Esercizi di ricapitolazione sulla Statistica Descrittiva

Esercizio 1. Arrotondare il numero 0.1702534 alla quarta e alla sesta cifra significativa.

Quarta cifra significatica: 0.1703

Sesta cifra significativa: 0.170253

Esercizio 2. Calcolare media, mediana e moda per la sequenza di dati: 1, 5, 7, 2, 1, 8, 3, 1.

Media
Sia X una variabile numerica. Consideriamo n dati: $x_1,\ x_2, ...,\ x_n$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\begin{aligned}\bar{x} &= \frac{1}{8}\sum_{i=1}^{8} x_i \\ &= \frac{1}{8}(1+5+7+2+1+8+3+1) \\ &= \frac{28}{8} \\ &= 3.5\end{aligned}$$

Mediana
Dopo aver ordinato gli n dati in modo crescente si ha:

$$m = \frac{1}{2}\ (x_{n/2} + x_{n/2+1}) \tag{2}$$

se n è pari e

$$m = x_{(n+1)/2} \tag{3}$$

se n è dispari

Dati ordinati:

1 1 1 2 3 5 7 8

$$
\begin{aligned}
m &= \frac{1}{2}\ (x_4 + x_5) \\
&= \frac{1}{2}(2+3) \\
&= \frac{5}{2} \\
&= 2.5
\end{aligned}
$$

Moda
Le mode sono i punti di massimo assoluto della distribuzione di frequenza

La distribuzione è unimodale con moda pari a 1.

Esercizio 3. Calcolare media, mediana e moda per la sequenza di dati: 2, 5, 5, 9, 2, 1, 8, 3, 4.

Media

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{9}(2+5+5+9+2+1+8+3+4) \\
&= \frac{39}{9} \\
&= 4.\bar{3}
\end{aligned}
$$

Mediana

Dati ordinati:

1 2 2 3 4 5 5 8 9

$$
\begin{aligned}
m &= x_5 \\
&= 4
\end{aligned}
$$

Moda

La distribuzione è bimodale con moda pari a 2 e 5.

Esercizio 4. Considerando i dati: 1, 5, 7, 2, 1, 8, 3, 1, calcolare la somma delle deviazioni dalla media.

La **somma delle deviazioni dalla media** è la somma di tutte le differenze tra i dati e la loro media, ovvero:

$$\sum_{i=1}^{n}(x_i - \bar{x})$$

$$\begin{aligned}\sum_{i=1}^{8}(x_i - \bar{x}) = \\ &= (1-3.5) + (5-3.5) + (7-3.5) + (2-3.5) + (1-3.5) + (8-3.5) + (3-3.5) + (1-3.5) \\ &= -2.5 + 1.5 + 3.5 - 1.5 - 2.5 + 4.5 - 0.5 - 2.5 \\ &= 0\end{aligned}$$

Esercizio 5. Considerando i dati: 1, 5, 7, 2, 1, 8, 3, 1, calcolare lo scarto medio assoluto rispetto alla media.

Lo **scarto medio assoluto rispetto alla media** è la somma dei valori assoluti delle deviazioni dalla media diviso n:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\begin{aligned}\frac{1}{8}\sum_{i=1}^{8}|x_i - \bar{x}| = \\ &= \frac{1}{8}\left(|1-3.5| + |5-3.5| + |7-3.5| + |2-3.5| + |1-3.5| + |8-3.5| + |3-3.5| + |1-3.5|\right) \\ &= \frac{1}{8}\left(2.5 + 1.5 + 3.5 + 1.5 + 2.5 + 4.5 + 0.5 + 2.5\right) \\ &= \frac{19}{8} \\ &= 2.375\end{aligned}$$

Esercizio 6. Considerando i dati: 1, 5, 7, 2, 1, 8, 3, 1, calcolare lo scarto medio assoluto rispetto alla mediana.

Lo **scarto medio assoluto rispetto alla mediana** è la somma dei valori assoluti delle deviazioni dalla media diviso n:

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - m|$$

$$\begin{aligned}
\frac{1}{8} \sum_{i=1}^{8} |x_i - m| = & \\
&= \frac{1}{8} \left(|1 - 2.5| + |5 - 2.5| + |7 - 2.5| + |2 - 2.5| + |1 - 2.5| + |8 - 2.5| + |3 - 2.5| + |1 - 2.5|\right) \\
&= \frac{1}{8} \left(1.5 + 2.5 + 4.5 + 0.5 + 1.5 + 5.5 + 0.5 + 1.5\right) \\
&= \frac{18}{8} \\
&= 2.25
\end{aligned}$$

Esercizio 7. Considerando i dati: 1, 5, 7, 2, 1, 8, 3, 1, calcolare la varianza e la deviazione standard.

La **varianza** è la somma dei quadrati delle deviazioni dalla media diviso n:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4}$$

$$\begin{aligned}
\sigma^2 &= \frac{1}{8} \sum_{i=1}^{8} (x_i - \bar{x})^2 \\
&= \frac{1}{8} (1 - 3.5)^2 + (5 - 3.5)^2 + (7 - 3.5)^2 + (2 - 3.5)^2 + (1 - 3.5)^2 + (8 - 3.5)^2 + (3 - 3.5)^2 + (1 - 3.5)^2 \\
&= \frac{1}{8} (-2.5)^2 + (1.5)^2 + (3.5)^2 + (-1.5)^2 + (-2.5)^2 + (4.5)^2 + (-0.5)^2 + (-2.5)^2 \\
&= \frac{56.0}{8} \\
&= 7.0
\end{aligned}$$

La **deviazione standard (o scarto quadratico medio)** è la radice quadrata della varianza:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

$$\begin{aligned}\sigma &= \sqrt{7} \\ &= 2.64\end{aligned}$$

Esercizio 8. Considerando i dati: 1, 5, 7, 2, 1, 8, 3, 1, calcolare l'indice di asimmetria e identificarela tipologia di asimmetria.

L'indice di skewness (= asimmetria) è un valore che fornisce una misura della mancanza di simmetria in una distribuzione di dati

$$s = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3 \tag{6}$$

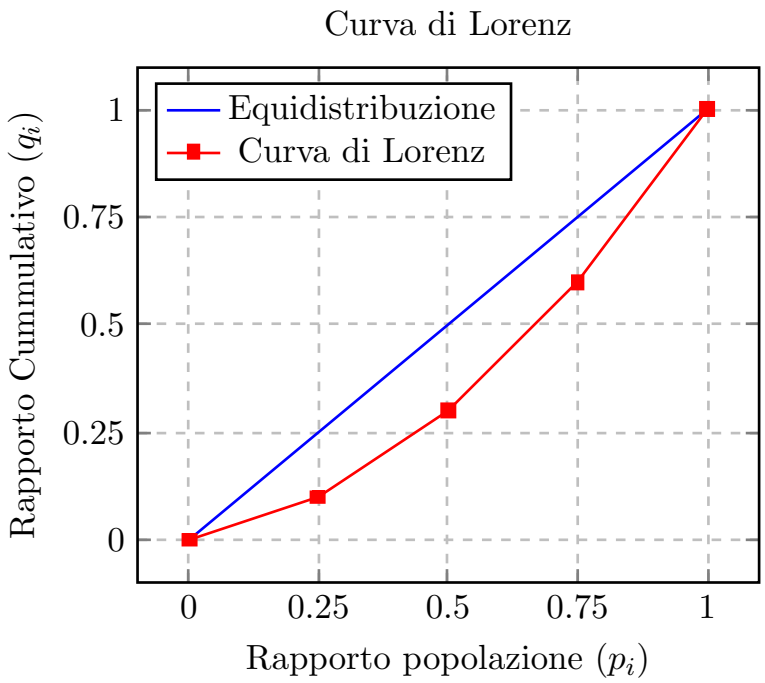
Questo indice misura dove si trova la "coda" della distribuzione:
$s > 0$ la "coda" è verso destra (asimmetrica a destra),
$s < 0$ la "coda" è verso sinistra (asimmetrica a sinistra),
$s = 0$ la distribuzione 'e (abbastanza) simmetrica rispetto alla media.

$$\begin{aligned}s &= \frac{1}{8}\sum_{i=1}^{8}\left(\frac{x_i - \bar{x}s}{\sigma}\right)^3 \\ &= \frac{1}{8}\,\frac{(1-3.5)^3 + (5-3.5)^3 + (7-3.5)^3 + (2-3.5)^3 + (1-3.5)^3 + (8-3.5)^3 + (3-3.5)^3 + (1-3.5)^3}{7\sqrt{7}} \\ &= \frac{1}{8}\,\frac{(-2.5)^3 + (1.5)^3 + (3.5)^3 + (-1.5)^3 + -(2.5)^3 + (4.5)^3 + (-0.5)^3 + (-2.5)^3}{7\sqrt{7}} \\ &= \frac{87.0}{56\sqrt{7}} \\ &= 0.59\end{aligned}$$

Il coefficente di asimmetria é pari a: 0.59, quindi la distribuzione è asimmetrica a destra.

Esercizio 9. Considerando una popolazione di quattro individui con i seguenti redditi: 500, 1000,1500, 2000, disegnare la curva di Lorenz e calcolare l'indice di disuguaglianza di Gini.

Indice (i)	Rapp. Popolazione (p_i)	Redito (r_i)	Redito Cumulativo (c_i)	Rapp. Cumulativo (q_i)
1	0.25	500	500	0.10
2	0.50	1000	1500	0.30
3	0.75	1500	3000	0.60
4	1.00	2000	5000	1.00
Totale	2.50	5000		2.00



Definizione **Indici di Gini**:

$$G = 1 - \frac{\sum_{i=1}^{n} q_i}{\sum_{i=1}^{n} p_i} \tag{7}$$

$$\begin{aligned} G &= 1.0 - \frac{\sum_{i=1}^{4} q_i}{\sum_{i=1}^{4} p_i} \\ &= 1.0 - \frac{2.0}{2.5} \\ &= 0.2 \end{aligned}$$

Esercizio 10. Considerando i dati: 3, 6, 7, 4, 4, 2, 7, 3, 2, 6, 2, 9, 5, 5, 6, si determini il valore (ovvero il percentile) corrispondente al 30% dei dati.

Siano $x_1 \leq x_2 \leq ... \leq x_n$, n dati osservati ordinati e sia $p \in [0,1]$. Il $p-esimo$ **quantile** (o $100p-esimo$ **percentile**) è

$$q_p = x_{\lceil np \rceil} \tag{8}$$

se np non è intero;

$$q_p = \frac{x_{np} + x_{np+1}}{2} \tag{9}$$

se np è intero.

Dati ordinati: 2 2 2 3 3 4 4 5 5 6 6 6 7 7 9

Vogliamo determinare il valore che identifica il 30% dei dati.

$$\begin{aligned} \lceil 0.3 \cdot 15 \rceil &= \lceil 4.5 \rceil \\ &= 5 \end{aligned}$$

$$q_{0.3} = x_5 = 3$$

Esercizio 11. Considerando i dati: 3, 6, 7, 4, 4, 2, 7, 3, 2, 6, 2, 9, 5, 5, 6, calcolare il valore dei quartili. Quanto vale la differenza interquartile?

Il $25-esimo$, $50-esimo$ e $75-esimo$ percentile, vengono indicati con Q_1, Q_2 e Q_3, rispettivamente, e sono detti **primo, secondo e terzo quartile**.

Dati ordinati: 2 2 2 3 3 4 4 5 5 6 6 6 7 7 9

$$\begin{aligned} \lceil 0.25 \cdot 15 \rceil &= \lceil 3.75 \rceil \\ &= 4 \end{aligned}$$

$$\begin{aligned} \lceil 0.50 \cdot 15 \rceil &= \lceil 7.50 \rceil \\ &= 8 \end{aligned}$$

$$\begin{aligned}\lceil 0.75 \cdot 15 \rceil &= \lceil 11.25 \rceil \\ &= 12\end{aligned}$$

$$\begin{aligned}Q_1 &= x_4 = 3 \\ Q_2 &= x_8 = 5 \\ Q_3 &= x_{12} = 6\end{aligned}$$

La **differenza interquartile** si calcola:

$$IQR = Q_3 - Q_1 \tag{10}$$

$$\begin{aligned}IQR &= Q_3 - Q_1 \\ &= 6 - 3 \\ &= 3\end{aligned}$$

Esercizio 12. Considerando i dati: 3, 6, 7, 4, 4, 2, 7, 3, 2, 6, 2, 9, 5, 5, 6, calcolare il valore dei percentili che suddividono i dati in tre parti uguali.

Dati ordinati: 2 2 2 3 3 4 4 5 5 6 6 6 7 7 9

$$\begin{aligned}q_{1/3} &= \frac{x_5 + x_6}{2} = 3.5 \\ q_{2/3} &= \frac{x_{10} + x_{11}}{2} = 6\end{aligned}$$

Esercizio 13. Considerando i dati: 3, 6, 7, 4, 4, 2, 7, 3, 2, 6, 2, 9, 5, 5, 6, disegnare un boxplot, avente come baffi il 10° e il 90° percentile, e dire quali sono gli outliers.

Dati ordinati: 2 2 2 3 3 4 4 5 5 6 6 6 7 7 9

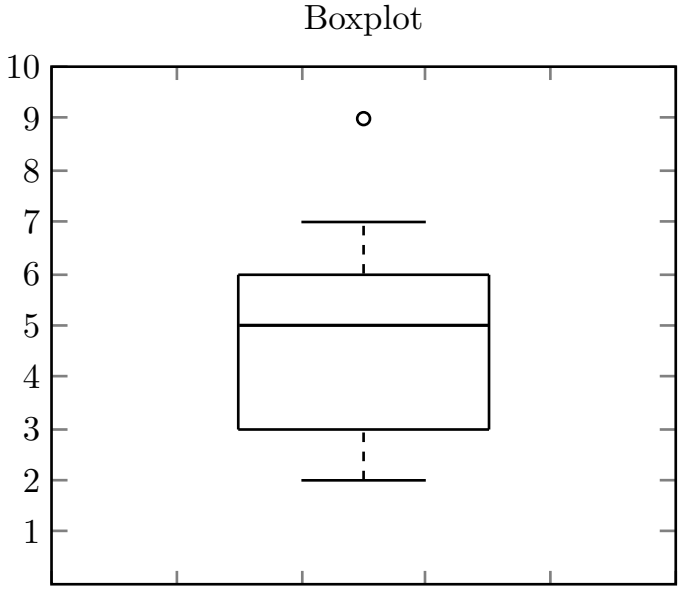
Determiniamo i baffi

$$\begin{aligned}\lceil 0.1 \cdot 15 \rceil &= \lceil 1.5 \rceil \\ &= 2\end{aligned}$$

$$\lceil 0.9 \cdot 15 \rceil = \lceil 13.5 \rceil$$
$$= 14$$

$$q_{0.1} = x_2 = 2$$
$$q_{0.9} = x_{14} = 7$$

Dunque $x_{15} = 19$ è l'unico outlier. Il primo, secondo e terzo quartili sono stati calcolati nell'esercizio precedente.



Esercizio 14. Si considerino due variabili statistiche X e Y su uno stesso campione di 5 unità statistiche. I dati raccolti sono i seguenti: (X) 1, 3, 5, 7, 9; (Y) 2, 4, 4, 5, 10. Calcolare il coefficiente di correlazione, individuando il tipo di correlazione che esiste tra i dati.

Covarianza
Sia $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ una successione di n osservazioni congiunte di due variabili. Si dice covarianza delle due variabili x e y la quantità

$$\sigma_{xy} = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{11}$$

La covarianza puoò anche essere calcolata nel seguente modo:

$$\sigma_{xy} = \frac{1}{n}\sum_{i}^{n} x_i y_i - \bar{x}\bar{y} \tag{12}$$

Cerchiamo di dimostrare eq. (12) partendo da eq. (11) facendo uso di semplici passaggi algebrici.

$$\begin{aligned}
\sigma_{xy} &= \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n}\sum_{i}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) \\
&= \frac{1}{n}\sum_{i}^{n} x_i y_i - \frac{1}{n}\sum_{i}^{n} x_i\bar{y} - \frac{1}{n}\sum_{i}^{n}\bar{x}y_i + \frac{1}{n}\sum_{i}^{n}\bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i}^{n} x_i y_i - \bar{y}\frac{1}{n}\sum_{i}^{n} x_i - \bar{x}\frac{1}{n}\sum_{i}^{n} y_i + \bar{x}\bar{y}\frac{1}{n}\sum_{i}^{n} 1 \\
&= \frac{1}{n}\sum_{i}^{n} x_i y_i - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i}^{n} x_i y_i - \bar{x}\bar{y}
\end{aligned}$$

Si dice **coefficiente di correlazione** delle due variabili x e y la quantità

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{13}$$

i	x_i	y_i	x_iy_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	1	2	2	16	9
2	3	4	12	4	1
3	5	4	20	0	1
4	7	5	35	4	0
5	9	10	90	16	25
Somma	25	25	159	40	36

$\bar{x}$	$\bar{y}$	σ_{xy}	σ_x	σ_y
5	5	6.80	2.83	2.68

$$\begin{aligned}\sigma_{xy} &= \frac{1}{5}\sum_i^5 x_i y_i - \bar{x}\bar{y} \\ &= \frac{159}{5} - 5 \cdot 5 \\ &= 31.8 - 25.0 \\ &= 6.8\end{aligned}$$

La media è calcolata come nell'esercizio 2 e la deviazione standard come nell'esercizio 7, sono riportati i passaggi rilevanti per esercizio

$$\begin{aligned}\rho_{xy} &= \frac{6.80}{2.83 \cdot 2.68} \\ &= 0.89\end{aligned}$$

Le variabile x e y sono fortemente direttamente correlate poicheè $\rho_{xy} = 0.89$.

Esercizio 15. Si considerino le variabili statistiche X e Y dell'esercizio precedente e si determini la retta di regressione e il valore previsto per $x = 15$.

La **retta di regressione** corrispondente alle osservazioni $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, ha equazione

$$y = ax + b \tag{14}$$

con

$$a = \frac{\sigma_{xy}}{{\sigma_x}^2} \tag{15}$$

$$b = \bar{y} - \bar{x}\frac{\sigma_{xy}}{{\sigma_x}^2} \tag{16}$$

Chiamiamo **valori stimati**, i numeri

$$\hat{y}_i = ax_i + b \tag{17}$$

Nella seguente tabella riportiamo gli indici calcolati nell'esercizio precedente e che ci servirano a calcolare i coefficenti a e b.

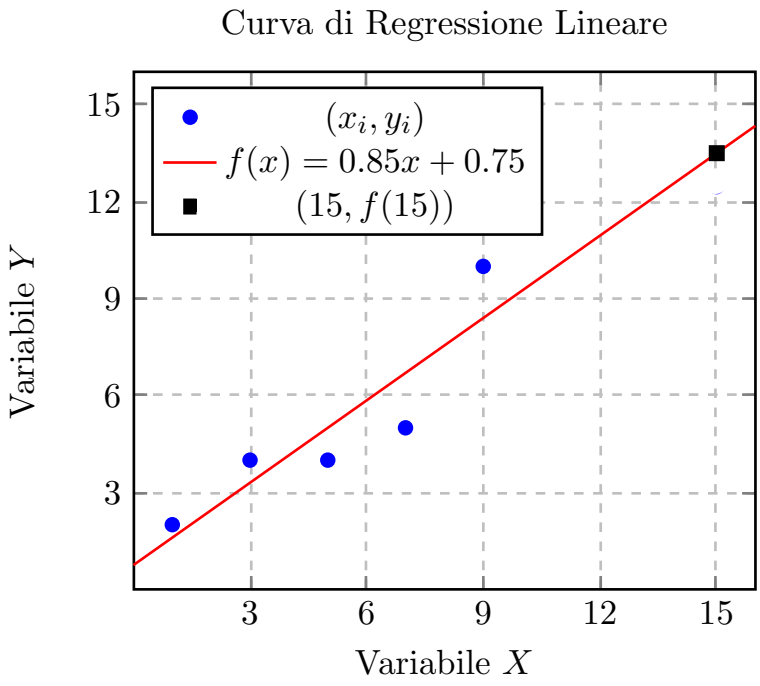
$\bar{x}$	$\bar{y}$	σ_{xy}	${\sigma_x}^2$
5	5	6.8	8

$$\begin{aligned} a &= \frac{6.8}{8} \\ &= 0.85 \end{aligned}$$

$$\begin{aligned} b &= 5 - 5 \cdot \frac{6.8}{8} \\ &= 0.75 \end{aligned}$$

$$y = 0.85x + 0.75$$

$$\begin{aligned} \hat{y} &= 0.85 \cdot 15 + 0.75 \\ &= 13.50 \end{aligned}$$


Curva di Regressione Lineare
(x_i, y_i)
$f(x) = 0.85x + 0.75$
$(15, f(15))$
Variabile Y
Variabile X
15
12
9
6
3
3
6
9
12
15