Statistica descrittiva (M.S. Bernabei)

- La parola Statistica deriva dal latino “status” e significa stato. La Statistica va ben oltre la semplice rappresentazione dei dati, infatti essa si occupa
  - stabilire gli obiettivi dell’ indagine
  - della raccolta dei dati;
  - della descrizione e sintesi dei dati,
  - dell’ interpretazione dei dati in modo da trarre conclusioni sul fenomeno in esame

I principi e le metodologie della statistica sono utili per rispondere a domande del tipo:

- Che tipo e quanti dati è necessario raccogliere?
- Come dovremmo organizzare e interpretare i dati?
- Come possiamo analizzare i dati e dedurre le conclusioni?
- Come stimare la potenza delle conclusioni e giudicare la loro incertezza?

Esempi

- Programma di formazione
- Richieste di controllo pubblicitario
- Riproduzione delle piante
- Radio

Anche se gli esempi precedenti provengono da campi molto lontani tuttavia ci sono delle caratteristiche che li accomuna:

- per acquisire nuova conoscenza si ha bisogno di raccogliere dati rilevanti per il fenomeno in esame.
- anche se spesso le osservazioni siano fatte nelle "stesse condizioni" tuttavia è impossibile eliminare la variabilità dei dati.
- L'accesso all'intero insieme di dati è praticamente impossibile o da un punto di vista pratico non facile per le limitazioni di tempo, delle risorse e delle facilitazioni, così che dobbiamo lavorare con una informazione incompleta, cioè con i dati che abbiamo a disposizione nel corso di una studio sperimentale. Esempio radio.

- **Unità:** una singola unità è di solito una persona o un oggetto le cui caratteristiche o variabilità sono di interesse per l'indagine.
- **Popolazione**: è l'insieme di tutte le unità su cui si svolge l'indagine statistica. Essa può essere finita o infinita.
- **Campione**: sottoinsieme finito di unità della popolazione che sono state raccolte nel corso dell'indagine. Il campione dovrebbe essere il più possibile "rappresentativo" per la popolazione.
- E' importante l' ampiezza del campione?

Obiettivi della Statistica:

- fare inferenza sulla popolazione dall'analisi delle informazioni che sono contenute nel campione dei dati. Questo include una stima dell'incertezza contenuta in tale inferenza,

- delineare il processo e l'ampiezza del campione in modo che le osservazioni formino una base per fare valide inferenze

Statistica descrittiva

- Popolazione: insieme su cui si svolge l'indagine statistica. Essa può essere finita o indefinita.
- Campione: sottoinsieme finito della popolazione.
- Variabile: X grandezza che varia all'interno di una popolazione. Essa può essere numerica se i valori che essa può assumere sono numeri, in particolare essa si dice discreta se l'insieme dei valori è finito o numerabile, p.e. numero di chiamate ad un centralino, e continua se è continuo, p.e. altezze di una popolazione. Se la variabile non è numerica si dice categorica, per esempio gruppi sanguinei.

Frequenza assoluta: f_i è il numero di volte in cui la variabile X assume il valore quantitativo o qualitativo x_i

Se la variabile X assume i valori $x_1, x_2, \ldots, x_k$, con frequenze
Rispettivamente $f_1, f_2, \ldots, f_k$ allora

$$f_1 + f_2 + \ldots + f_k = n$$

Dove n è il numero totale degli oggetti del campione k è il
numero delle classi

Frequenza relativa: è il rapporto tra la frequenza assoluta E il numero degli elementi del campione, cioè

$$p_i = f_i / n, \ (i=1,\ldots,k)$$

$$p_1 + p_2 + \ldots + p_k = 1 .$$

Frequenza cumulata: p_{ci} fino al valore x_i è la somma delle Frequenze relative fino alla p_i, cioè

$$p_{ci} = p_1 + p_2 + \ldots + p_i \quad (i=1,2,\ldots,k) .$$

Tabella di frequenza per variabile numerica discreta

ESEMPIO 2 (variabile numerica discreta): Numeri e percentuali di donne inglesi di 40 anni e più intervistate sul numero di figli avuti.

Numero di figli	Numero delle donne Freq. Ass.	% delle donne Freq. rel.	Percentuale cumulata
0	354	12,5	12,5
1	414	14,6	27,1
2	1130	39,9	67,0
3	567	20,0	87,0
4	246	8,7	95,7
5	66	2,3	98,0
6	33	1,2	99,2
7	11	0,4	99,6
8	6	0,2	99,8
9	1	0,0	99,8
10	1	0,0	99,8
TOTALE	2829	99,8	

Fonte: General Household Survey 1995-96

TABELLA 2

Distribuzione di frequenza per variabili continue

Nel caso in cui la variabile X assume valori in un certo intervallo.

ESEMPIO 3: La seguente tabella rappresenta il consumo di alcool in unità di alcool data da un bicchierino di cognac o da un boccale di birra forniti dall' Ufficio Nazionale di Statistica, 1998.

Classi di etˆ	freq. assol. f_i	freq.rel. p_i	freq.cumulata
16-24	1850	0,12	0,12
25-44	5800	0,37	0,49
45-64	4724	0,30	0,79
65-84	3281	0,21	1,00
TOTALE	156555	1,00	

Esercizio

Conteggio del numero di errori di stampa per pagina (variabile discreta) riscontrati su un testo di 45 pagine:

5 6 3 4 7 2 3 2 3 2 6 4 3 9 3

2 0 3 3 4 6 5 4 2 3 6 7 3 4 2

5 1 3 4 3 7 0 2 1 3 1 5 0 4 5

Distribuzione di frequenze assolute e relative (arrotondate) degli errori delle 45 pagine:

classe(x_i)	freq.ass.f_i	freq.rel.p_i	freq.cumul. p_{ci}
0	3	0,0667	0,0667
1	3	0,0667	0,1333
2	7	0,1556	0,2889
3	12	0,2667	0,5556
4	7	0,1556	0,7111
5	5	0,1111	0,8222
6	4	0,0889	0,9111
7	3	0,0667	0,9778
8	0	0,0000	0,9778
9	1	0,0222	1,0000
	45	1	

TABELLA 2

Grafici di distribuzioni di frequenze

Le distribuzioni contenute nella distribuzione di frequenza possono essere rappresentate graficamente. Ciò permette di sintetizzare i dati.

Distribuzione di frequenza per variabili continue

Nel caso in cui la variabile X assume valori in un certo intervallo.

ESEMPIO 3: Raggruppamento in classi di una variabile continua, altezza in centimetri di 40 piante

111-119-130-170-143-156-126-113-127-107-83-
100-128-143-127-117-125-64-119-130-120-108-95-
192-124-129-143-198-131-163-152-104-119-161
178-135-146-158-176-98

Procedura:

•individuare il valore minimo (64) e massimo (198)

•stabilire l' intervallo di variazione, cioè un intervallo che comprenda tutti i dati, i cui estremi non si discostino troppo dai valori minimo e massimo, per esempio (60, 200), la cui ampiezza è 200-60=140.

•sulla base di $\sqrt{n}$ (n è l'ampiezza del campione, $\sqrt{40}$ nell'esempio è compreso tra 6 e 7) si decide il numero di classi (con ampiezza 20=140/7).

•Ogni classe ha la stessa ampiezza.

I passi da fare per costruire una distribuzione di frequenza sono i seguenti:

- individuare il valore minimo e quello massimo nell'insieme dei dati;
- scegliere intervalli di uguale lunghezza che ricoprono tutti i dati dal minimo al massimo senza sovrapposizione. Tali intervalli sono chiamati **classi.** Per far ciò si potrebbe calcolare l' intervallo di variazione, cioè la differenza tra il massimo e il minimo. Per convenienza si potrebbe allargare tale intervallo di variazione, tenendo conto che i suoi estremi non dovrebbero discostarsi troppo dai valori minimo e massimo. Il numero di classi è stabilito sulla base di $\sqrt{n}$ *(naturalmenet si prende la parte intera)*. Per ottenere l' ampiezza di una classe si divide l'intervallo di variazione per il numero di classi.
- Si calcola il numero di dati che appartengono a coascuna classe. Tale numero è la **frequenza della classe.**

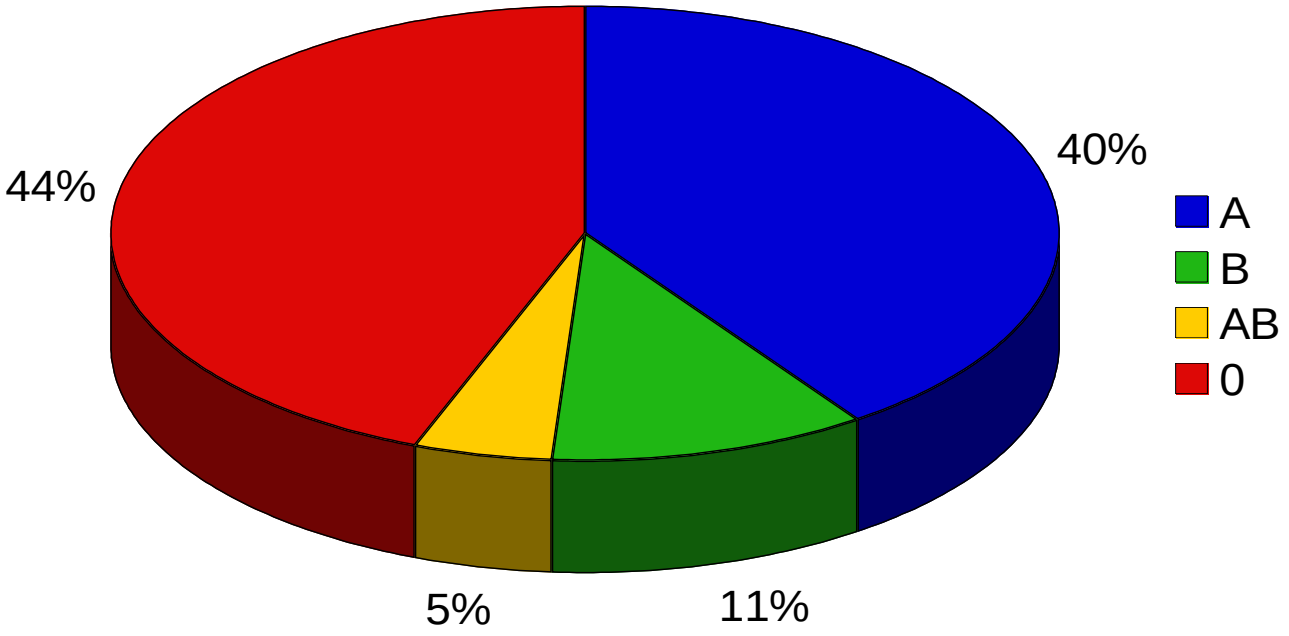
• Con la tabella distribuzione di frequenza si perde l'informazione di come sono distribuiti i dati all'interno di ciascuna classe. Per questo si prende come punto di riferimento il valore centrale di ciascuna classe.

class	val.cent $.x_i$	freq.as-sol.f_i	freq.rel. p_i	freq.cumulata
60-80	70	1	0,025	0,025
80-100	90	3	0,075	0,1
100-120	110	10	0,25	0,35
120-140	130	12	0,3	0,65
140-160	150	7	0,175	0,825
160-180	170	5	0,125	0,95
180-200	190	2	0,05	1
		40	1	

TABELLA 3

Aerogrammi a torta

L'area di ciarcun settore è proporzionale alla frequenza:
$a_i:360° = f_i : n$



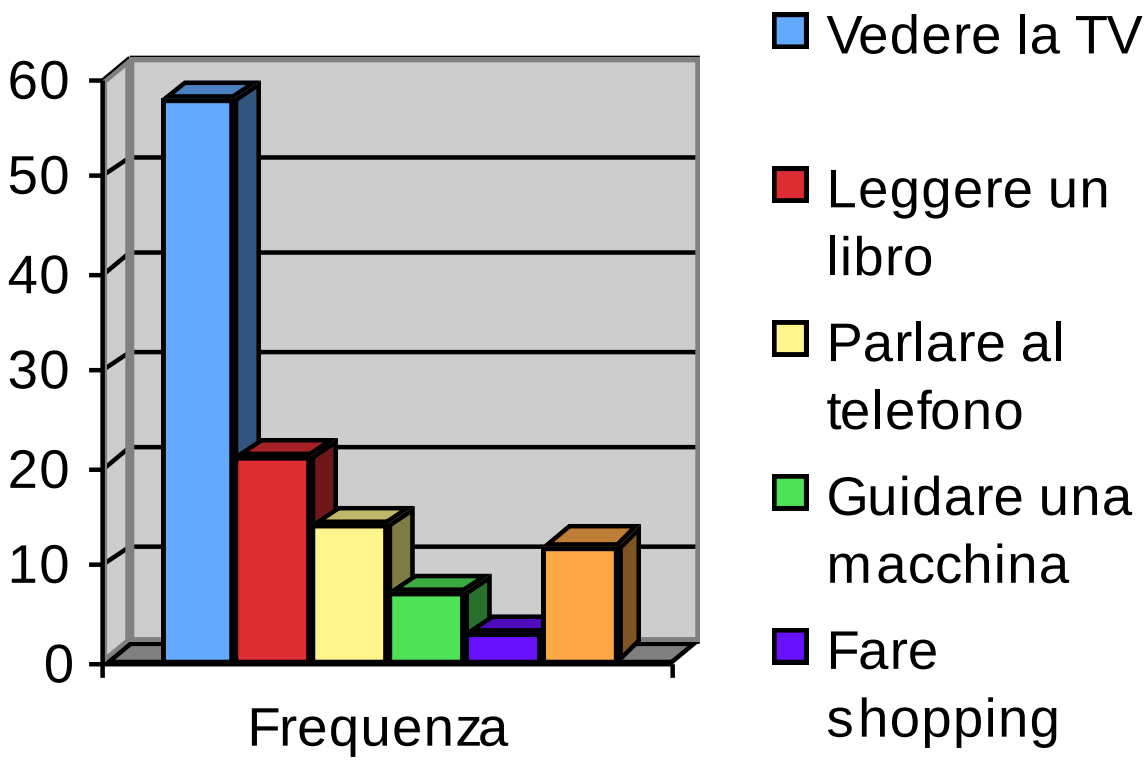
Diagrammi di Pareto

Principio di Pareto è sintetizzabile nell'affermazione: la maggior parte degli effetti è dovuta ad un numero ristretto di cause.

Esempio

Attività	Frequenza
Vedere la TV	58
Leggere un libro	21
Parlare al telefono	14
Guidare una macchina	7
Fare shopping	3
Altro	12

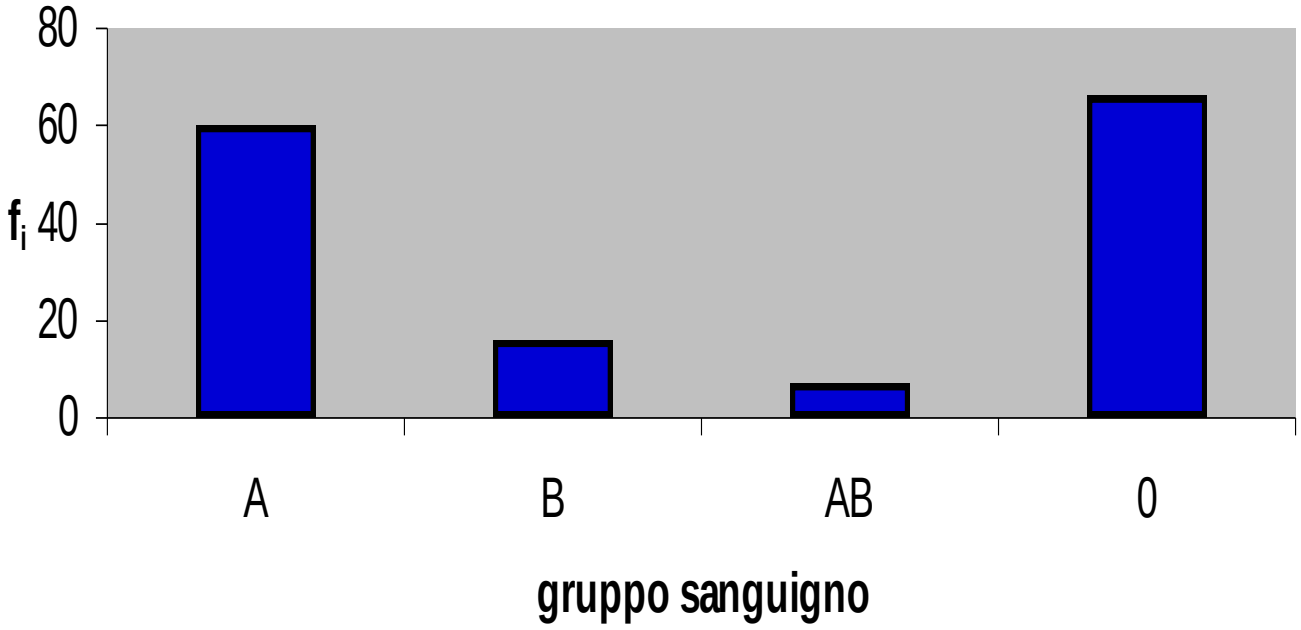
Pareto


60
50
40
30
20
10
0
Frequenza
Vedere la TV
Leggere un libro
Parlare al telefono
Guidare una macchina
Fare shopping
Altro

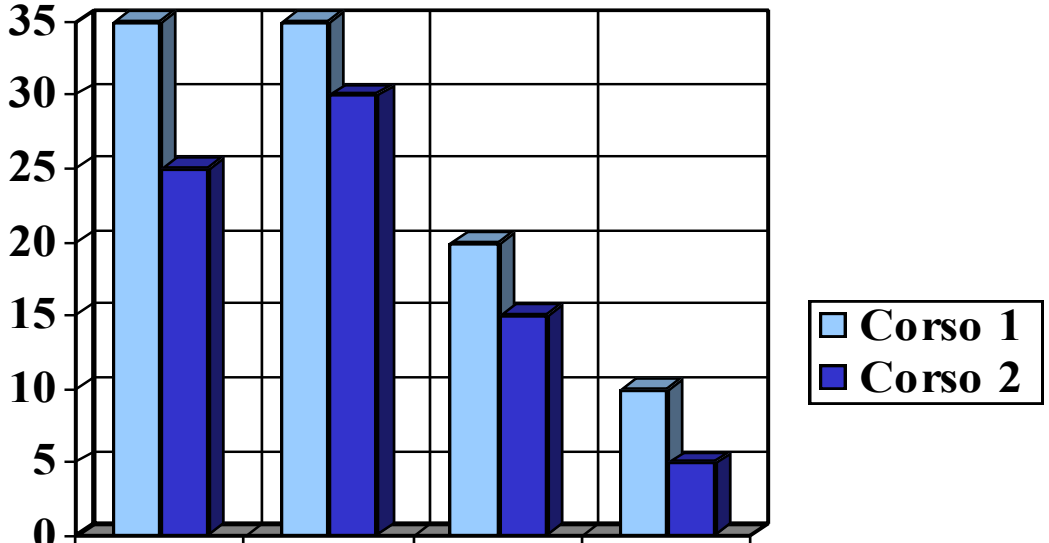
DIAGRAMMI A BARRE E ISTOGRAMMI :

Classe corrisponde una barra la cui base (per tutte uguali) non ha significato, mentre l'altezza rappresenta la frequenza della classe. Le barre non sono adiacenti per ricordare che sull'asse delle x non c'è alcuna unita di misura, e l'ordine delle barre ha significato nel caso in cui i valori della variabile si possono ordinare.

Consideriamo l'esempio 1. Il relativo diagramma a barre è il seguente:



Il seguente diagramma rappresenta i risultati di un esame per due corsi distinti:



	Insuff.	Suff.	buono	ottimo
Corso 1	35	35	20	10
Corso 2	25	30	15	5

Oppure si potrebbe rappresentare nel seguente modo:

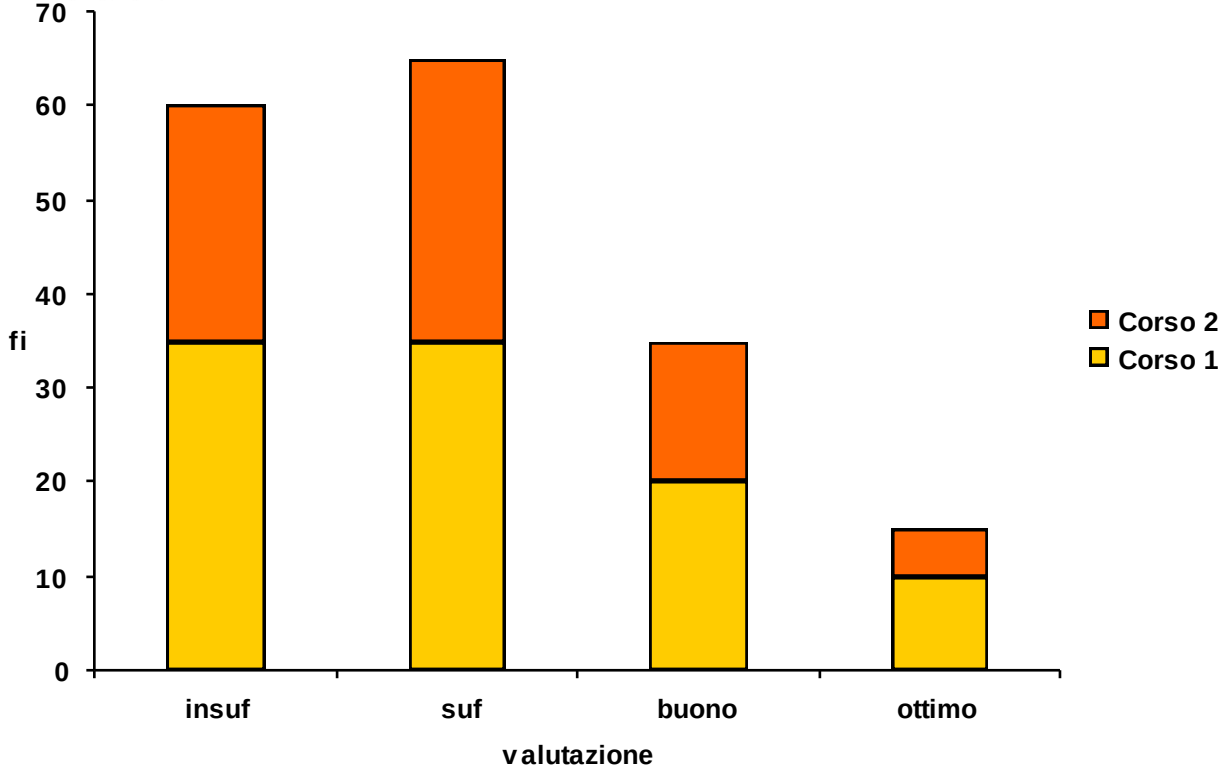
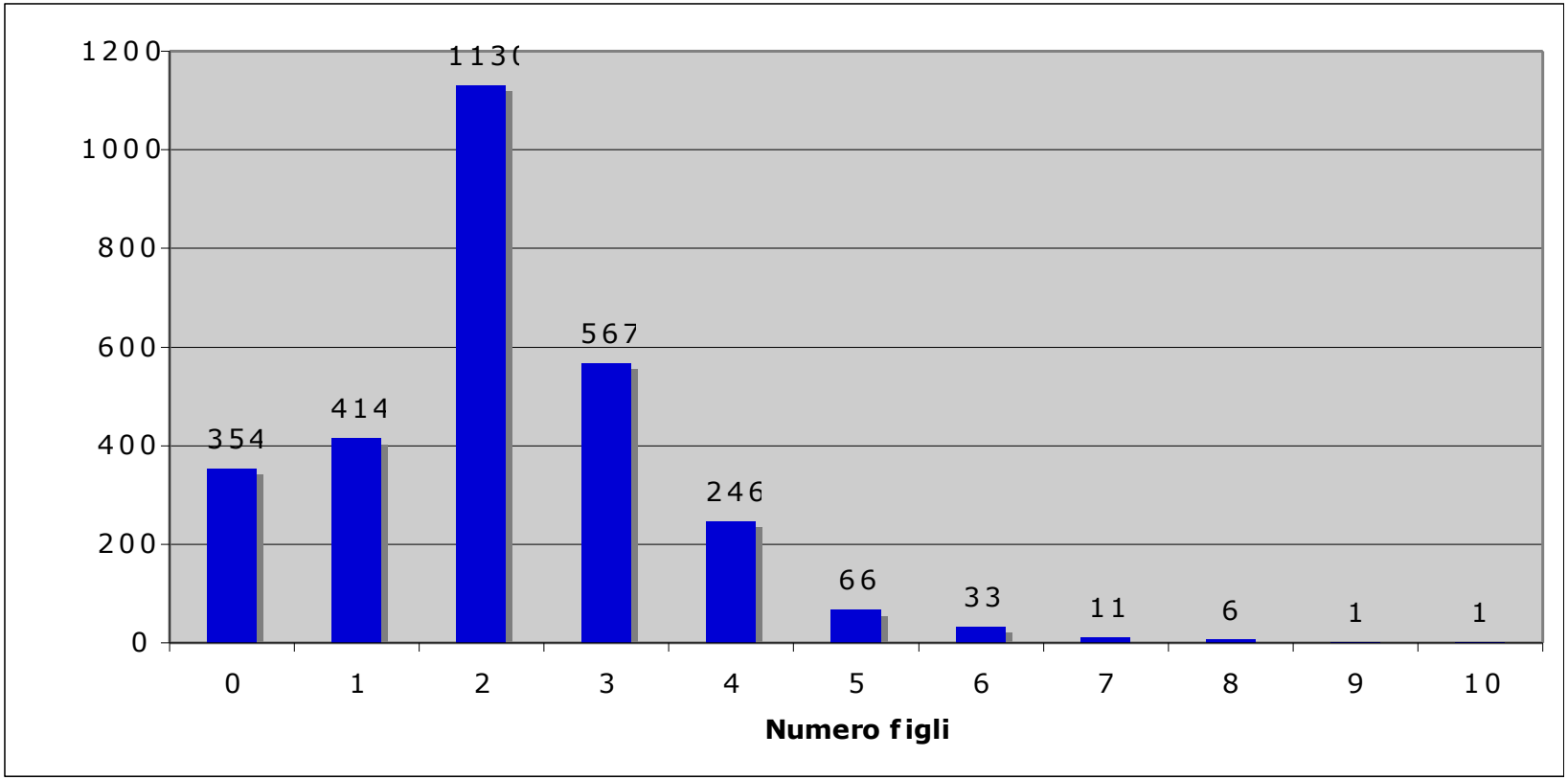
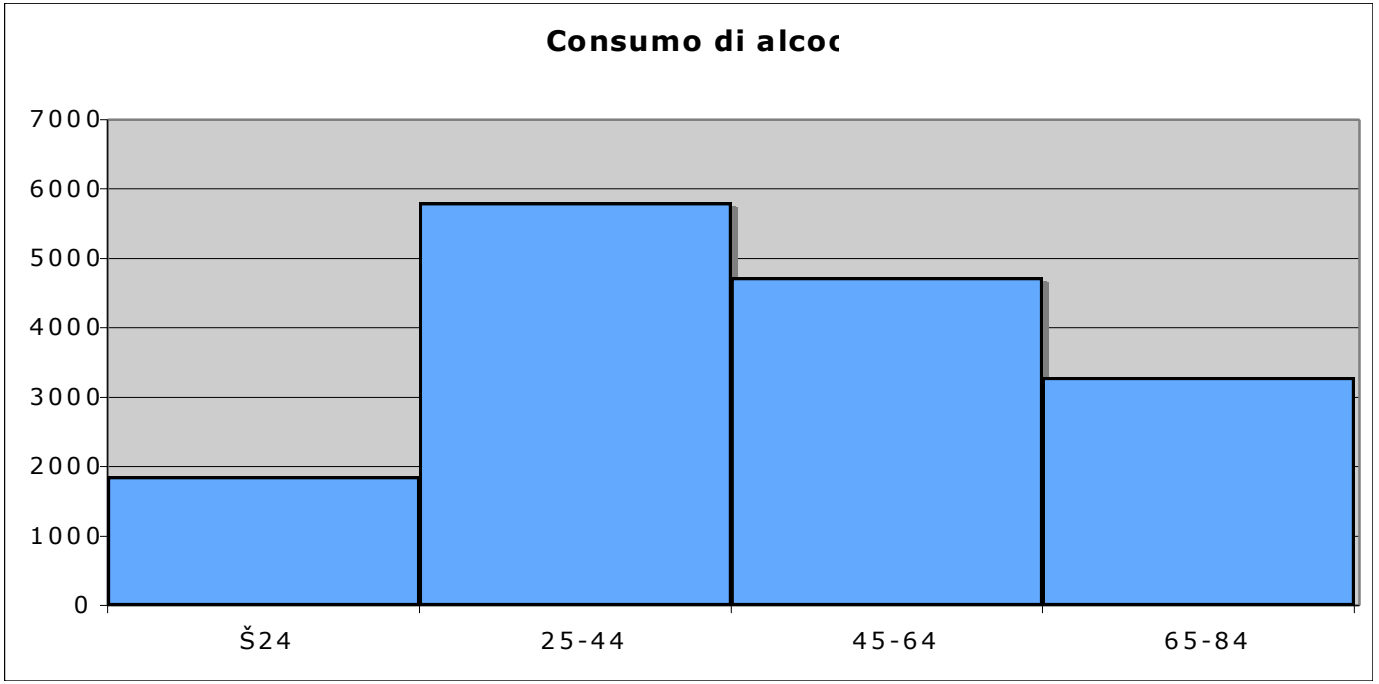


Diagramma relativo all'esempio 2:



Istogrammi.

- L'istogramma è costituito mediante rettangoli adiacenti, le cui basi sono gli intervalli che definiscono le classi, e le altezze rappresentano le frequenze relative, o le densità di frequenza, cioè
- $F_i = p_i$ / ampiezza classe
- Avendo le classi supposte di uguale ampiezza, si potrebbe considerare la frequenza assoluta f_i al posto della densità di frequenza F_i.
- L'istogramma relativo all'esempio 3 è:


Consumo di alcoc
7000
6000
5000
4000
3000
2000
1000
0
Š24
25-44
45-64
65-84

Indici Sintetici dei dati

Essi danno delle informazioni quantitative sull'ordine di grandezza delle osservazioni (misure di posizione), sulla variabilità delle osservazioni (misure di dispersione, misure di forma).

OSSERVAZIONE: nel caso di una distribuzione di frequenza con dati raggruppati $x_1, x_2, \ldots, x_k$ rappresentano i valori centrali delle classi, $f_1, f_2, \ldots, f_k$ le corrispondenti frequenze assolute e r è il numero di classi.

MEDIA

MEDIA ARITMETICA:

- per dati semplici:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

- per dati ponderati:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i = \frac{1}{n}\left(x_1 f_1 + x_2 f_2 + \ldots + x_k f_k\right)$$

OSSERVAZIONE: la media aritmetica sarà indicata con μ quando si riferisce alla popolazione, quando si riferisce al campione sarà indicata con $\bar{x}$

Nella Tabella 2 la media è:

$$\bar{x} = \frac{1}{2829}\left(0 \cdot 354 + 1 \cdot 414 + 2 \cdot 1130 + 3 \cdot 567 + 4 \cdot 246 + 5 \cdot 66 + 6 \cdot 33 + 7 \cdot 11 + 8 \cdot 6 + 9 \cdot 1 + 10 \cdot 1\right) = 2{,}13$$

MEDIANA

- E’ il valore che occupa la posizione centrale in un insieme ordinato di dati, essa permette di ripartire la distribuzione in due parti, in ciascuna delle quali cade il 50% delle osservazioni.
- Non è influenzata dai valori estremi.
- Si usa per attenuare l’effetto dei valori estremi molto alti o molto bassi
- Per calcolare la mediana bisogna ordinare i valori.
- Se il campione ha un numero dispari di valori la mediana è il valore che occupa la posizione centrale. Per esempio la mediana di 1, 4, 6, 7, 8 è 6. Se il campione ha un numero pari di valori, la mediana è il valor medio dei due valori che occupano la posizione centrale. Per esempio la mediana di 1, 4, 6, 7, 8, 9 è (6+7)/2=6.5.

Confronto tra media e mediana

Il numero di giorni di sopravvivenza per i primi sei pazienti che hanno avuto un trapianto di cuore a Stanford sono stati:
15, 3, 46, 623, 126, 64. Il valor medio è

$$\bar{x} = \frac{15+3+46+623+126+64}{6} = \frac{877}{6} = 146{,}2$$

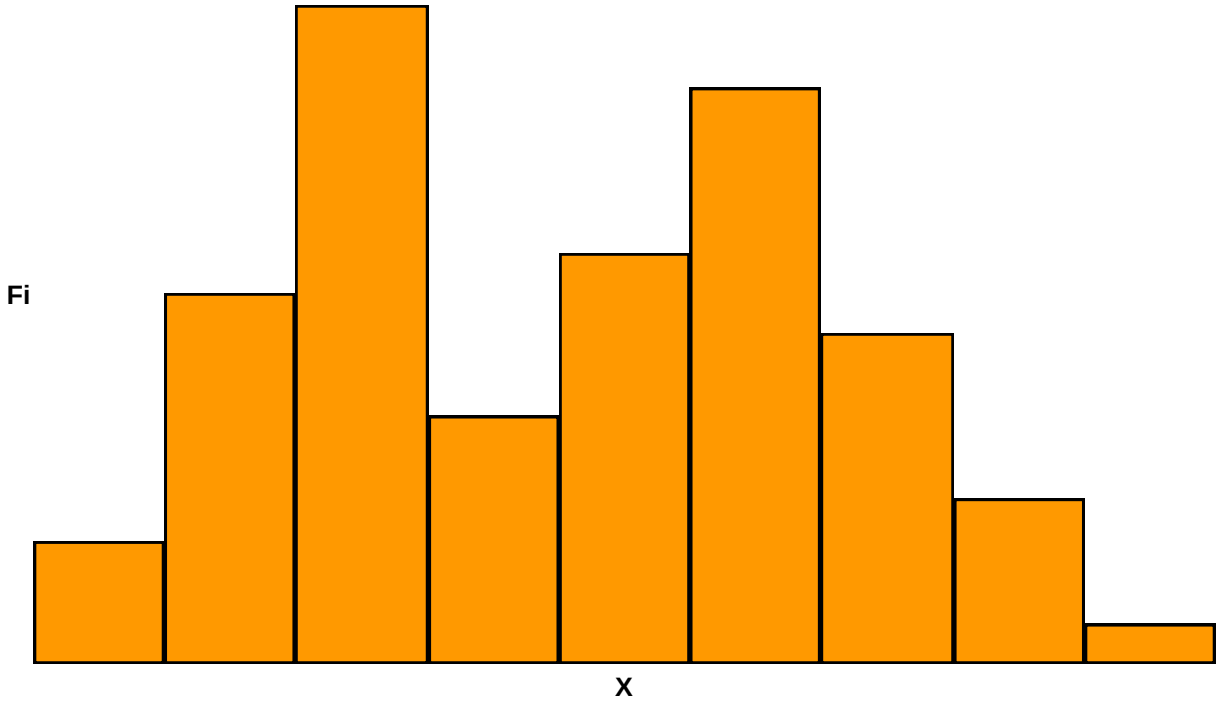
La mediana e' invece (46+64)/2 =55. Infatti i valori centrali della distribuzione ordinata sono 46 e 64 il cui valore
medio è 55:

3, 15, 46, 64, 123, 623

MODA

- E’ il valore più frequente di una distribuzione

Esempio di distribuzione bimodale



MISURE DI DISPERSIONE O VARIABILITA'

CAMPO DI VARIAZIONE

E' la differenza tra il valore massimo e il valore minimo

VARIANZA

Media dei quadrati degli scarti dei valori dalla loro media diviso il numero di osservazioni. Misura la dispersione dei dati rispetto alla media.

- per i dati semplici

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

- per i dati ponderati

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot f_i$$

Altro modo per calcolare la varianza

Per i dati semplici:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Per i dati ponderati:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot x_i^2 - \bar{x}^2$$

ESEMPIO: Calcolare la varianza nei due mo di descritti dei valori: 5, 6, 7, 7, 8, 10.

$$\bar{x} = \frac{5+6+7+7+8+10}{6} = \frac{43}{6} = 7.1\overline{6} \approx 7{,}17$$

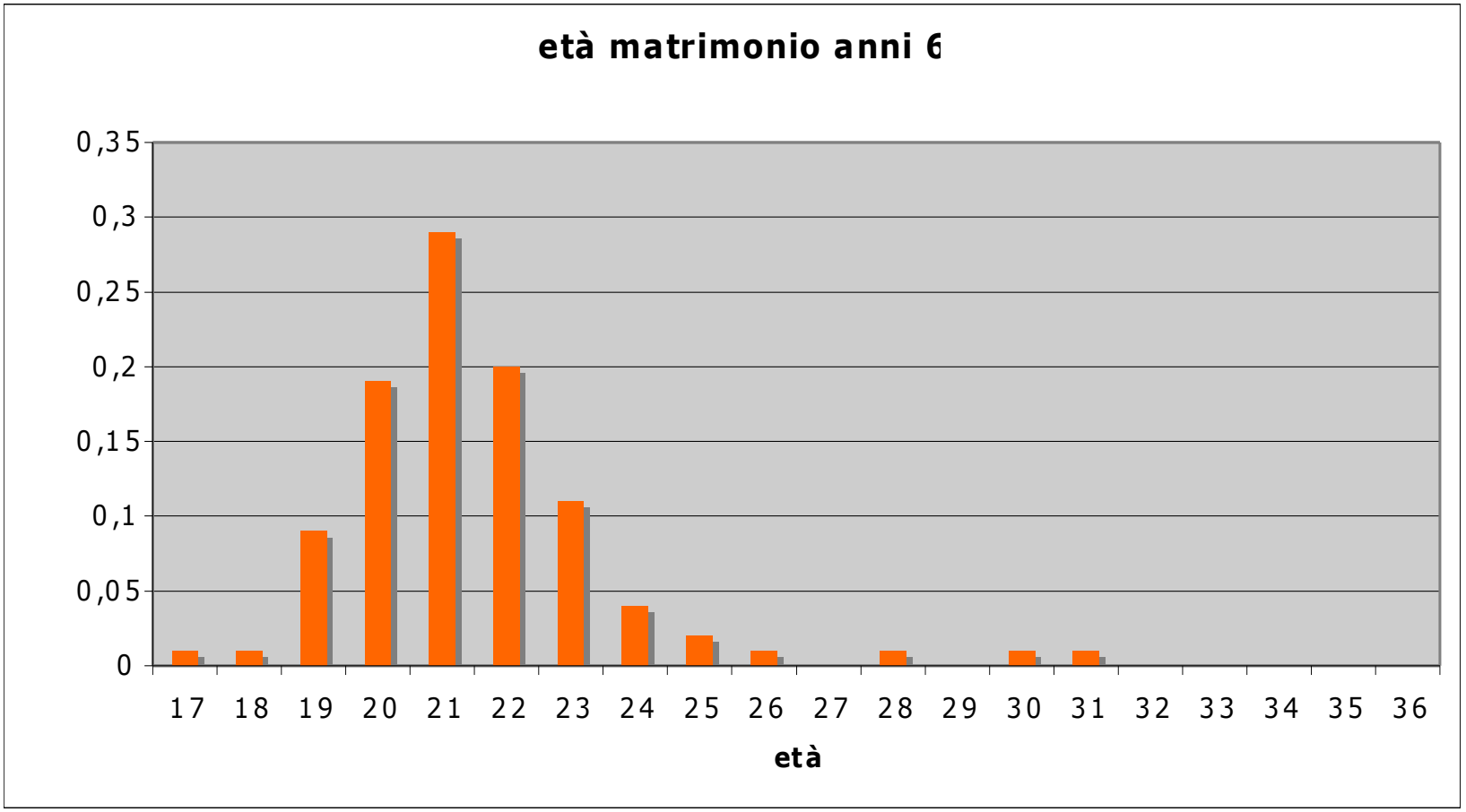
Allora si ha:

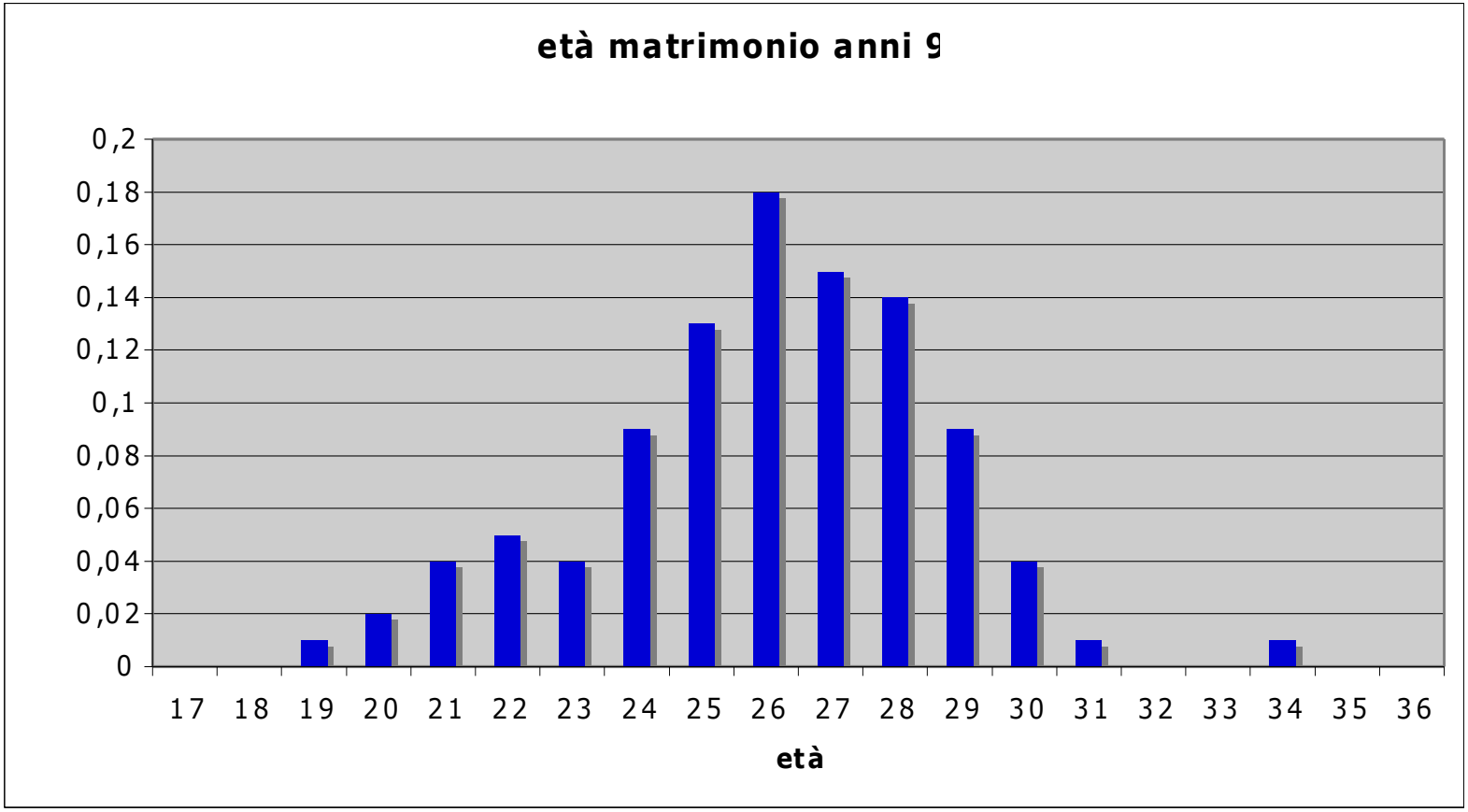
I METODO:

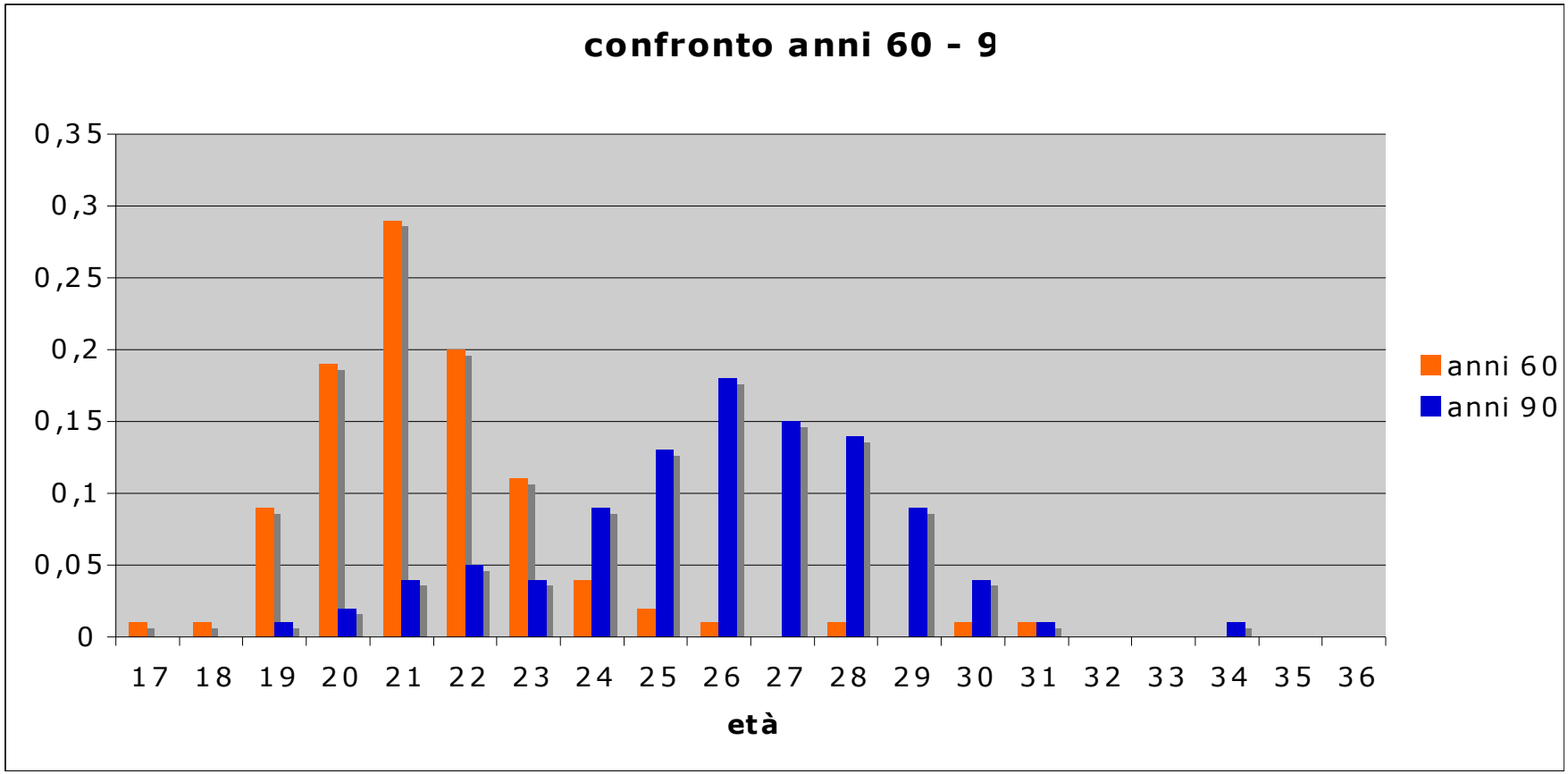
$$s^2 = \frac{1}{6}\left[(5-7.17)^2 + (6-7.17)^2 + (7-7.17)^2 + (7-7.17)^2 + (8-7.17)^2 + (10-7.17)^2\right] \approx 2{,}47$$

II METODO:

$$s^2 = \frac{1}{6}\left[5^2 + 6^2 + 7^2 + 7^2 + 8^2 + 10^2\right] - 7.17^2 \approx 2{,}42$$


età matrimonio anni 6
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
età


età matrimonio anni 9
0,2
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
età



Quale distribuzione ha varianza maggiore?
I valori medi sono rispettivamente 21,5 e 26.

DEVIAZIONE STANDARD (O SCARTO QUADRATICO MEDIO) DEL CAMPIONE

E' la radice quadrata della varianza

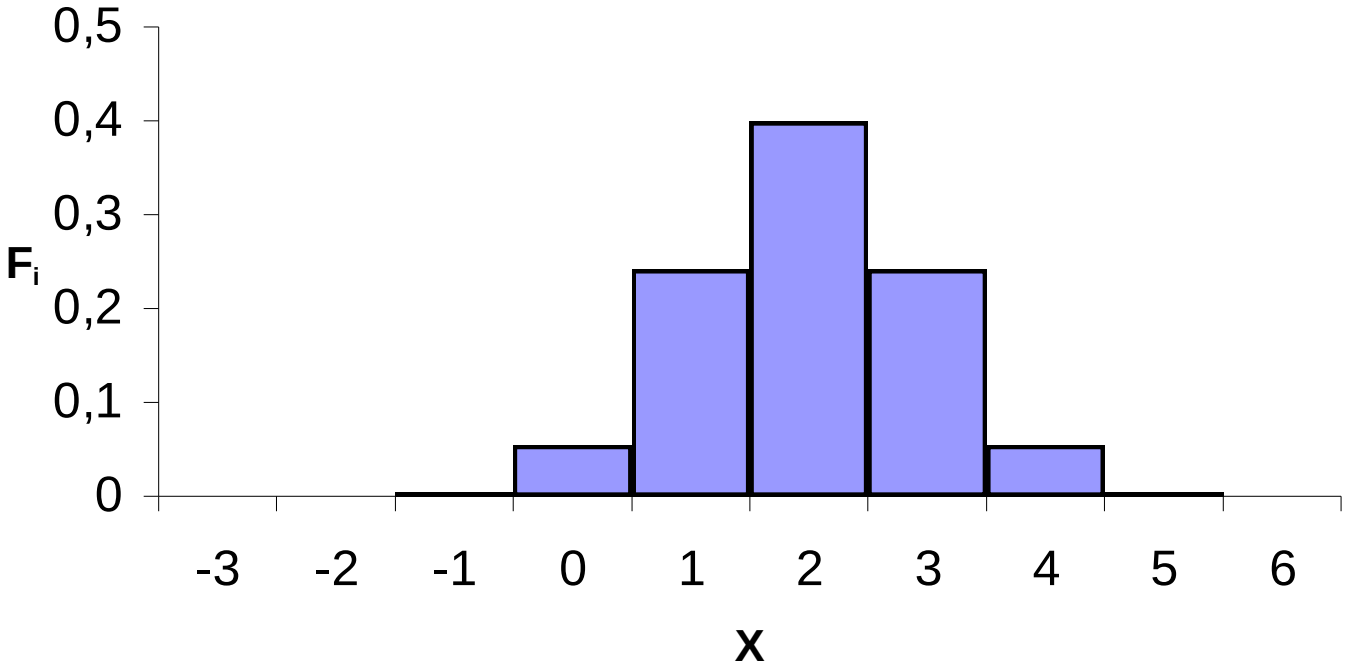
$$s = \sqrt{s^2}$$

Analogamente per la deviazione standard campionaria con la correzione di Student si ha

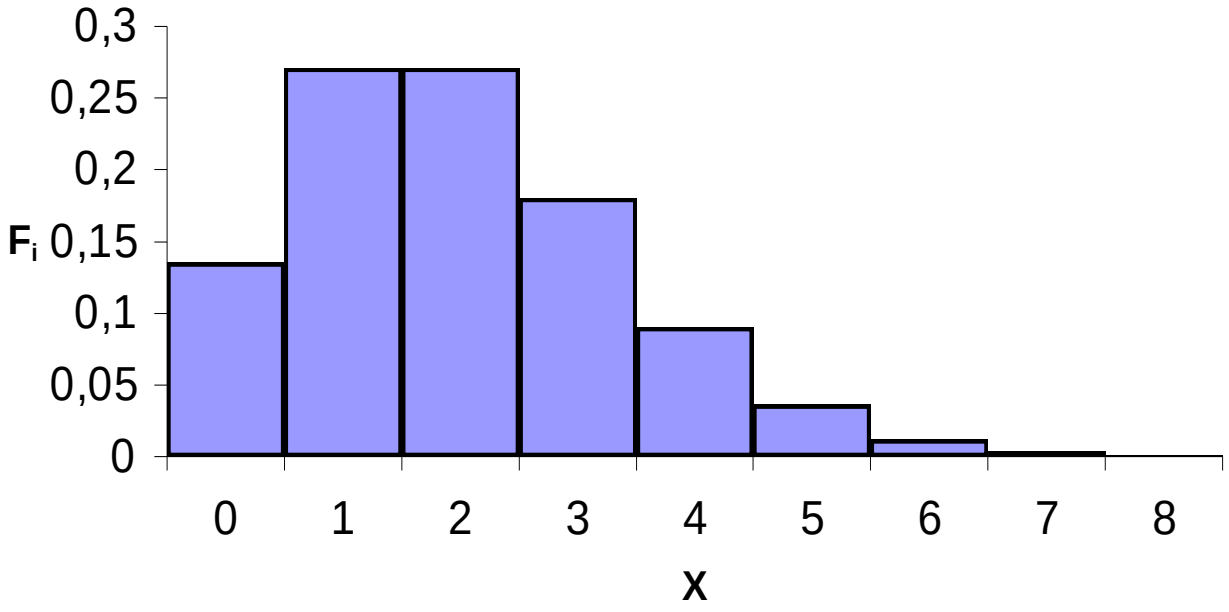
$$s' = \sqrt{s'^2}$$

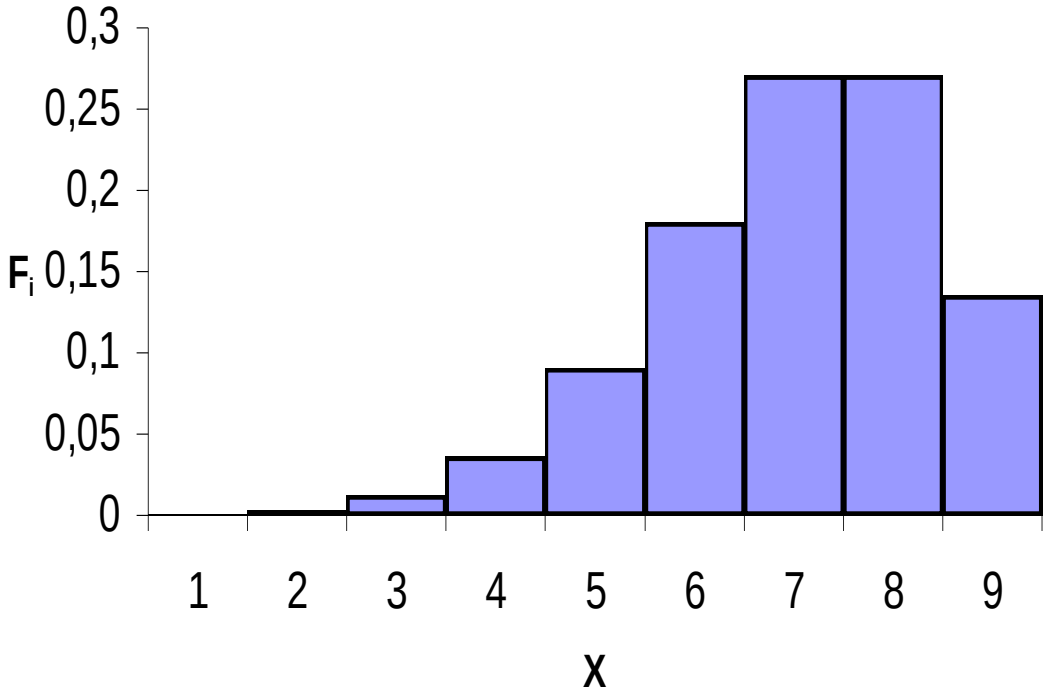
MISURE DI FORMA: Simmetria e asimmetria di una distribuzione

- SIMMETRIA: quando media e mediana coincidono;
- ASIMMETRIA POSITIVA quando la media è maggiore della moda, quindi si ha una coda più lunga a destra;
- ASIMMETRIA NEGATIVA quando la media è minore della moda, quindi si ha una coda più lunga a sinistra.



ESEMPIO DI DISTRIBUZIONE SIMMETRICA


0,3
0,25
0,2
Fi 0,15
0,1
0,05
0
0 1 2 3 4 5 6 7 8
X



NEGATIVA

INDICE DI ASIMMETRIA (Skewness)

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3$$

- >0 coda a destra (simmetrica positivamente)
- <0 coda a sinistra (simmetrica negativamente)
- =0 “quasi” simmetrica

Trasformazione dei dati

Supponiamo di voler confrontare due alunni in base ai voti ottenuti alla fine dell' esame sostenuto in 4 materie da due studenti A e B

Materia	Voto A	Voto B
Italiano	7	
Storia	7	7
Filosofia	6,5	6
Matematica		6,5
Fisica		6,6
Inglese	7,5	
	6,9	6,6

Standardizzare i dati.

$$Z_i = \frac{X_i - \bar{X}}{s}$$

dove

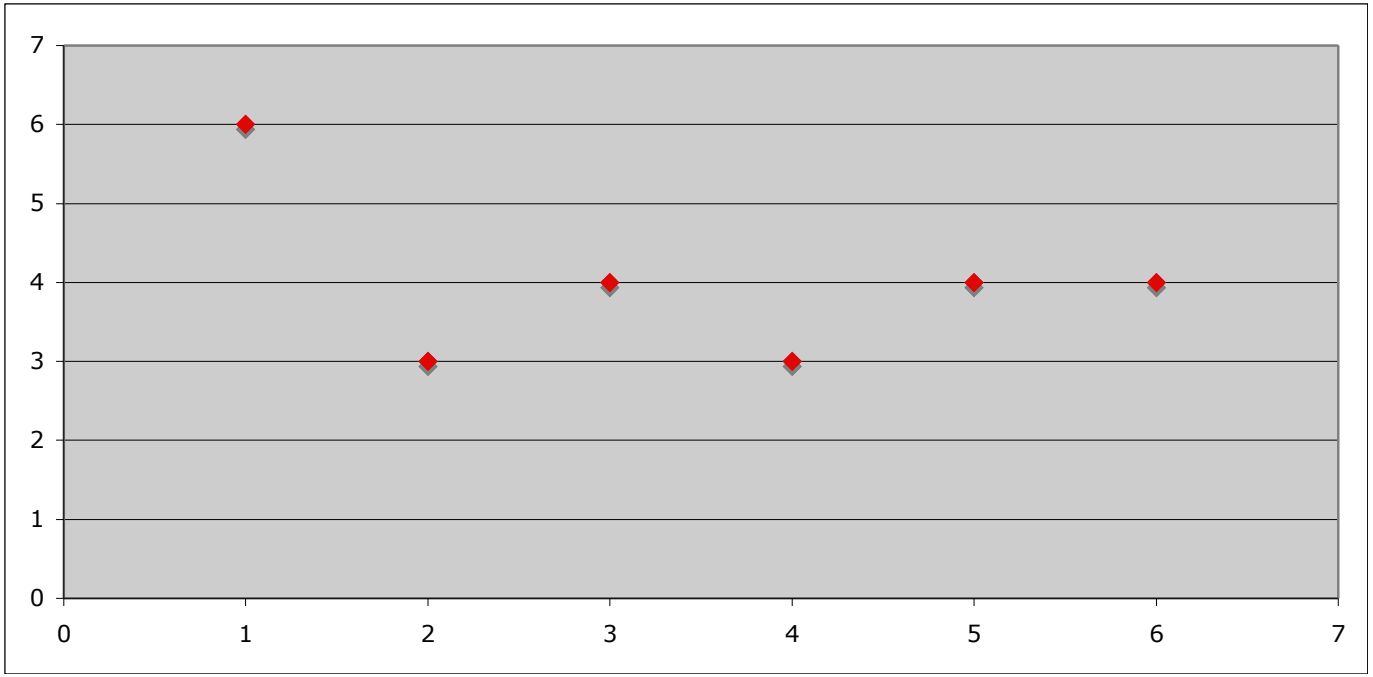
- X_i è la i-esima osservazione, i=1,2,…,n.
- è il valor medio di esse.
- s è lo scarto quadratico medio di esse.

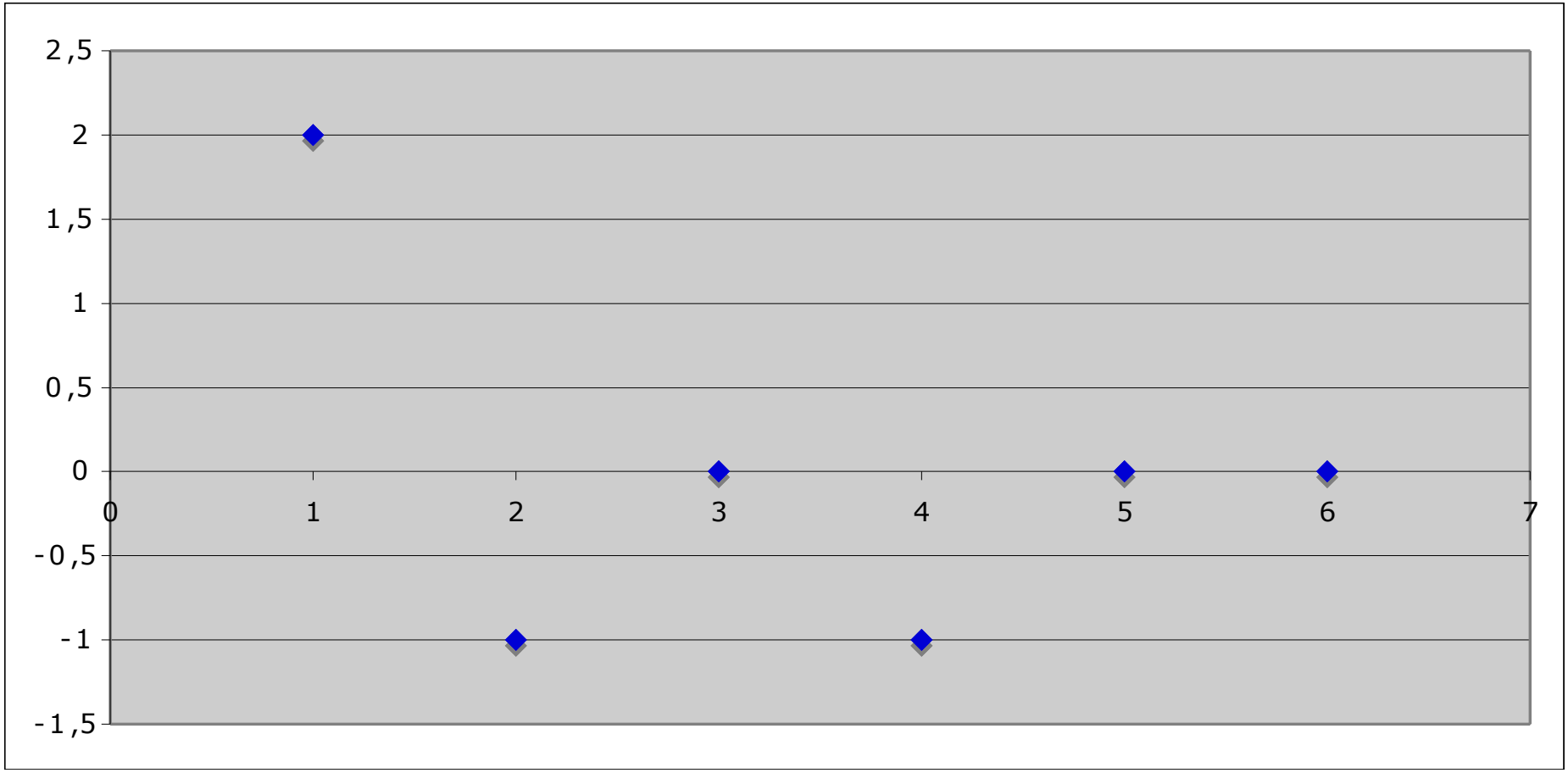
Supponiamo che 5 atlete dopo un allenamento abbiano perso ciascuna 5 chili:

Peso prima	Peso dopo
50	45
54	49
58	53
61	56
62	57
Media 57	52
Dev. Stand. 5	5

Osserviamo che l'operazione di sottrarre una stessa costante trasforma I dati in nuovi dati con media pari alla media originale meno la costante, lasciando inalterata lo scarto quadratico medio

Esempio. Dati I numeri 6,3,4,3,4,4, il loro valor medio è 4. Sottraendo la media otteniamo 2,-1,0,-1,0,0 che ha media nulla




2,5
2
1,5
1
0,5
0
-0,5
-1
-1,5
0
1
2
3
4
5
6
7

51,3=0,9*57

4,5=0,9*5

Consideriamo le 5 atlete che dopo lÕallenamento hanno perso un decimo del proprio peso:

Peso prima	Peso dopo
50	45
54	48,6
58	52,2
61	54,9
62	55,8
Media 57	51,3
Dev. Stand. 5	4,5

Supponiamo ora di moltiplicare ciascun dato per una stessa costante. In tal caso la media sarà moltiplicata per la stessa costante e lo stesso accade per il valor medio.

Esempio

Materia	Voto A	Voto B
Italiano	7	
Storia	7	7
Filosofia	6,5	6
Matematica		
Fisica		6,6
Inglese	7,5	
	7	6,5

Materia	media	scarto quadratico
Italiano	7	0,1
Storia	7	0,2
Filosofia	6	0,5
Matematica	6	0,9
Fisica	5,8	1,1
Inglese	7,8	0,84

Se ora standardizziamo I dati si ottiene e quindi la media dei voti standardizzati risulta essere maggiore per lo studente B, contrariamente alla media di ogni studente dei voti non standardizzati.

Materia	Voto norm A	Voto norm B
Italiano	0	
Storia	0	0
Filosofia	1	0
Matematica		0,556
Fisica		0,727
Inglese	-0,357	
	0,16	0,32

Quando il voto standardizzato viene 0, ciò vuol dire che lo studente sta sulla media, se viene 1 si trova ad una scato sopra la media e così via.

Excel: Analisi dei dati

Analisi dei dati: è un'applicazione di Excel che permette di fare un'analisi statistica dei dati forniti in input insieme a dei parametri, attraverso delle funzioni macro statistiche. I risultati sono visualizzati sottoforma di tabelle o grafici.

Come aprire l'applicazione: scegliere **Analisi dei dati** dal menu **Strumenti.** Se non compare selezionare **Analisi dei dati** dall'opzione **Componenti aggiuntive.**

Statistica Descrittiva

Scegliendo dalla finestra di dialogo l'opzione Statistica Descrittiva viene fatta una analisi statistica dei dati presi da una colonna di una tabella di Excel (vedi istruzioni per costruire un foglio elettronico).

Opzioni:

- Intervallo di input: dare le coordinate dei dati da analizzare, per esempio A1:A15, cioè tutti i dati contenuti nelle celle della colonna A dalla prima alla quindicesima riga.
- Intervallo di output: dare le coordinate della cella superiore sinistra della tabella di output.

Riepilogo statistiche: genera una tabella di output con le statistiche: Media, Errore standard (della media), Mediana, Moda, Deviazione Standard, Varianza, Curtosi, Asimmetria, Intervallo, Min, Max, Somma, Conteggio

Istogramma

L'opzione Istogramma raggruppa un insieme di dati in classi e calcola le frequenze assolute, e quelle cumulate.

Opzioni dalla finestra di dialogo:

- Intervallo di input: dare le coordinate dei dati da analizzare.
- Intervallo di classe (facoltativo): dare un intervallo di celle contenenti un insieme di valori limite che definiscano gli estremi destri di degli intervalli delle classi. Se non viene dato le classi vengono create automaticamente. Se i dati sono discreti ogni intervallo ha ampiezza 1.
- Intervallo di output: dare le coordinate della cella superiore sinistra della tabella di output.

- Etichetta: tale opzione viene scelta quando la prima riga della colonna data è stata utilizzata per un titolo della colonna.
- Percentuale cumulata: fornisce il grafico della frequenza relativa cumulata.
- Grafico: fornisce il grafico delle frequenze.