STATISTICHE CAMPIONARIE

Quando i dati sono molti e *illeggibili* nella forma grezza, si rende necessario introdurre quantità numeriche che possano essere usate per sintetizzarli. Queste misure riassuntive dei dati si chiamano **statistiche campionarie (indici)**.

Una statistica è una quantità numerica il cui valore è determinato dai dati.

Per esempio la **moda** (valore con frequenza piu alta) è una statistica così come la **media**, la **mediana**, la **varianza** e la **deviazione standard**.

Gli indici (statistiche) servono per misurare quantitativamente caratteristiche che possono essere osservate qualitativamente sui grafici.

• **indici di posizione** che misurano la tendenza centrale dell'insieme dei dati: (**media, moda, mediana**)

• **indici di dispersione**: che danno una misura di quanto i valori siano lontani dal centro: (**varianza, deviazione standard**)

MODA

Si chiama **moda** campionaria il valore che si verifica con maggiore frequenza.

Unimodale

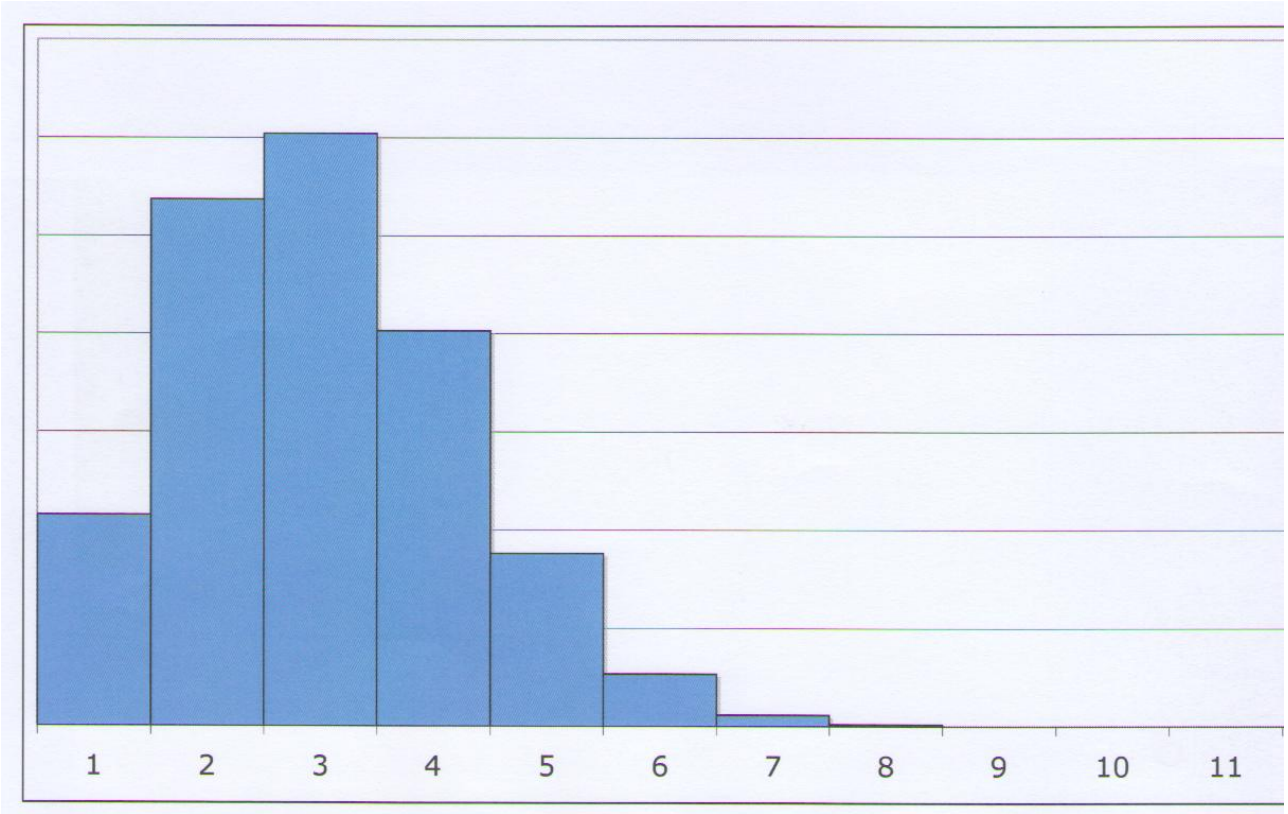
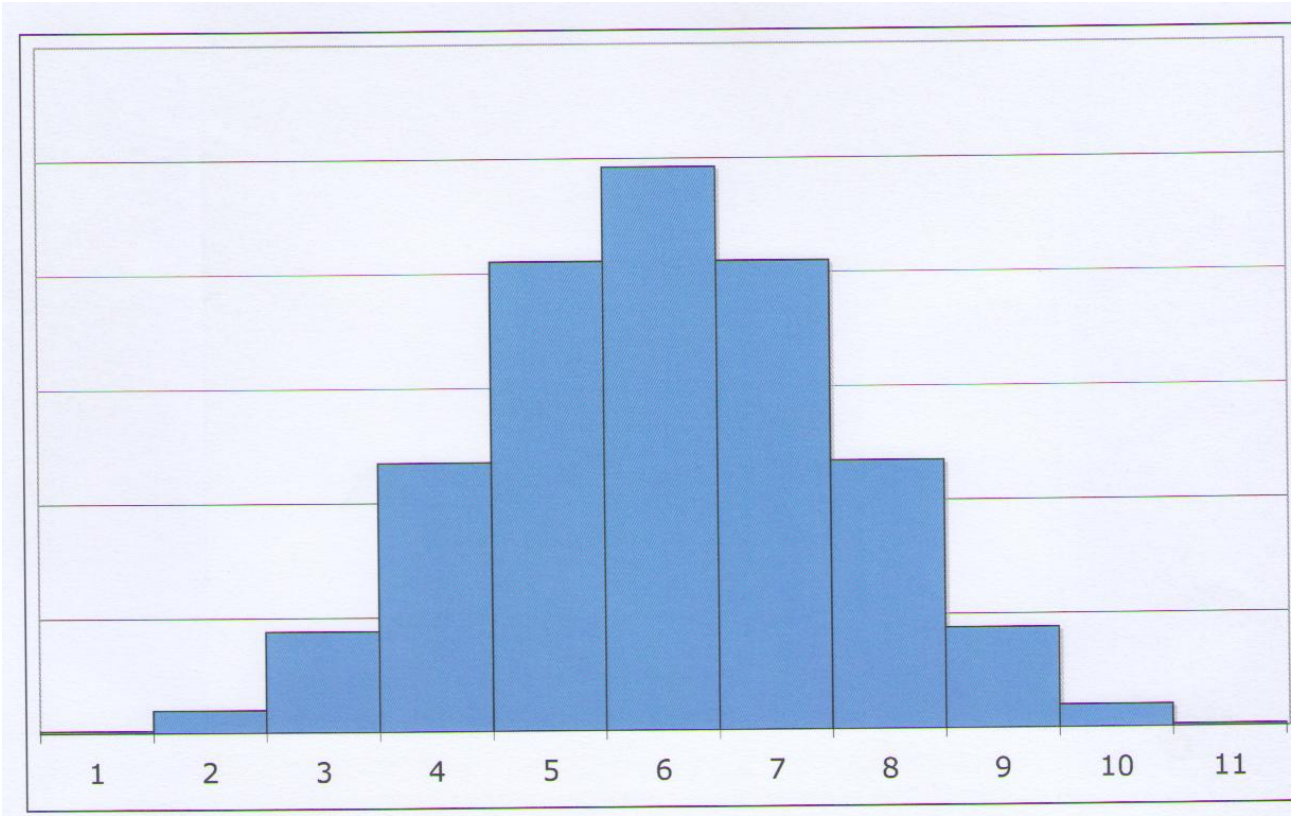
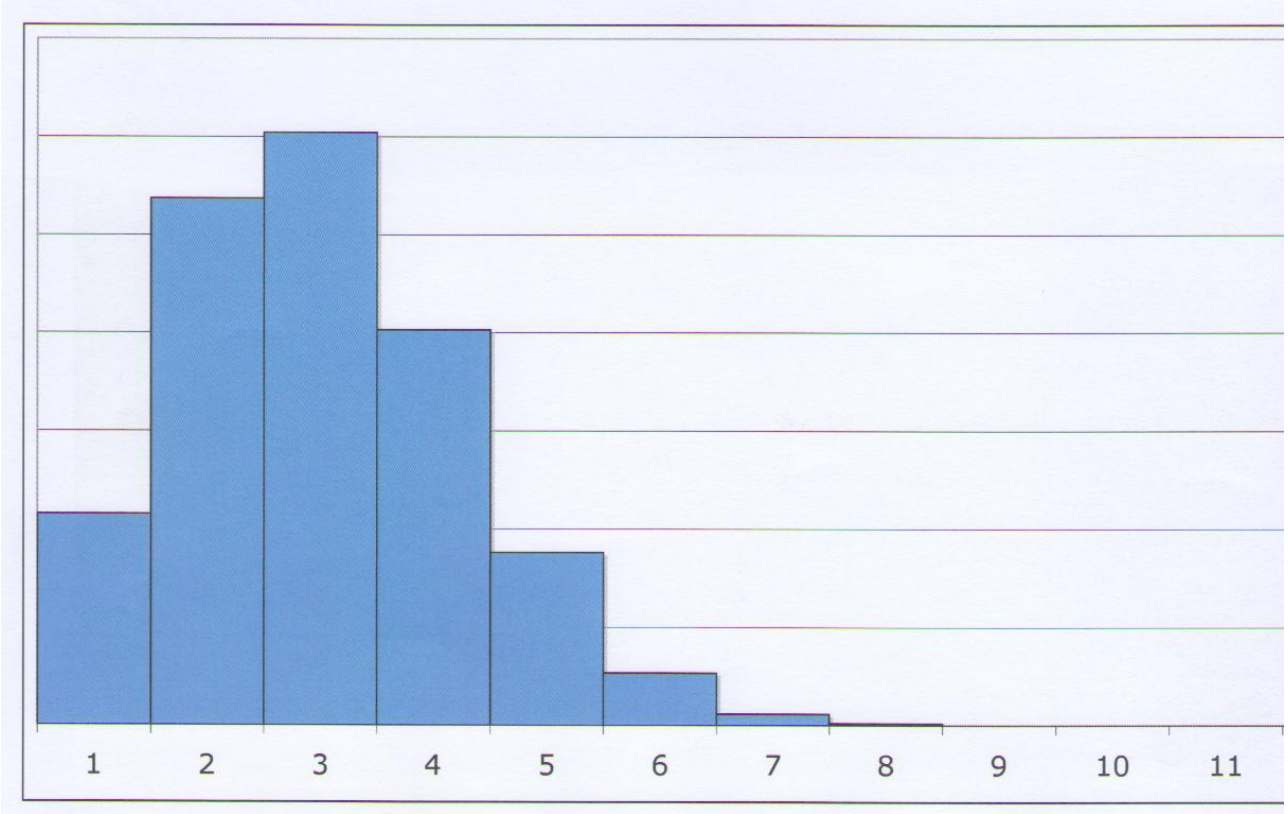


Grafico unimodale **simmetrico**: la simmetria è centrata intorno alla moda



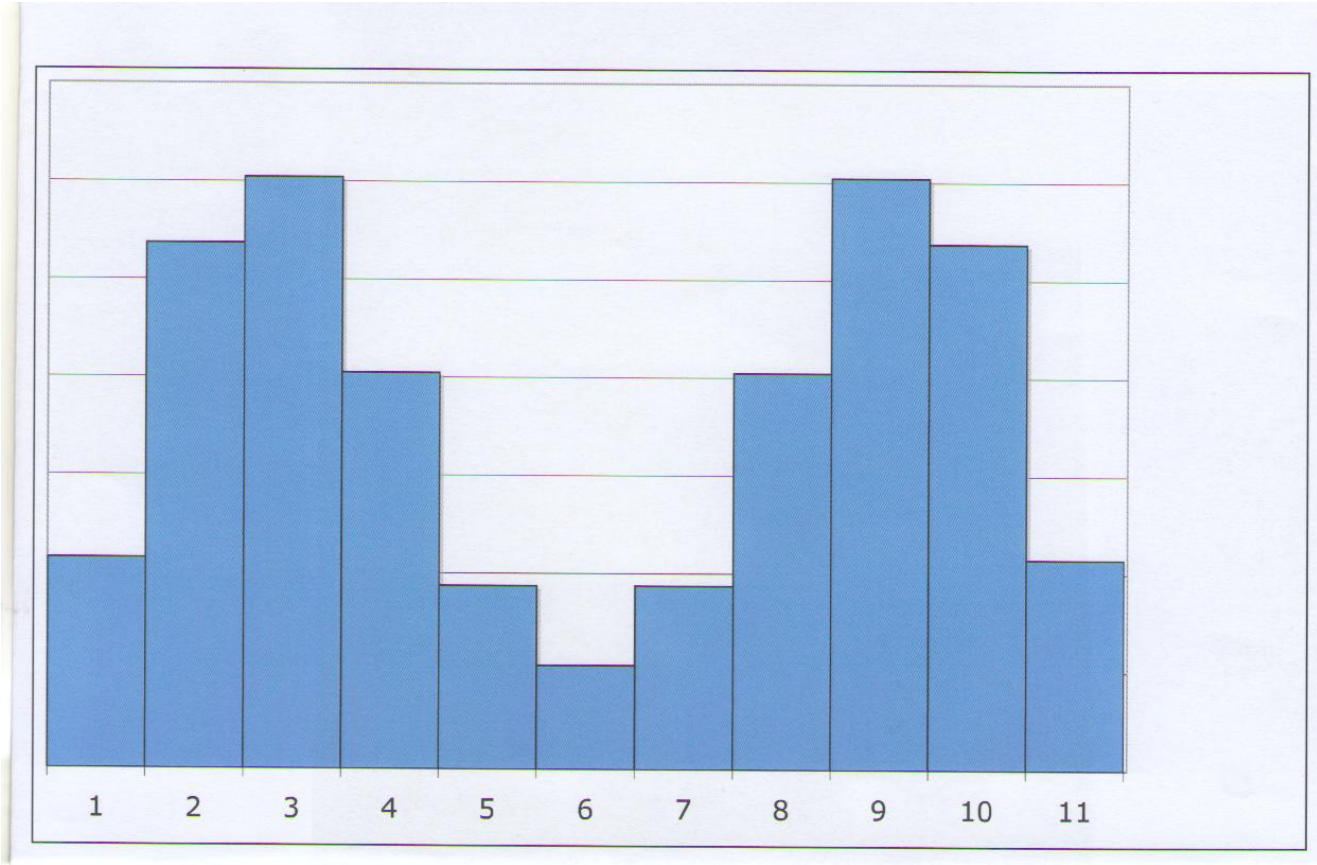
mentre per valori lontani dalla moda le frequenze sono piccole.

Grafico unimodale **asimmetrico**:



anche in questo caso le frequenze sono piccole per valori lontani dalla moda.

Bimodale



La moda assume due valori (ci sono di due distinte classi modali).

MEDIA

Si definisce **media campionaria** di un campione di taglia n, la media aritmetica dei dati

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Esempio. La rilevazione della temperatura massima (in gradi centigradi) fatta a Roma il 1 agosto per 10 anni ha fornito i seguenti dati:

30.1 28.2, 31.3, 22.6, 27.5, 29.4, 24.5, 27.8, 33.2, 22.8

La media campionaria vale

$$\bar{x} = \frac{1}{10}\Big(30.1 + 28.2 + 31.3 + 22.6 + 27.5 + 29.4 + 24.5 + 27.8 + 33.2 + 22.8\Big) = \frac{277.4}{10} = 27.74$$

PROPRIETÀ DELLA MEDIA

Se ciascun valore viene incrementato di una costante c (positiva o negativa) allora anche la media campionaria viene incrementata di c.

In formule: si definisce la variabile y con valori y_1,...y_n,

$$y_i = x_i + c, \qquad i = 1, \ldots n$$

allora, indicando con $\bar{y}$ la media campionaria della variabile y,

$$\bar{y} = \bar{x} + c$$

La dimostrazione di questa asserzione è una semplice verifica:

$$\bar{y} = \frac{1}{n}\Big[(x_1 + c) + (x_2 + c) + \ldots (x_n + c)\Big] = \frac{1}{n}\Big(x_1 + x_2 + \ldots x_n + nc)\Big) = \bar{x} + c$$

Esempio: x di taglia $n = 5$ con valori 6, 7, 5, 5, 7. La media campionaria è

$$\bar{x} = \frac{6 + 7 + 5 + 5 + 7}{5} = \frac{30}{5} = 6$$

Sommiamo $c = -5$ ed otteniamo la variabile y con valori

$$1 \quad 2 \quad 0 \quad 0 \quad 2$$

La media campionaria della variabile y è

$$\bar{y} = \frac{1 + 2 + 0 + 0 + 2}{5} = 1$$

per cui

$$\bar{y} = \bar{x} + c$$

Se ciascun valore viene moltiplicato per una costante a (positiva o negativa) allora anche la media campionaria viene moltiplicata per a.
In formule: si definisce la variabile y con valori y_1,...y_n,

$$y_i = a\,x_i, \qquad i = 1, \ldots n$$

allora, indicando con $\bar{y}$ la media campionaria della variabile y,

$$\bar{y} = a\bar{x}$$

La dimostrazione di questa asserzione è una semplice verifica:

$$\begin{aligned}\bar{y} &= \frac{1}{n}\left(a\,x_1 + a\,x_2 + \ldots a\,x_n\right) \\ &= \frac{a}{n}\left(x_1 + x_2 + \ldots x_n\right) = a\bar{x}\end{aligned}$$

Questa proprietà è utile "per cambiare unità di misura".

Nel 1994 le entrate medie mensili in Italia per la voce "turismo" sono state 3.192.310 lire. Per conoscerle in Euro basta usare che

$$\text{Euro} = \frac{\text{Lire}}{1936,27}$$

Quindi

$$\bar{y} = \frac{3.192.310}{1936,27} = 1.648,69$$

Riassumendo. Se a partire dai valori $x_1, \ldots x_n$ della variabile x si definisce la variabile y con valori

$$y_i = ax_i + c, \qquad a, c \text{ numeri qualsiasi}$$

si ha

$$\bar{y} = a\bar{x} + c$$

CALCOLO DELLA MEDIA CON LE FREQUENZE

Esempio. Abbiamo i seguenti dati disposti in una tabella delle frequenze assolute

Valore	**Freq. assoluta**
4	1
6	4
7	2
TOTALE	**7**

Quindi l'insieme dei dati originali è composto da 7 valori che disposti in modo crescente sono:

$$4 \quad 6 \quad 6 \quad 6 \quad 6 \quad 7 \quad 7$$

La media campionaria è dunque

$$\bar{x} = \frac{4+6+6+6+6+7+7}{7} = \frac{1 \cdot 4 + 4 \cdot 6 + 2 \cdot 7}{7} = 6$$

In generale sia x una variabile numerica di taglia n con k valori distinti ordinati $x_1 < x_2 <... \ x_k$ con frequenze assolute n_1, n_2, n_3,.....n_k $(n_1 + n_2 + \cdots + n_k = n)$. La media campionaria è data da

$$\bar{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \cdots + n_k \cdot x_k}{n}$$

Si osservi che

$$\bar{x} = \frac{n_1}{n} x_1 + \frac{n_2}{n} x_2 + \cdots + \frac{n_k}{n} x_k$$

e poichè $f_i = \frac{n_i}{n}$ è la frequenza relativa, la media puo' essere calcolata tramite le frequenze relative usando la formula

$$\bar{x} = f_1 x_1 + f_2 x_2 + \cdots + f_k x_k$$

Esercizio 1. Supponiamo di sapere che la metà dei valori di un campione sono uguali a 10, un sesto sono uguali a 20 e un terzo sono uguali a 30. Quanto vale la media campionaria?

$$\bar{x} = \frac{1}{2}\,10 + \frac{1}{6}\,20 + \frac{1}{3}\,30 = 18,33$$

Esercizio 2. Il numero di settimane trascorse per un campione di 7 persone da quando hanno completato il corso di guida a quando hanno ottenuto la patente sono

$$2 \quad 110 \quad 5 \quad 7 \quad 6 \quad 7 \quad 3$$

Calcolare la media campionaria:

$$\bar{x} = \frac{2 + 110 + 5 + 7 + 6 + 7 + 3}{7} = \frac{140}{7} = 20$$

Tutti i valori tranne 1 sono molto minori della media. Un punto debole della media come indicatore del centro di un insieme di dati. è che il suo valore è ampiamente influenzato da un valore estremo.

MEDIANA

Si introduce un'altro indicatore che indichiamo con m= **mediana campionaria**.

Si dispongono i valori degli n dati in ordine crescente:

- Se n è dispari allora m è il **valore intermedio**.
- Se n è pari allora m è la **media dei due valori intermedi**.

In altri termini, i dati x_1,... x_n sono disposti in ordine crescente

- se n è **dispari** $m = \mathbf{x}_{\frac{n+1}{2}}$
- se n è **pari** $m = \frac{1}{2}[\mathbf{x}_{\frac{n}{2}} + \mathbf{x}_{\frac{n}{2}+1}]$

La definizione assicura che a destra della mediana cadono lo stesso numero di dati che a sinistra.

Esempio 1. Calcoliamo la mediana per i dati dell'esercizio 2

$$2 \quad 110 \quad 5 \quad 7 \quad 6 \quad 7 \quad 3$$

dove $n = 7$. In ordine crescente

$$2 \quad 3 \quad 5 \quad 6 \quad 7 \quad 7 \quad 110$$

La mediana campionaria è il **quarto** valore, quindi $m = 6$.
Esempio 2. Il campione è di taglia $n = 6$ ed i dati in ordine crescente sono

$$4 \quad 5 \quad 7 \quad 8 \quad 9 \quad 50$$

La mediana campionaria è la media aritmetica dei valori intermedi:

$$m = \frac{7+8}{2} = \frac{15}{2} = 7.5$$

La media campionaria di questi dati è

$$\bar{x} = \frac{4+5+7+8+9+50}{6} = \frac{83}{6} = 13,8$$

Osserva che la mediana campionaria non risente dei valori estremi. Infatti se il dato 50 viene sostituito con 10:

$$4 \quad 5 \quad 7 \quad 8 \quad 9 \quad 10$$

la mediana è la stessa mentre **la media diventa 7,1**

$$\bar{x} = \frac{4+5+7+8+9+10}{6} = \frac{43}{6} = 7,1$$

E' informativo calcolare entrambe le statistiche per descrivere la tendenza centrale di un insieme di dati. La media campionaria prende in considerazione tutti i valori, invece la mediana considera soltanto 1 o 2 valori centrali dei dati e quindi non dipende dai valori estremi.

Esercizio 1. I dati di una variabile numerica di taglia $n = 11$ sono:

Valore	**Freq. assoluta**
10	3
11	1
16	3
21	1
23	3
TOTALE	**11**

Determinare la media campionaria e la mediana campionaria.
Poichè $n = 11$ è dispari la mediana è

$$m = x_{\frac{11+1}{2}} = x_6 = 16$$

$$\bar{x} = \frac{1}{11}[3x10 + 11 + 3x16 + 21 + 3x23] = 16,27$$

Per dati approssimativamente simmetrici rispetto alla mediana la media e la mediana sono vicine.

Esercizio 2.
Una compagnia di assicurazioni ha rilevato il numero di incidenti nel periodo 1996-2000 relativo a 25 assicurati

$$0, 1, 0, 2, 5, 0, 1, 4, 3, 2, 0, 1, 0, 5, 2, 0, 0, 6, 1, 1, 0, 3, 1, 2, 2$$

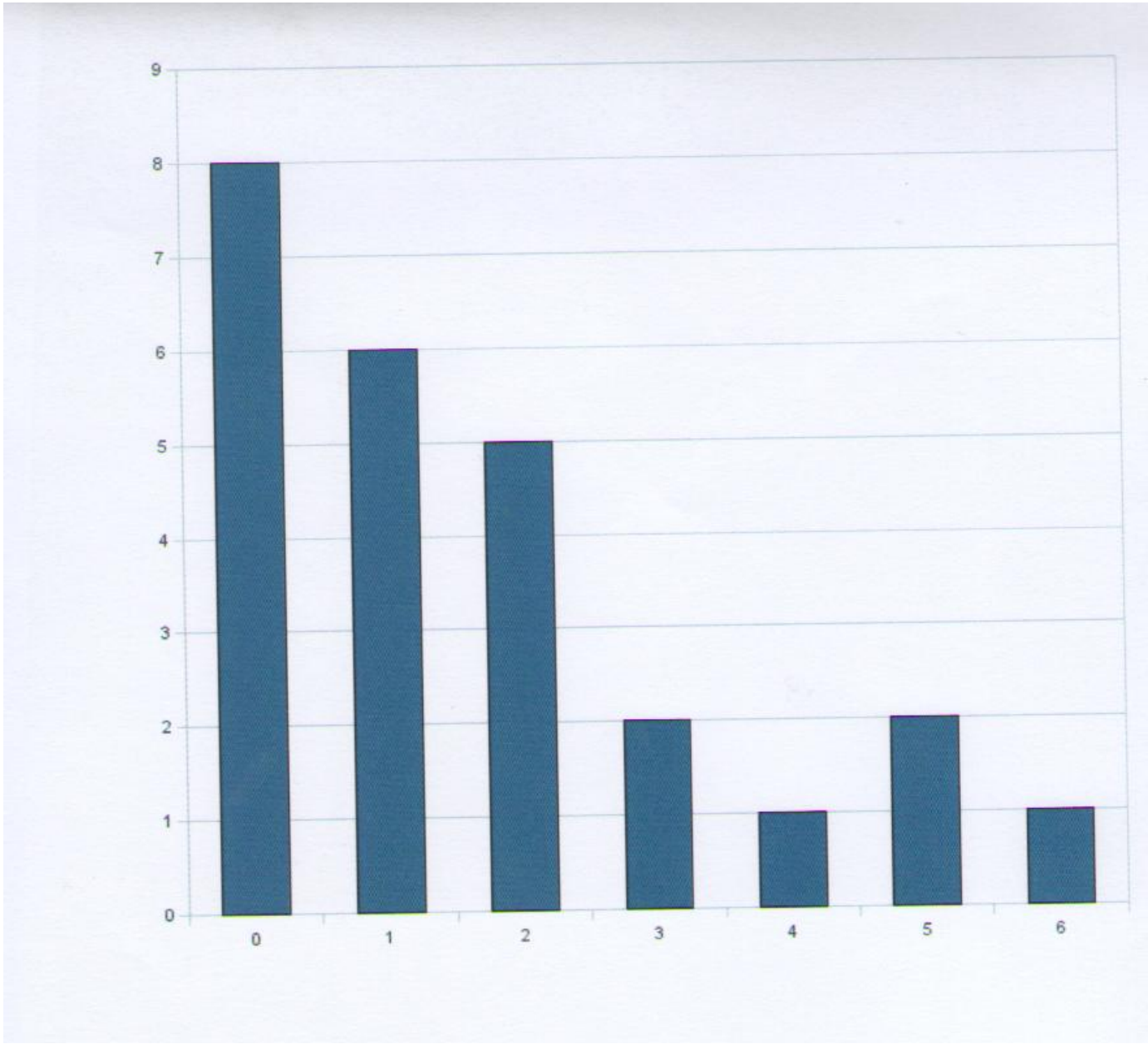
(1) Rappresentare i dati in una tabella delle frequenze e con un diagramma a barre.
(2) Calcolare la media, la mediana e la moda campionarie.
(3) Con quale frequenza non si è dovuto risarcire più di un sinistro?

Ordiniamo i dati

$$0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 5, 5, 6$$

Valori	**Freq. assoluta**	**Freq. relativa**	**Freq. percentuale**
0	8	0.32	32
1	6	0.24	24
2	5	0.2	20
3	2	0.08	8
4	1	0.04	4
5	2	0.08	8
6	1	0.04	4
TOTALE	**25**	**1**	**100**

Diagramma a barre



La media campionaria è :

$$\bar{x} = \frac{6 + 2 \cdot 5 + 3 \cdot 2 + 4 + 5 \cdot 2 + 6}{25} = \frac{42}{25} = 1,68$$

La taglia del campione è $n = 25$ quindi la mediana è il valore corrispondente al tredicesimo valore nella lista ordinata:
$m = x_{13} = 1$

La moda è il valore con la frequenza più alta ed è quindi uguale a 0.

Per calcolare con quale frequenza relativa non si è dovuto risarcire più di un sinistro, osserviamo che il numero degli assicurati che ha subito al più un incidente (cioè o 0 o 1) è dato dalla somma delle frequenze assolute del valore 0 e del valore 1 = 8+6=14.
Quindi la frequenza cercata è $\frac{14}{25} = 0.56$.