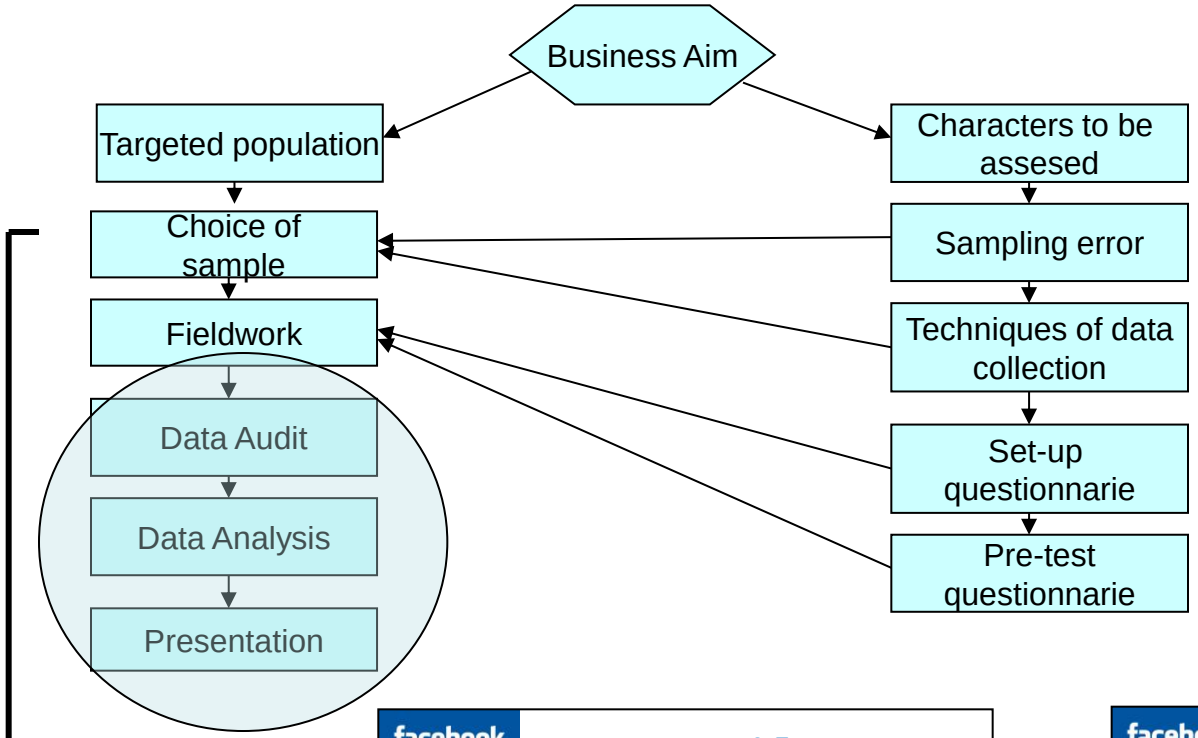
Metodi Quantitativi per Economia, Finanza e Management

Lezione n°3

Analisi Univariata

Quantitative Market Research

Set-up Protocol



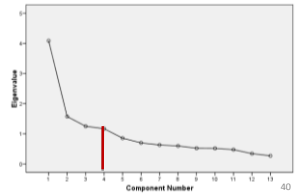
facebook

4 Factors

Our choice was consistent with the following criteria:
-The proportion between the maximum number of variables and the chosen factors is in the acceptable range (4/13 < 30%)
-The Eingenvalues are all bigger than 1
-The Cumulative Variance Explained is over 60%
-Communalities homogeneous values

NUMBER OF FACTORS (K/P)	OK
EINGENVALUE (> 1)	OK
% global explained variance (Between 60 and 75%)	62% - OK
Communalities - Low difference between highest and lowest (+-30%)	79% - 50% = 29% - OK

Scree Plot



As the Scree Plot confirms, only after 4 components the slope of the curve sensibly decreases.

40

facebook

The 5 Clusters

- **Cool Hunters (28%)**: More than all, they are users absolutely interested on Broadening.
- **PR's (7%)**: Interested above all in Public Relations and express some attachment to Spying, but not related at all with Keeping Up.
- **Detached (20%)**: Apart from some light interest on Broadening, they do not express any involvement with the Facebook use (in particular with Public Relations).
- **Functional (18%)**: Above all, interested in Keeping up with their network of friends and use Public Relations inside this network. Besides, they do not care at all about Spying and Broadening.
- **Gossipers (27%)**: They are also interested in Keeping up, but above all in Spying their network. Furthermore, they are not interested in Public Relations and Broadening.

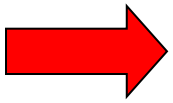
Each single Cluster was then crossed with socio-demographic and usage variables, through the contingency table tool, in order to better understand their main characteristics. The following slides sum-up the most relevant results of these crossings for each single cluster.

68

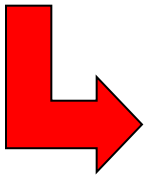
Univariate descriptive statistics

In the univariate descriptive statistics we analyze one variable at a time.

N_ID	D_8_2
H1	0.1
H2	0
H3	0
H4	0.2
H5	0.05
H6	0.2
H7	0.1
H8	0.1
H9	0.2
H10	0.05
H11	0
H12	0
H13	0
H14	0.15
H15	0
H16	0.1
H17	0
H18	0.2
H19	0
H20	0.05
H21	0.2
H22	0.2
⋮	⋮
H234	0.2
H235	0.1
H236	0.1



- Frequency distribution
- Synthesis measures
  - *Measures of location*
  - *Measures of spread*
  - *Measures of shape*



- Data Audit
  - Input errors
  - Missing values
  - Outliers
- Basic insights

Le distribuzioni di frequenza

- *Frequenza assoluta*: è un primo livello di sintesi dei dati- consiste nell'associare a ciascuna categoria, o modalità, il numero di volte in cui compare nei dati

- *Distribuzione di frequenza*: insieme delle modalità e delle loro frequenze

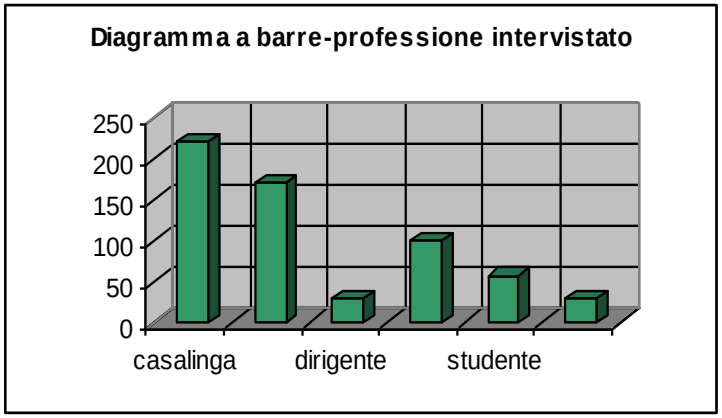
- *Frequenza relativa*: rapporto tra la frequenza assoluta ed il numero complessivo delle osservazioni effettuate.

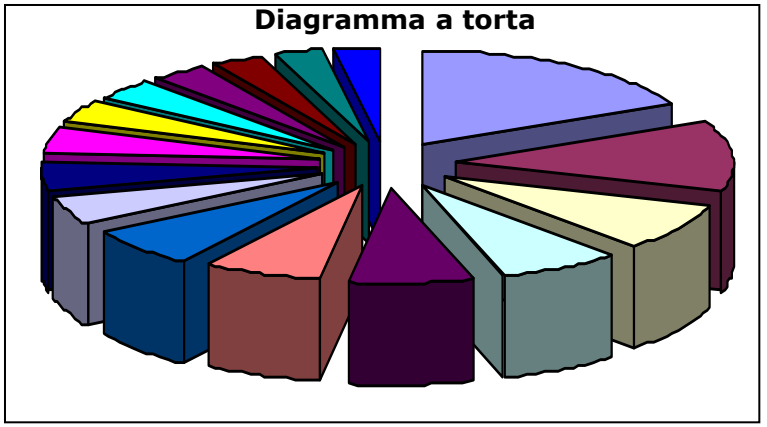
$$p_i = n_i / N$$

I due tipi di frequenze vengono usati con dati qualitativi e quantitativi discreti.

Le distribuzioni di frequenza

- *Rappresentazione grafica var.qualitative:*



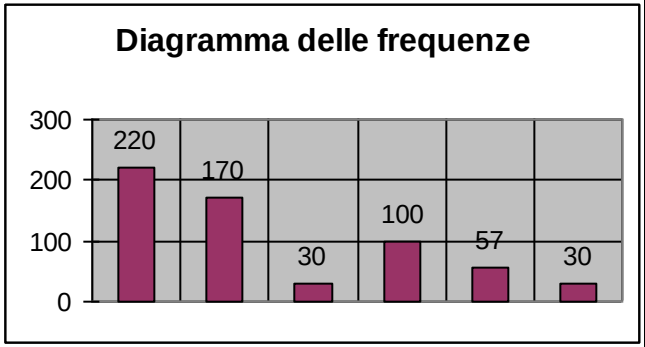


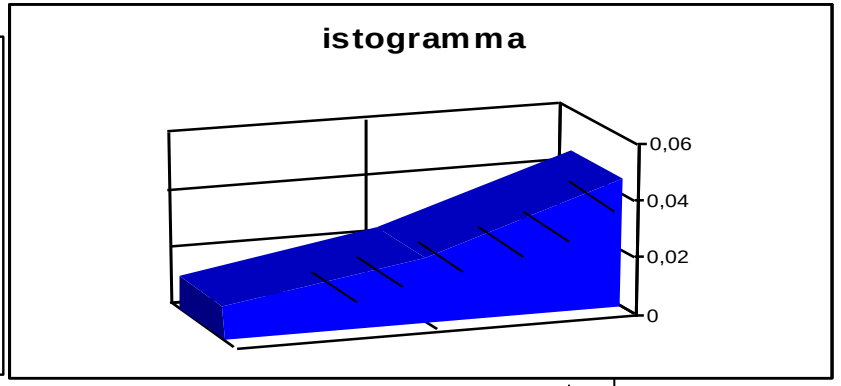
Diagr. a barre: nell'asse delle ascisse ci sono le categorie, senza un ordine preciso; in quello delle ordinate le frequenze assolute/relative corrispondenti alle diverse modalità

Diagr. a torta: la circonferenza è divisa proporzionalmente alle frequenze

Le distribuzioni di frequenza

- *Rappresentazione grafica var.quantitative discrete:*





Diagr. delle frequenze: nell'asse delle ascisse ci sono i valori assunti dalla var. discreta (quindi ha un significato quantitativo); l'altezza delle barre è proporzionale alle frequenze relative o assolute del valore stesso

*Istogramma:*nell'asse delle ascisse ci sono le classi degli intervalli considerati; l'asse delle ordinate rappresenta la densità di frequenza; l'area del rettangolo corrisponde alla frequenza della classe stessa.

Misure di sintesi

Misure di tendenza centrale:

- Media aritmetica
- Mediana
- Moda

Misure di tendenza non centrale:

- Quantili
- Percentili

Misure di dispersione:

- Campo di variazione
- Differenza interquantile
- Varianza
- Scarto quadratico medio
- Coefficiente di variazione

Misure di forma della distribuzione:

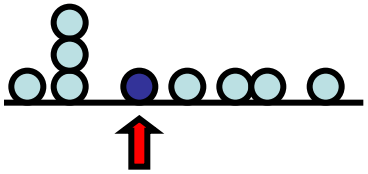
- Skewness
- Kurtosis

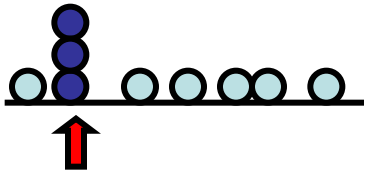
Misure di Tendenza Centrale

Tendenza Centrale

- **Media**
- **Mediana**
- **Moda**

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





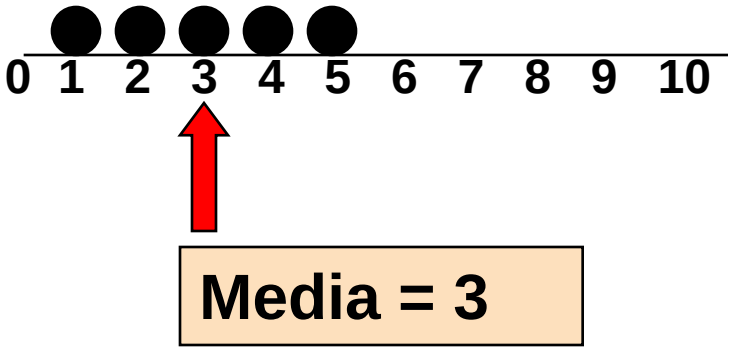
Media Aritmetica

Valore centrale delle osservazioni ordinate

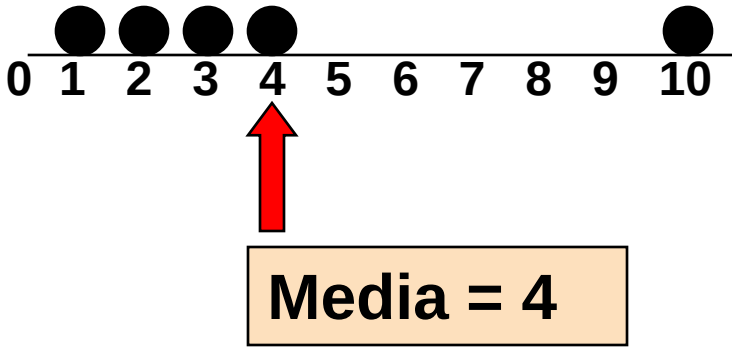
Valore più frequente

Media Aritmetica

- La misura di tendenza centrale più comune
- Media = somma dei valori diviso il numero di valori
- Influenzata da valori estremi (outlier)



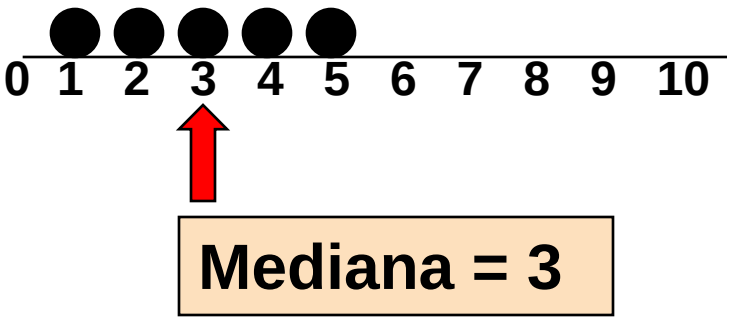
$$\frac{1+2+3+4+5}{5}=\frac{15}{5}=3$$

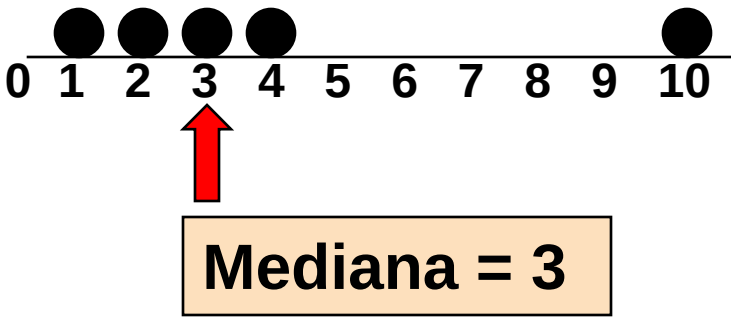


$$\frac{1+2+3+4+10}{5}=\frac{20}{5}=4$$

Mediana

- In una lista ordinata, la mediana è il valore “centrale” (50% sopra, 50% sotto)

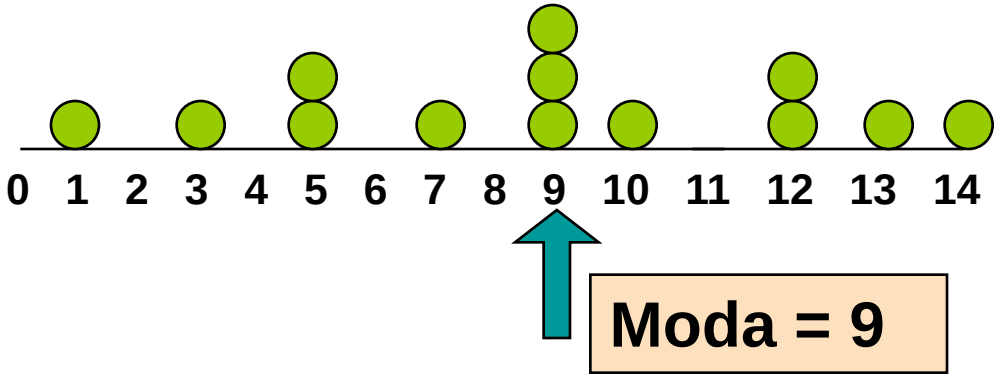


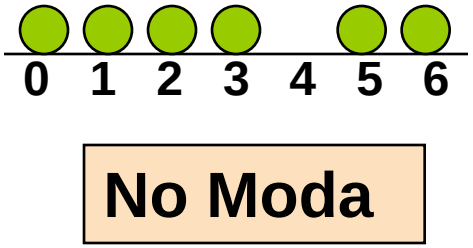


- Non influenzata da valori estremi

Moda

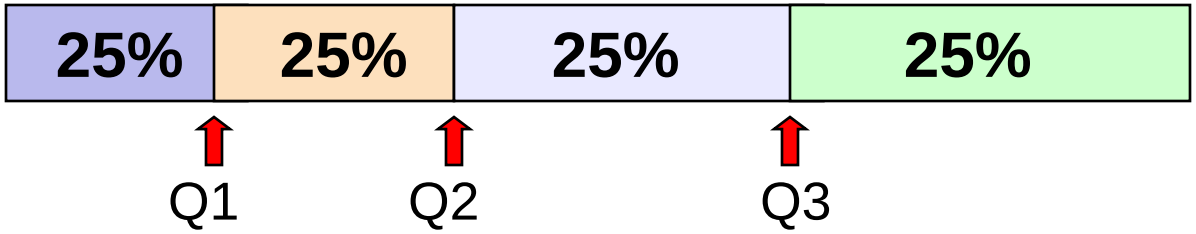
- Valore che occorre più frequentemente
- Non influenzata da valori estremi
- Usata sia per dati numerici che categorici
- Può non esserci una moda
- Ci può essere più di una moda





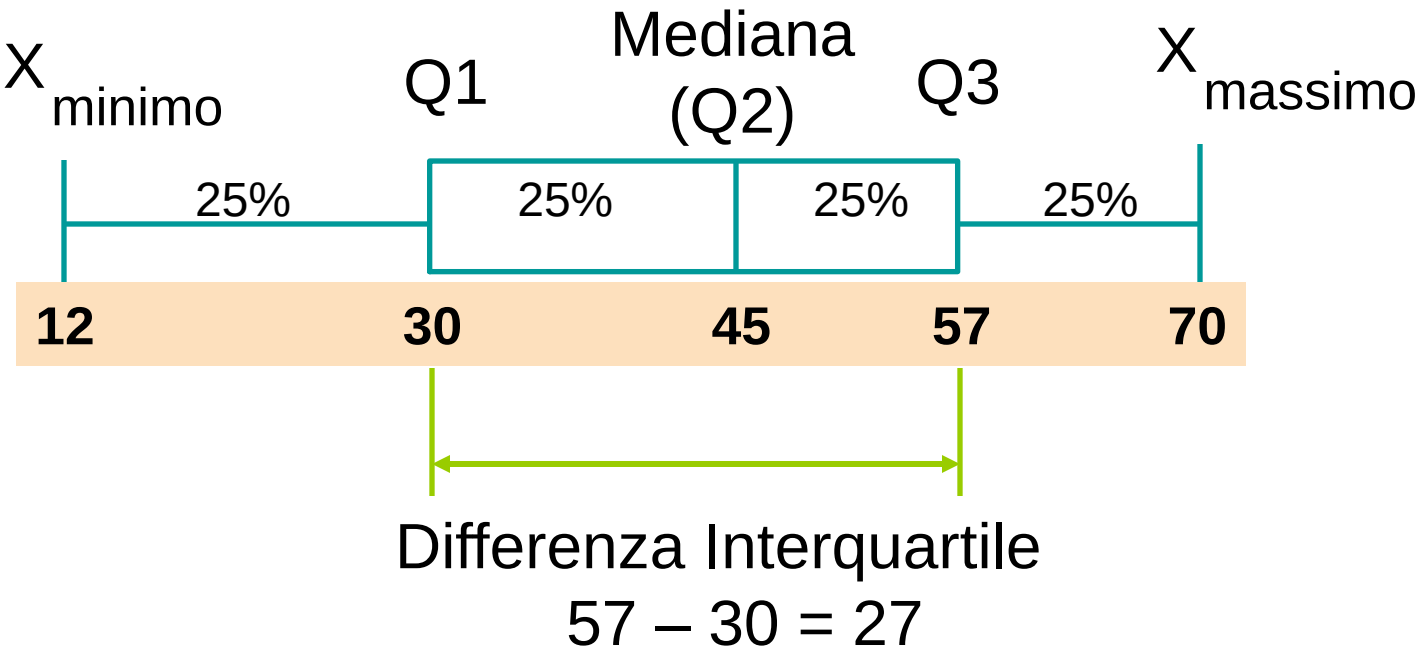
Misure di Tendenza Non Centrale

- I Quartili dividono la sequenza ordinata dei dati in 4 segmenti contenenti lo stesso numero di valori



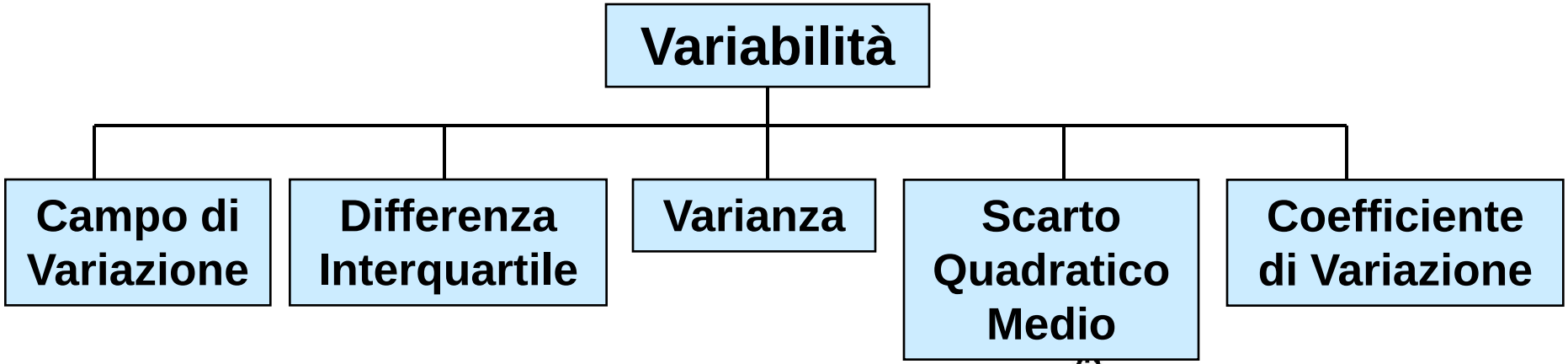
- Il primo quartile, Q_1, è il valore per il quale 25% delle osservazioni sono minori e 75% sono maggiori di esso
- Q_2 coincide con la mediana (50% sono minori, 50% sono maggiori)
- Solo 25% delle osservazioni sono maggiori del terzo quartile

Box Plot

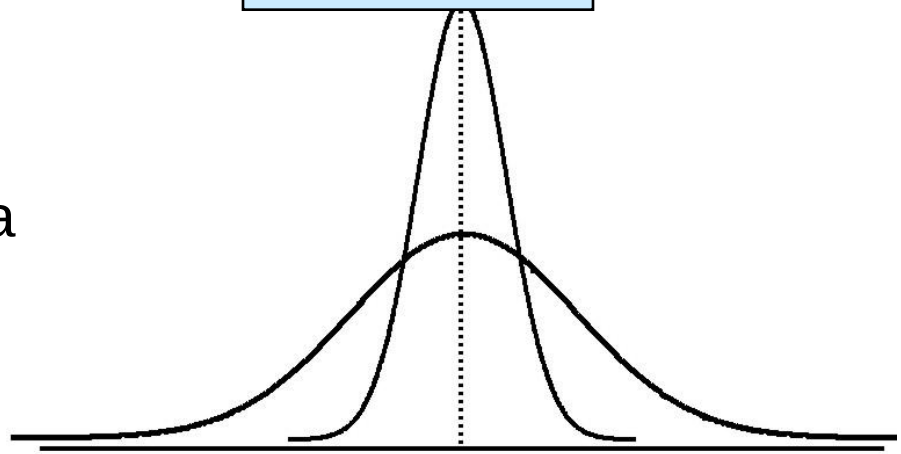


<u>OUTLIERS</u>:	Q1 - 1,5 * Differenza interquartile
	Q3 + 1,5 * Differenza interquartile

Misure di Variabilità



- Le misure di variabilità forniscono informazioni sulla **dispersione** o **variabilità** dei valori.



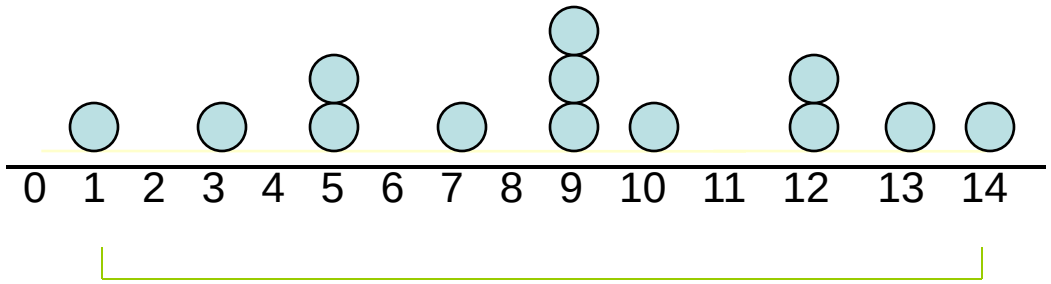
Stesso centro,
diversa variabilità

Campo di Variazione

- La più semplice misura di variabilità
- Differenza tra il massimo e il minimo dei valori osservati:

$$\text{Campo di variazione} = X_{\text{massimo}} - X_{\text{minimo}}$$

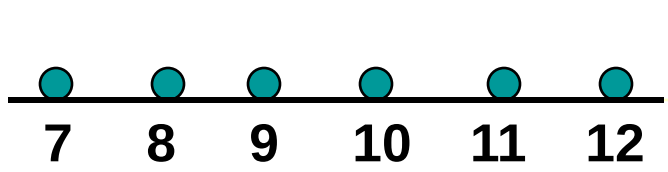
Esempio:



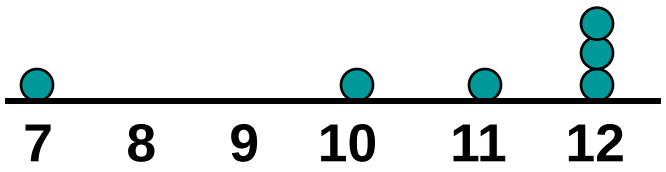
Campo di Variazione = 14 - 1 = 13

Campo di Variazione

- Ignora il modo in cui i dati sono distribuiti



Campo di Var. = 12 - 7 = 5



Campo di Var. = 12 - 7 = 5

- Sensibile agli outlier

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Campo di Var. = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Campo di Var = 120 - 1 = 119

Differenza Interquartile

- Possiamo eliminare il problema degli outlier usando la differenza interquartile
- Elimina i valori osservati più alti e più bassi e calcola il campo di variazione del 50% centrale dei dati
- Differenza Interquartile = 3º quartile – 1º quartile

$$IQR = Q_3 - Q_1$$

Varianza

- Media dei quadrati delle differenze fra ciascuna osservazione e la media

  – Varianza della Popolazione:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

dove μ = media della popolazione

N = dimensione della popolazione

x_i = i^{imo} valore della variabile X

Scarto Quadratico Medio

- Misura di variabilità comunemente usata
- Mostra la variabilità rispetto alla media
- Ha la stessa unità di misura dei dati originali

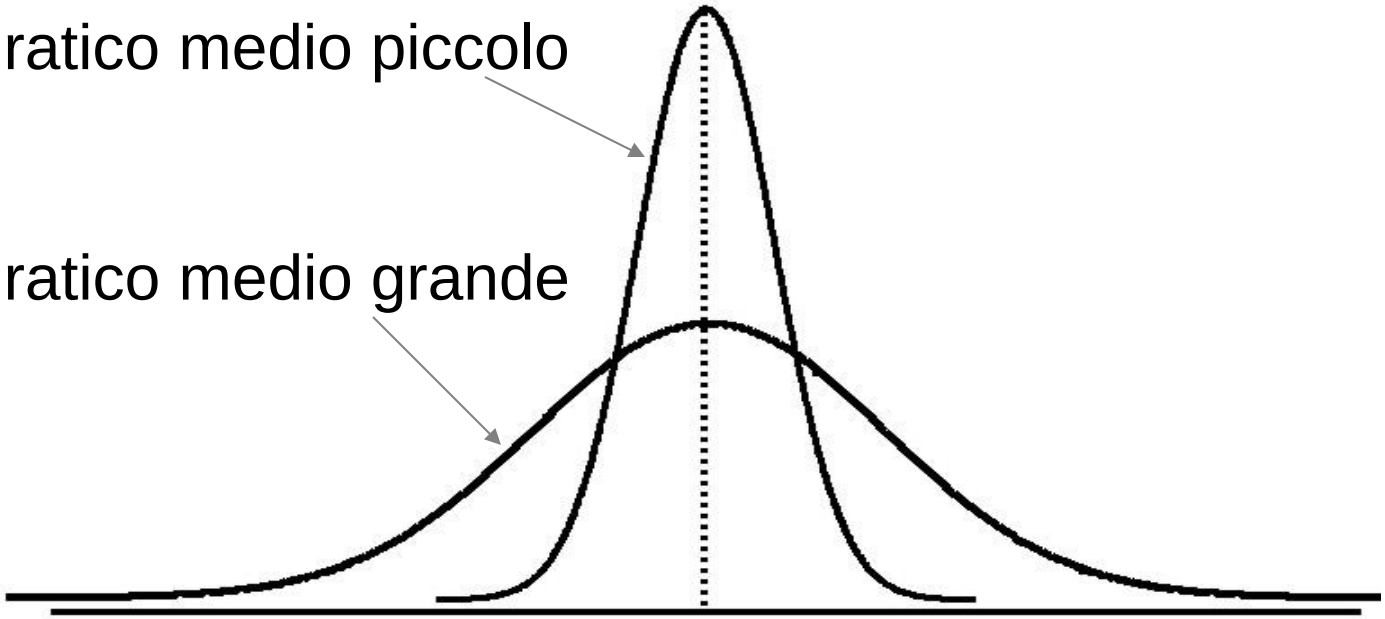
  – Scarto Quadratico Medio della Popolazione:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

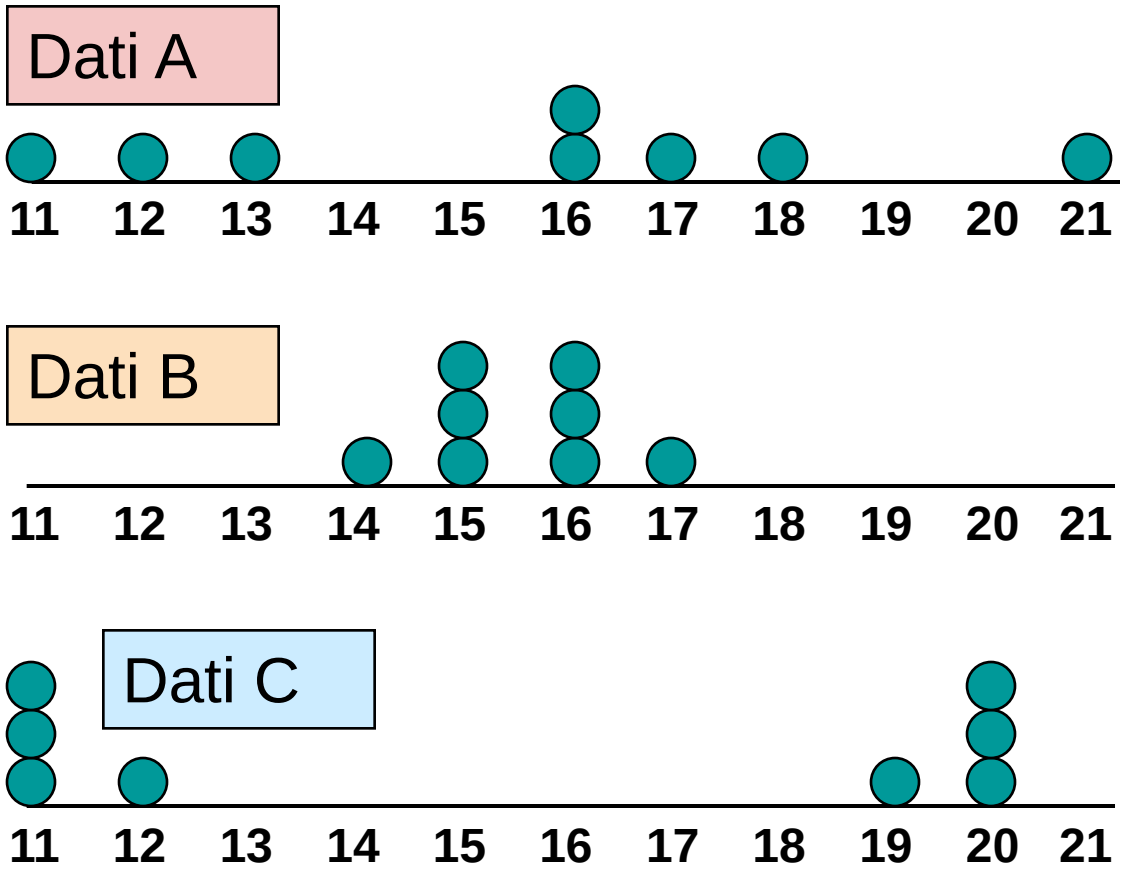
Scarto Quadratico Medio

Scarto quadratico medio piccolo

Scarto quadratico medio grande



Scarto Quadratico Medio



Media = 15.5
s = 3.338

Media = 15.5
s = 0.926

Media = 15.5
s = 4.570

Scarto Quadratico Medio

- Viene calcolato usando tutti i valori nel set di dati
- Valori lontani dalla media hanno più peso
  (poichè si usa il quadrato delle deviazioni dalla media)
- Le stesse considerazioni valgono anche per il calcolo della Varianza

Coefficiente di Variazione

- Misura la variabilità relativa
- Sempre in percentuale (%)
- Mostra la variabilità relativa rispetto alla media
- Può essere usato per confrontare due o più set di dati misurati con unità di misura diversa

$$CV = \left(\frac{s}{|\bar{x}|} \right) \cdot 100\%$$

Coefficiente di Variazione

- Azione A:
  - Prezzo medio scorso anno = $50
  - Scarto Quadratico Medio = $5

$$CV_A = \left(\frac{s}{|\bar{x}|} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Azione B:
  - Prezzo medio scorso anno = $100
  - Scarto Quadratico Medio = $5

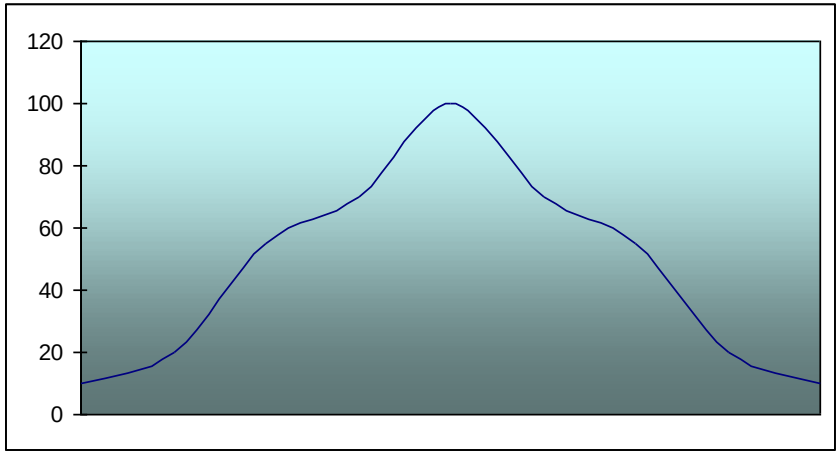
$$CV_B = \left(\frac{s}{|\bar{x}|} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

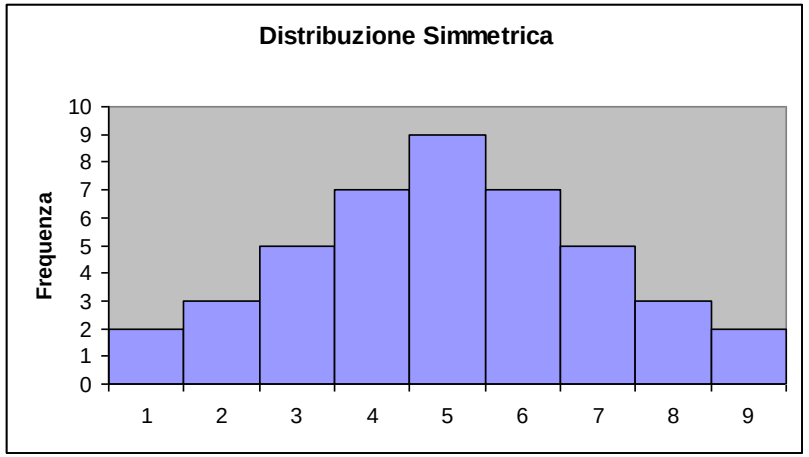
Entrambe le azioni hanno lo stesso scarto quadratico medio, ma

l'azione B è meno variabile rispetto al suo prezzo

Forma della Distribuzione

- La forma della distribuzione si dice simmetrica se le osservazioni sono bilanciate, o distribuite in modo approssimativamente regolare attorno al centro.

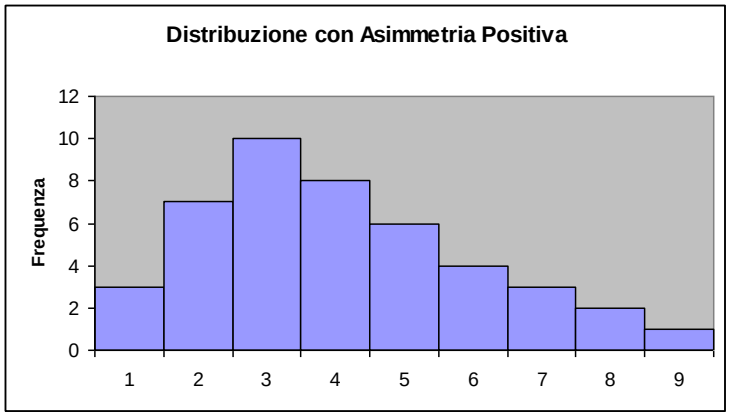




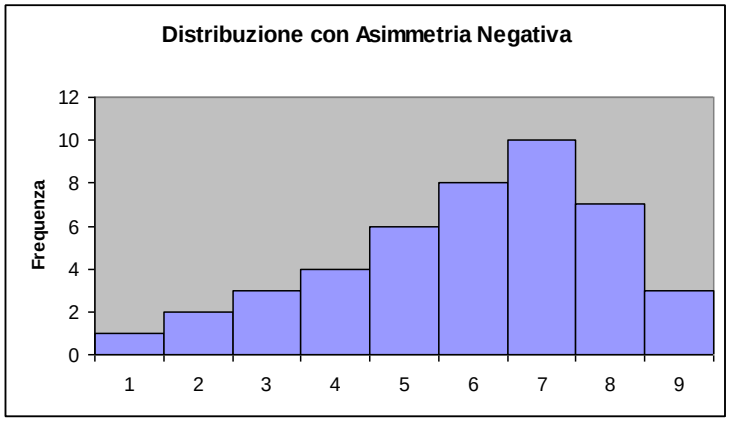
Forma della Distribuzione

- La forma della distribuzione è detta asimmetrica se le osservazioni non sono distribuite in modo simmetrico rispetto al centro.

Una distribuzione con asimmetria positiva (obliqua a destra) ha una coda che si estende a destra, nella direzione dei valori positivi.



Una distribuzione con asimmetria negativa (obliqua a sinistra) ha una coda che si estende a sinistra, nella direzione dei valori negativi.

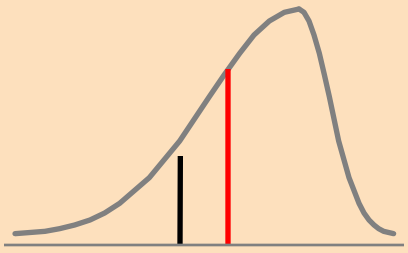


Misure di Forma della Distribuzione

- Descrive come i dati sono distribuiti
- Misure della forma
  - Simmetrica o asimmetrica

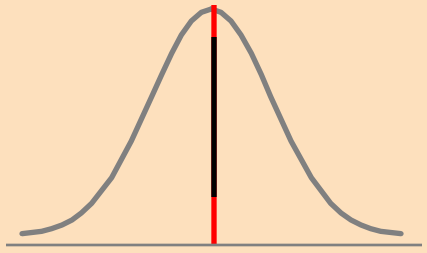
Obliqua a sinistra

Media < Mediana



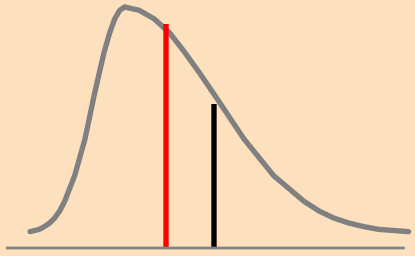
Simmetrica

Media = Mediana



Obliqua a destra

Mediana < Media



Misure di Forma della Distribuzione

Skewness: indice che informa circa il grado di simmetria o asimmetria di una distribuzione.

- $\gamma=0$ ditribuzione simmetrica;
- $\gamma<0$ asimmetria negativa (mediana>media);
- $\gamma>0$ asimmetria positiva (mediana<media).

Kurtosis: indice che permette di verificare se i dati seguono una distribuzione di tipo Normale (simmetrica).

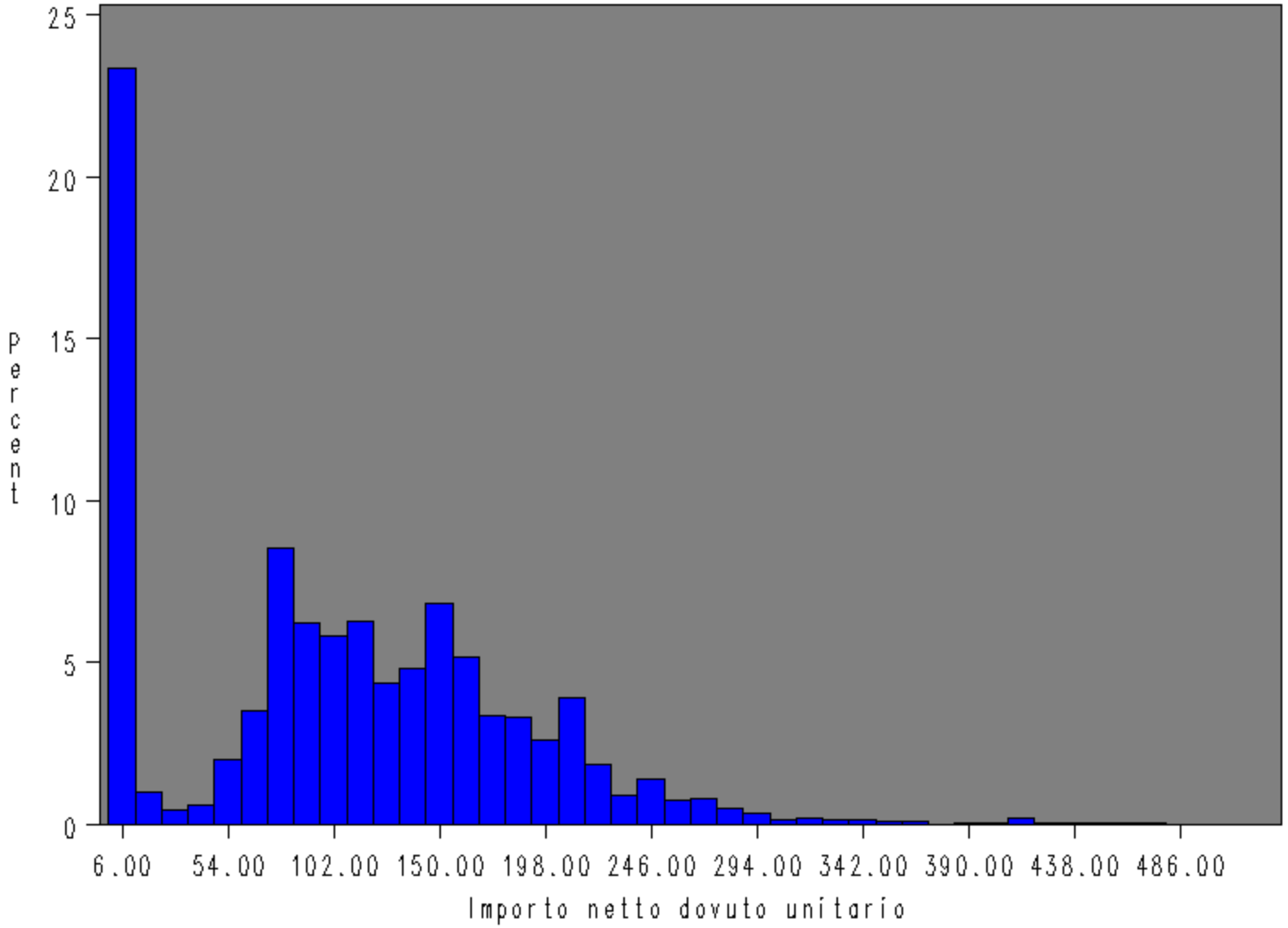
- $\beta=3$ se la distribuzione è “Normale”;
- $\beta<3$ se la distribuzione è iponormale (rispetto alla distribuzione di una Normale ha densità di frequenza minore per valori molto distanti dalla media);
- $\beta>3$ se la distribuzione è ipernormale (rispetto alla distribuzione di una Normale ha densità di frequenza maggiore per i valori molto distanti dalla media).

IMPORTO NETTO UNITARIO

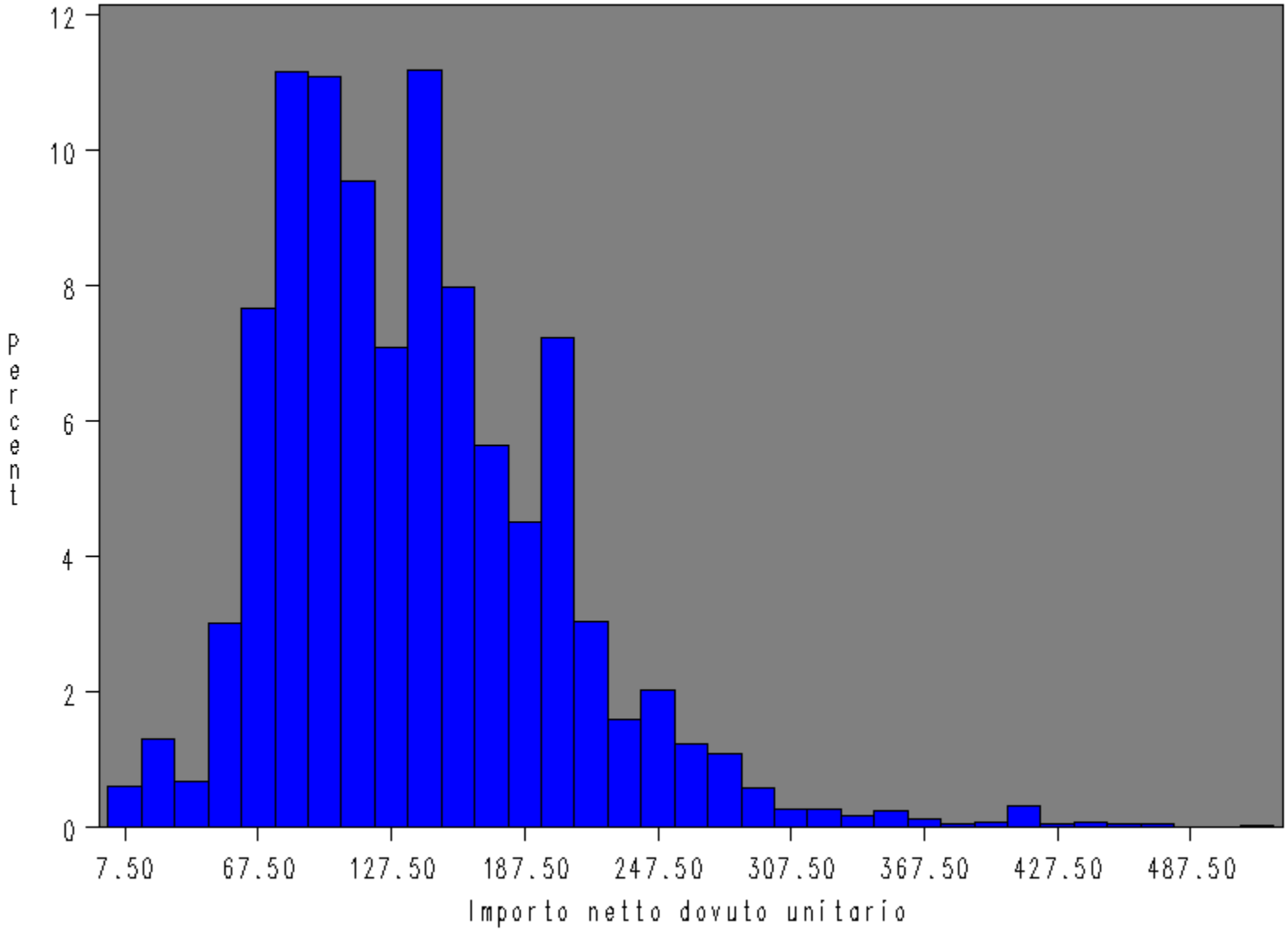
Basic Statistical Measures

Location		Variability	
Mean	106.1410	**Std Deviation**	81.01306
Median	103.2900	**Variance**	6563
Mode	0.0000	**Range**	523.69000
		Interquartile Range	118.62500

IMPORTO NETTO UNITARIO



IMPORTO NETTO UNITARIO



IMPORTO NETTO UNITARIO

Basic Statistical Measures

Location		Variability	
Mean	138.0247	**Std Deviation**	64.29397
Median	129.1100	**Variance**	4134
Mode	149.0000	**Range**	521.77000
		Interquartile Range	82.62000