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

Modulo: Laboratorio di Metodi Matematici e Statistici

Appello del 20 Gennaio 2017

Esercizio 1

Si considerino i seguenti dati, che si riferiscono al numero di giorni di malattia chiesti dai 13 dipendenti di una certa azienda nel 2016:

$$2,\ 0,\ 1,\ 1,\ 0,\ 2,\ 3,\ 0,\ 1,\ 10,\ 3,\ 2,\ 2$$

1) Di che tipo di variabile si tratta? Calcolare la moda, la media e la varianza dei dati.
2) Rappresentare graficamente la distribuzione di frequenza dei dati.
3) Calcolare i quartili e rappresentare il box-plot. Ci sono valori grandi in modo anomalo? Giustificare la risposta.

Esercizio 2

In un campione casuale di 150 soggetti affetti da morbillo è emerso che 22 presentano tra i sintomi la febbre alta.
1) Fornire una stima della proporzione di malati di morbillo che possono presentare la febbre alta come sintomo.
2) Si estraggano a caso, con reimmissione, 10 soggetti dal campione: calcolare la probabilità che meno di 2 abbiano la febbre alta.
3) Determinare l'intervallo di confidenza di livello 99% per la proporzione di malati di morbillo che possono presentare come sintomo la febbre alta.
4) C'è abbastanza evidenza nel campione per dimostrare ad un livello di significatività $\alpha = 0.05$ che la proporzione di malati di morbillo che possono presentare tra i sintomi la febbre alta è superiore al 12%?
5) Senza ulteriori conti, quale sarebbe la risposta al punto 4) se fosse $\alpha = 0.01$? Giustificare la risposta.

Esercizio 3

Per uno studio sull'inquinamento dell'aria a Milano sono stati rilevati, per 30 giorni, il valore medio orario di PM10, y_i in $\mu g/m^3$, e la temperatura massima giornaliera, x_i in $°C$, ottenendo i seguenti risultati:

$$\begin{array}{lll} \text{Temperatura:} & \sum_{i=1}^{30} x_i = 255.9, & \sum_{i=1}^{30} x_i^2 = 2346.8 \\ \text{PM10:} & \sum_{i=1}^{30} y_i = 1146.0, & \sum_{i=1}^{30} y_i^2 = 53763.2 \\ & \sum_{i=1}^{30} x_i y_i = 10867.8 & \end{array}$$

1) Calcolare media e varianza dei dati relativi a ciascun fenomeno: quale dei due presenta la maggiore variabilità?
2) I due fenomeni sono correlati? Come e quanto?
3) Si discuta l'opportunità di adattare ai dati un modello di regressione lineare $Y_i = a + b\,x_i + \epsilon_i$ e, in ogni caso, si forniscano le stime dei parametri incogniti del modello.
4) Si valuti la significatività della regressione al livello del 5% e si determini l'intervallo di confidenza del 95% per la pendenza.

Soluzioni

Esercizio 1

La variabile "numero di giorni di malattia" è una variabile discreta, positiva. Sarà utile in tutto l'esercizio aver riordinato i 13 dati in ordine crescente: 0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 3, 3, 10.

1. I dati presentano come unica moda il valore 2. Per la media e la varianza si ha: $\bar{x} = (3\times 0+3\times 1+4\times 2+2\times 3+10)/13 = 2.077$ e $\sigma^2 = (3\times 0^2+3\times 1^2+4\times 2^2+2\times 3^2+10^2)/13 - 2.077^2 = 6.225$.

2) La rappresentazione corretta è quella tramite un diagramma ad aste (o bastoncini) come in Figura 1, con l'altezza delle aste definita dalle frequenze relative (o assolute).

3) Con riferimento ai dati riordinati, i quartili sono i valori nelle posizioni: $(n+1)/4 = 3.5$, $(n+1)/2 = 7$, $3(n+1)/4 = 10.5$ e, pertanto, valgono: $Q_1 = (0+1)/2 = 0.5$, $Q_2 = 2$ e $Q_3 = (2+3)/2 = 2.5$ rispettivamente. La differenza interquartile (IQR) vale $Q_3 - Q_1 = 2.5 - 0.5 = 2$ e la lunghezza massima dei baffi del box-plot vale, quindi, $1.5 \times IQR = 1.5 \times 2 = 3$. Il baffo inferiore si arresterà, dunque, a 0 (il dato minimo), mentre il baffo superiore si arresterà a 3 (il dato più vicino a 5.5), mentre il dato pari a 10 costituisce un *outlier*, o dato grande in modo anomalo, essendo esterno al baffo (in particolare, si tratta di un *outlier* estremo). Il boxplot è rappresentato in Figura 2.

Esercizio 2

1) La proporzione campionaria di malati di morbillo con la febbre alta tra i sintomi è: $\widehat{p} = 22/150 = 0.147$

2) Trattandosi di estrazioni *con reimmissione* sono indipendenti e, pertanto, la variabile casuale X che conta, tra i 10 estratti, il numero di malati di morbillo con la febbre alta come sintomo è una Binomiale di parametri $n = 10$ e $p = \widehat{p} = 0.147$. La probabilità cercata è $P(X < 2) = P(X = 0) + P(X = 1) = (1-0.147)^{10} + 10 \times 0.147^1 \times (1-0.147)^9 = 0.555$.

3) Essendo soddisfatte le condizioni per l'utilizzo dell'intervallo di confidenza asintotico per la proporzione:

$$\left(\widehat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}; \widehat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$$

per $\alpha = 0.01$ si ha $z_{\frac{\alpha}{2}} = 2.5758$, e quindi

$$(0.147 - 2.5758 \cdot 0.029,\ 0.147 + 2.5758 \cdot 0.029)$$

$$(0.073,\ 0.221)$$

4) Si vuole sottoporre a verifica l'ipotesi nulla $H_0 : p = 0.12$ contro l'alternativa $H_1 : p > 0.12$. Osserviamo che $\widehat{p} = 0.147 > 0.12$. La statistica test è data da

$$\frac{\widehat{p} - 0.12}{\sqrt{\frac{0.12\cdot(1-0.12)}{150}}} = 1.018$$

ed essendo inferiore al valore critico $z_{0.05} = 1.6448$, non si può rifiutare l'ipotesi nulla al livello richiesto, cioè **non** c'è abbastanza evidenza per dimostrare che la percentuale di malati di morbillo che possono manifestare la febbre alta tra i sintomi è superiore al 12%.

5) Se non si può rifiutare l'ipotesi al livello del 5% non la si rifiuta nemmeno al livello dell'1%, perchè riducendo α si aumenta il valore critico del test.

Esercizio 3

1) Con riferimento alla tabella del testo, si ha:

Temperatura: $\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{255.9}{30} = 8.53$, $\sigma_x^2 = \frac{1}{30}\sum_{i=1}^{30} x_i^2 - \bar{x}^2 = \frac{2346.8}{30} - 8.53^2 = 5.466$

PM10: $\bar{y} = \frac{1}{30}\sum_{i=1}^{30} y_i = \frac{1146.0}{30} = 38.2$, $\sigma_y^2 = \frac{1}{30}\sum_{i=1}^{30} y_i^2 - \bar{y}^2 = \frac{53763.2}{30} - 38.2^2 = 332.867$

La variabilità dei dati si confronta tramite il coefficiente di variazione: $\mathrm{CV}_x = \frac{\sigma_x}{\bar{x}} = \frac{\sqrt{5.466}}{8.53} = 0.274$ e $\mathrm{CV}_y = \frac{\sigma_y}{\bar{y}} = \frac{\sqrt{332.867}}{38.2} = 0.478$ pertanto il PM10 medio orario ha una variabilità superiore alla temperatura massima giornaliera.

2) Valutiamo tramite il coeff. di correlazione lineare ρ_{xy} l'eventuale presenza di una relazione lineare tra il PM10 medio orario e la temperatura massima giornaliera. Siccome

$$\rho_{xy} = \frac{cov_{x,y}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{n^{-1}\sum_i x_i y_i - \bar{x}\bar{y}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{10867.8/30 - 8.53 \times 38.2}{\sqrt{5.466 \times 332.867}} = \frac{36.414}{\sqrt{5.466 \times 332.867}} = 0.854$$

possiamo affermare che tra i dati esiste una correlazione lineare positiva.
3) La bontà dell'adattamento ai dati di un modello lineare è misurata da $R^2 = \rho_{xy}^2 = 0.854^2 = 0.729 > 0.70$ e quindi sembra ragionevole modellizzare la relazione tra i dati con il modello di regressione lineare semplice. Le stime dei parametri sono:

$$\hat{b} = \frac{cov_{xy}}{\sigma_x^2} = \frac{36.414}{5.466} = 6.662, \quad \hat{a} = 38.2 - \hat{b} \cdot 8.53 = -18.627,$$

$$\text{ed} \quad s^2 = \frac{n}{n-2}(1 - R^2)\sigma_y^2 = \frac{30}{28}(1 - 0.729) \cdot 332.867 = 96.650.$$

4) L'intervallo di confidenza al 95% per la pendenza è dato da:

$$\hat{b} \mp t(n-2)_{\frac{\alpha}{2}} \sqrt{\frac{s^2}{n\sigma_x^2}}, \text{ quindi: } \left(6.662 \mp 2.0484 \sqrt{\frac{96.650}{30 \cdot 5.466}}\right) = (5.089, 8.235)$$

ed essendo un intervallo che non contiene il valore 0, possiamo rifiutare l'ipotesi $H_0 \ : \ b = 0$ al livello del 5%. Infatti, il valore della statistica test è

$$\frac{\hat{b}}{\sqrt{\frac{s^2}{n\sigma_x^2}}} = \frac{6.662}{\sqrt{\frac{96.650}{30 \cdot 5.466}}} = 8.678$$

e siccome $8.678 > t(n-2)_{\alpha/2} = 2.0484$ la regressione è significativa al livello del 5%.

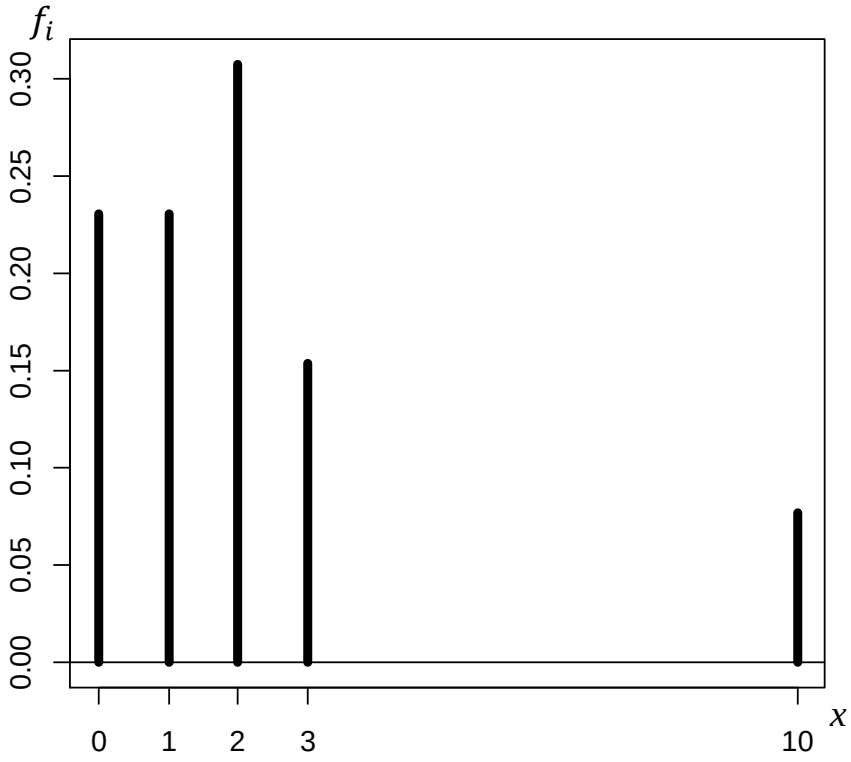

Figura 1 – Esercizio 1. Diagramma ad aste/bastoncini della distribuzione di frequenza (relative) dei dati.

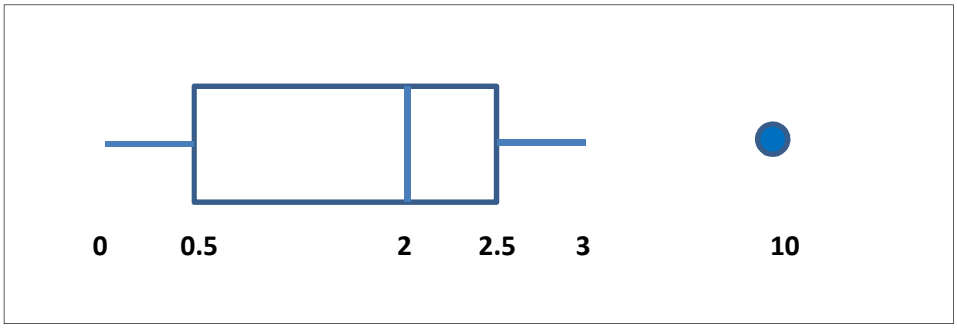

Figura 2 – Esercizio 1. Boxplot.