1}^n x_i y_i = 69408.81$

1) I due fenomeni sono correlati? Come e quanto?
2) La relazione tra i due indici è stata studiata con un modello di regressione lineare, ottenendo la seguente tabella di risultati:

Variabile	Coefficiente	Dev. standard	Statistica t	$p-value$
Intercetta	0.9167	1.1908	0.77	0.445
AHI	0.9974	0.0322	30.97	0

Quale dei due indici è stato utilizzato come variabile indipendente e quale come variabile dipendente? Qual è la retta di regressione stimata?
3) Quali sono i parametri della regressione che risultano significativi? E a quale livello di significatività?
4) Fornite una misura della bontà di adattamento ai dati del modello lineare stimato e una stima della varianza dell'errore, formulando le corrette ipotesi sulla distribuzione dell'errore.
5) Alla luce dei risultati ottenuti, ritenete che si debbano utilizzare entrambi gli indici nello studio della patologia?

Soluzione.
Dai dati si ha che $\bar{x} = \frac{1338.7}{50} = 26.774, \bar{y} = 27.62$ e quindi $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = 648.678$ e dunque le due variabili sono correlate positivamente. Inoltre: $\sigma_x^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = 650.395$ e $\sigma_y^2 = 679.345$ e quindi il coefficiente di correlazione lineare vale $\rho = \frac{648.678}{\sqrt{650.395 \times 679.345}} = 0.976$ e pertanto c'è una forte correlazione lineare, positiva.
b) La variabile indipendente è AHI, mentre ODI è la dipendente. La retta di regressione stimata è $y_i = 0.9167 + 0.9974 x_i + \epsilon_i$.
3) Solo la pendenza (coefficiente di AHI) risulta significativamente diversa da 0 ($p - valore = 0 < 0.05$) mentre l'intercetta non è significativamente diversa da 0 ($p - valore = 0.445 > 0.05$). La pendenza risulta significativamente diversa da 0 a tutti i livelli di significatività.
4) Una misura di adattamento è data dall'indice R^2 che si può calcolare come $R^2 = \rho^2 = 0.976^2 = 0.952$, indicatore di un adattamento molto buono. La varianza σ^2 comune agli errori ϵ_i, variabili aleatorie che supponiamo indipendenti, gaussiane di media 0, è stimata da $s^2 = \frac{n}{n-2}(1 - R^2)\sigma_y^2 = 33.967$
5) Siccome la regressione lineare è fortemente significativa, si può usare AHI per predire ODI, nel campione, e quindi non è necessario usare entrambi gli indici per lo studio della patologia.