Statistica (5 crediti)

Alessandro Lubisco

Informazioni generali

- Orario delle lezioni (Via S. Giacomo, 12)
  - Giovedì 14-16
  - Venerdì 9-11
- Orario di ricevimento
  - Giovedì 10-12

    4° piano
    c/o Dip.to di Scienze statistiche
    V. Belle Arti, 41 – Bologna

    mail: alessandro.lubisco@unibo.it

    web: www2.stat.unibo.it/lubisco

Indicazioni generali sul corso

- Obiettivi
  - Introduzione ai concetti e alle metodologie di base della statistica
- Requisiti richiesti
  - Nessuno (a parte una calcolatrice…)
- Ore di lezione: 42
- Obbligo di frequenza

Indicazioni generali sul corso

- Modalità di svolgimento dell'esame
  - Prova scritta (obbligatoria)
  - Prova orale (facoltativa)

- Libro di testo

  Montanari A., Agati P., Calò D.G.

  Statistica

  Masson, Milano, 1998
  ISBN 88-214-0092-1

Cosa NON è la Statistica

- Prima di tutto, lo studioso di Statistica è uno Statistico e non uno Statista
- Statistica e Statistiche non sono la stessa cosa
- La Statistica non è ciò che dice: "Se tu hai un pollo e io non ne ho, in media abbiamo mezzo pollo a testa"

A cosa serve la Statistica

- Strumento essenziale per la scoperta di leggi e relazioni tra fenomeni
- Interviene in tutte le situazioni nelle quali occorre assumere decisioni in condizioni di incertezza
  - Ricerca scientifica
  - Pianificazione economica
  - Azione politica
  - …

A cosa serve la Statistica

- La Statistica analizza in termini quantitativi i fenomeni collettivi, cioè fenomeni il cui studio richiede l'osservazione di un insieme di manifestazioni individuali

  Analizzare in termini quantitativi significa basarsi su dati e non su idee o ipotesi

Fenomeno collettivo

- Consumo di un determinato bene in un periodo prefissato
- Reddito di un insieme di individui
- Peso di un gruppo di oggetti o persone
- ...

Un fenomeno può essere individuato da una pluralità di caratteri

Caratteri

Se il fenomeno che stiamo analizzando è il curriculum vitae degli studenti, esempi di carattere sono:

- tipo di maturità
- voto di maturità
- anno di conseguimento di maturità
- età (o data di nascita)
- sesso
- sport praticati
- …

Caratteri e Modalità

Nome	Età	Sesso	Maturità	Punt.	Anno	Sport
Rossi A.	21	F	Classica	95	2003	Ritmica
Bianchi D.	19	M	Scientifica	88	2005	Calcio
Verdi G.	24	F	Sociale	84	2000	Nuoto
Gialli S.	22	F	Linguistica	96	2002	Atletica
Neri M.	25	M	Scientifica	98	2003	Pallanuoto

- A ogni riga corrisponde un individuo del quale sono stati rilevati alcuni **caratteri**
- In corrispondenza di ogni individuo, ciascun carattere assume una determinata **modalità**

Domanda …

Carattere rilevato: colore degli occhi

Quali sono le modalità che ci si aspetta di osservare?

- Nero
- Marrone
- Verde
- Azzurro

Analisi statistica

Numero di caratteri oggetto di studio	Analisi statistica
1	Univariata
2	Bivariata
Più di 2	Multivariata

Collettivo in esame

Il collettivo su cui si è osservato il fenomeno può essere:

- Popolazione di interesse
- Sottoinsieme della popolazione di interesse: il campione

Collettivo statistico o Popolazione

Una popolazione è un qualsiasi insieme di elementi, reale o ipotetico, presente o futuro, che forma oggetto di uno studio statistico.

Una popolazione non costituisce necessariamente un insieme "biologico".

Collettivo statistico o Popolazione

Esempi di collettivo statistico:

- Popolazione residente a Bologna
- Esercizi commerciali di Rimini

Collettivi di stato

- Automobili vendute in Italia nel 2005
- Lampadine prodotte nell'ultimo mese

Collettivi di movimento

Popolazione di interesse

Per ora non operiamo la distinzione tra popolazione finita e popolazione infinita.

Se osserviamo l'intera popolazione di interesse l'obiettivo è:

- Descrivere le caratteristiche della popolazione

Analisi statistica descrittiva

Popolazione di interesse

Se osserviamo un sottoinsieme della popolazione di interesse, cioè un campione, l'obiettivo è:

- Sfruttare le caratteristiche osservate nel sottoinsieme per cercare di conoscere le caratteristiche della popolazione

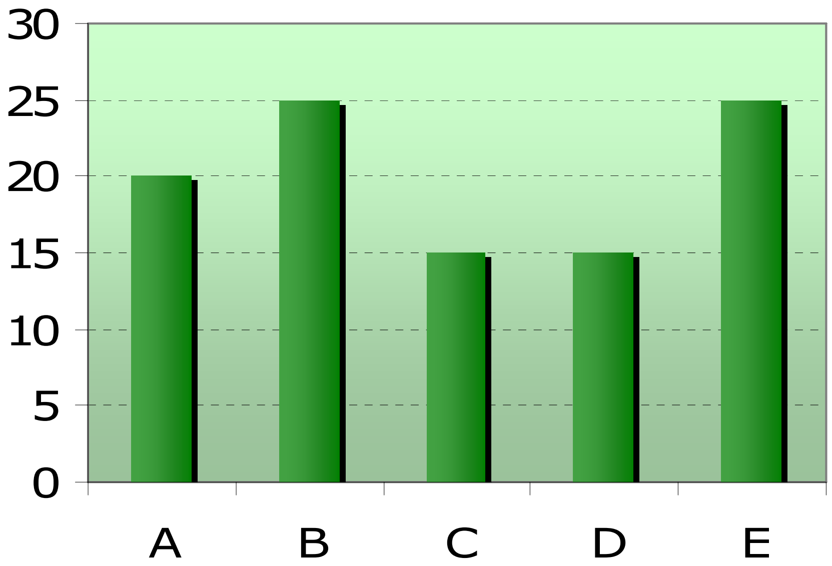
Analisi statistica inferenziale

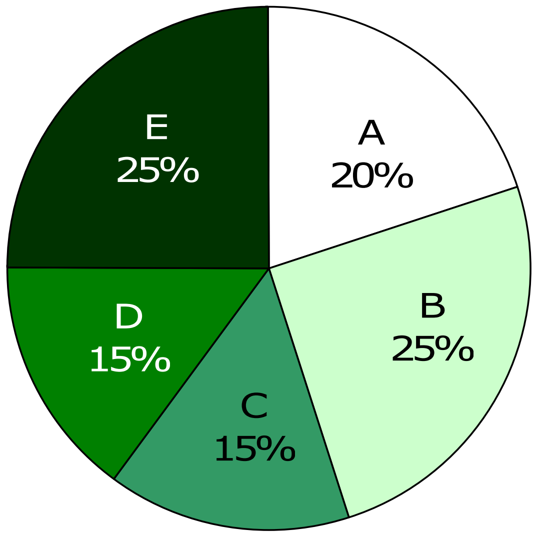
Analisi statistica descrittiva

- Gli strumenti della statistica descrittiva servono a sintetizzare i dati raccolti
  - Strumenti descrittivi grafici
  - Strumenti descrittivi numerici

Strumenti descrittivi grafici

- Diagrammi a barre
- Istogrammi
- Diagrammi a torta





Strumenti descrittivi numerici

- Media
- Mediana
- Moda
- Valore minimo
- Valore massimo

→ Statistiche

Analisi statistica

Collettivo osservato	Analisi statistica
Popolazione di interesse	Descrittiva 1
Campione	Inferenziale 2

Concetti introdotti

- Fenomeno collettivo
- Carattere
- Modalità
- Collettivo statistico o Popolazione
  - Collettivi di stato
  - Collettivi di movimento
- Campione

Unità statistica

Nome	Età	Sesso	Maturità	Punt.	Anno	Sport
Rossi A.	21	F	Classica	95	2003	Ritmica
Bianchi D.	19	M	Scientifica	88	2005	Calcio
Verdi G.	24	F	Sociale	84	2000	Nuoto
Gialli S.	22	F	Linguistica	96	2002	Atletica
Neri M.	25	M	Scientifica	98	2003	Pallanuoto

Si definiscono unità statistiche gli oggetti descritti tramite un insieme di dati.
Le unità statistiche possono essere persone, ma anche animali o cose.

Domanda …

Nome	Età	Sesso	Maturità	Punt.	Anno	Sport
Rossi A.	21	F	Classica	95	2003	Ritmica
Bianchi D.	19	M	Scientifica	88	2005	Calcio
Verdi G.	24	F	Sociale	84	2000	Nuoto
Gialli S.	22	F	Linguistica	96	2002	Atletica
Neri M.	25	M	Scientifica	98	2003	Pallanuoto

Quante sono le unità statistiche di questo collettivo?

E quanti sono i caratteri rilevati?

Protocollo elementare

- **Protocollo elementare**: è l'insieme dei valori assunti dal carattere oggetto di indagine nelle unità statistiche del collettivo in esame.

Collettivo in esame: 88 individui iscritti al corso di Statistica

Carattere osservato: Voto conseguito all'esame di Statistica

{29, 29, 24, 20, 22, 28, 19, 19, 21, 26, 20, 24, 21, 19, 25, 25, 23, 28, 22, 29, 26, 23, 28, 30, 20, 27, 22, 27, 20, 24, 25, 18, 26, 29, 29, 23, 23, 24, 22, 25, 27, 26, 23, 18, 19, 26, 22, 25, 20, 26, 22, 24, 20, 22, 21, 29, 30, 19, 24, 24, 26, 26, 29, 30, 29, 25, 28, 26, 22, 27, 27, 29, 26, 26, 22, 27, 24, 29, 30, 20, 24, 24, 21, 18, 22, 28, 23, 21}

Definizione di Carattere

- Carattere: è un aspetto del fenomeno oggetto di studio, rilevato o misurato sulle unità statistiche.

E' tipico dei fenomeni reali di interesse statistico che i caratteri assumano modalità differenti nelle varie unità statistiche

Definizione di Modalità

- Modalità: è l'espressione concreta del carattere nelle unità statistiche, cioè il valore che l'unità statistica manifesta.

L'elenco di tutte le possibili modalità di un carattere si dice esaustivo se è completo.

Le modalità si dicono disgiunte se una unità statistica può manifestare il carattere in una ed una sola modalità tra quelle indicate.

Esempio

Elenco di modalità per il carattere “tinta degli occhi”

- Nera
- Marrone
- Verde
- Azzurra
- Chiara
- Scura

Questo elenco è esaustivo perché comprende tutti i possibili colori.

Tuttavia le modalità non sono disgiunte in quanto una unità statistica può assumere sia la modalità Nera che la modalità Scura.

Elenchi di modalità disgiunte

Gruppi sanguigni	Gare nuoto Stile libero	Voti negli esami
0	50m	18
A	100m	19
B	200m	…
AB	400m	29
	800m	30
	1500m	30L

Tipologie di Carattere

- **Caratteri qualitativi**
  - Colore degli occhi
  - Genere
  - Categoria sportiva
- **Caratteri quantitativi**
  - Reddito
  - Altezza
  - Numero di esami sostenuti

Caratteri qualitativi

I dati qualitativi sono la forma più semplice di dati.

Il carattere è detto qualitativo se non assume valori numerici, ma ammette gradi o attributi distinti.

Carattere qualitativo → Mutabile statistica

Carattere qualitativo

- Ordinabile: tra i gradi è possibile stabilire una relazione d'ordine
  - Ordinabile rettilineare (categoria sportiva, titolo di studio)
  - Ordinabile ciclico (mese, stagione)
- Sconnesso: non esiste un ordinamento degli attributi del carattere
  - (nazione di nascita, laurea conseguita, colore degli occhi, sport praticato)

Caratteri quantitativi

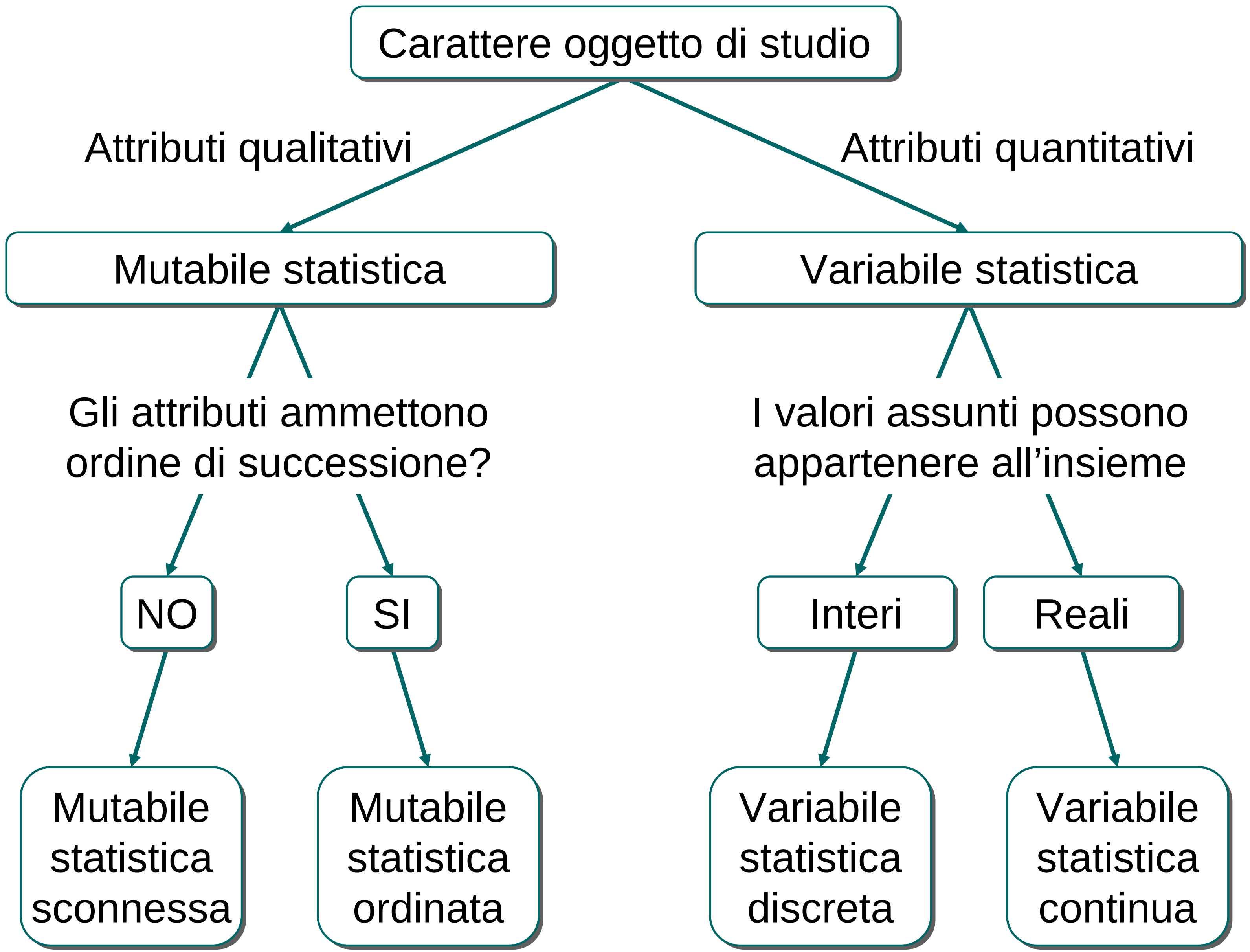
I dati quantitativi sono intrinsecamente numerici e per essi hanno senso operazioni come la somma o la media.

Quindi, un carattere si dice quantitativo se assume valori numerici.

Carattere quantitativo → Variabile statistica

Carattere quantitativo

- Discreto o enumerabile: può assumere solo valori interi
  - (Numero di componenti la famiglia, numero di dipendenti di un'azienda)
- Continuo o misurabile: può assumere tutti i valori di un intervallo
  - (statura, temperatura, tempo di percorrenza di una distanza)


Carattere oggetto di studio
Attributi qualitativi
Attributi quantitativi
Mutabile statistica
Variabile statistica
Gli attributi ammettono ordine di successione?
I valori assunti possono appartenere all'insieme
NO
SI
Interi
Reali
Mutabile statistica sconnessa
Mutabile statistica ordinata
Variabile statistica discreta
Variabile statistica continua

Simbologia

- **Mutabili**: si indicano con le prime lettere dell’alfabeto (A, B, …)
- **Variabili**: si indicano con le ultime lettere dell’alfabeto (X, Y, …)
- Con le corrispondenti lettere minuscole si indicano le loro determinazioni *a*, *b*, *x*, *y* in una unità statistica

Simbologia

- L'insieme degli attributi o degli stati di grandezza di un carattere rilevati su un insieme di n unità statistiche è così indicato:

Per la mutabile A: $\{a_1, a_2, \ldots, a_j, \ldots, a_n\}$

Per la variabile X: $\{x_1, x_2, \ldots, x_j, \ldots, x_n\}$

Domanda …

Collettivo in esame → 88 individui iscritti al corso di Statistica

Carattere osservato → Voto conseguito all'esame di Statistica

Protocollo elementare

{29, 29, 24, 20, 22, 28, 19, 19, 21, 26, 20, 24, 21, 19, 25, 25, 23, 28, 22, 29, 26, 23, 28, 30, 20, 27, 22, 27, 20, 24, 25, 18, 26, 29, 29, 23, 23, 24, 22, 25, 27, 26, 23, 18, 19, 26, 22, 25, 20, 26, 22, 24, 20, 22, 21, 29, 30, 19, 24, 24, 26, 26, 29, 30, 29, 25, 28, 26, 22, 27, 27, 29, 26, 26, 22, 27, 24, 29, 30, 20, 24, 24, 21, 18, 22, 28, 23, 21}

Che tipo di carattere stiamo osservando?

Che simbologia si adotta?

Risposta

Collettivo in esame → 88 individui iscritti al corso di Statistica

Carattere osservato → Voto conseguito all’esame di Statistica

Protocollo elementare {29, 29, 24, 20, 22, ...}

Il carattere è di tipo quantitativo e i possibili valori sono interi, quindi esso descrive una variabile statistica discreta

n=88

X=Voto conseguito all’esame di Statistica

Distribuzione di frequenza

- La distribuzione di frequenza è il primo passo dell'elaborazione statistica
- Non è altro che una classificazione delle n unità statistiche in k classi (dove k≤n) formate sulla base delle modalità del carattere osservato nel collettivo

Distribuzione di frequenza

- Nella distribuzione di frequenza sono raggruppate nella medesima classe tutte le unità statistiche che hanno la medesima modalità del carattere considerato
- Ogni classe della distribuzione è definita da una coppia di elementi
  - Modalità del carattere
  - Corrispondente frequenza

Esempio di distribuzione di frequenza

Modalità → ← Frequenze

Voti	Studenti
18	3
19	5
20	7
21	5
22	10
23	6
24	10
25	6
26	11
27	6
28	5
29	10
30	4
Totale	88

Esempio di protocollo elementare

Collettivo in esame → 38 clienti di cartomante

Carattere osservato → A=Segno zodiacale ?

Protocollo elementare

{Scorpione, Cancro, Leone, Toro, Sagittario, Capricorno, Scorpione, Vergine, Ariete, Cancro, Pesci, Capricorno, Sagittario, Vergine, Pesci, Toro, Leone, Capricorno, Ariete, Ariete, Toro, Pesci, Acquario, Capricorno, Bilancia, Ariete, Leone, Gemelli, Sagittario, Scorpione, Vergine, Toro, Acquario, Acquario, Cancro, Vergine, Scorpione, Vergine}

Esempio di distribuzione di frequenza

Segno zodiacale	Numero di clienti
Ariete	4
Toro	4
Gemelli	1
Cancro	3
Leone	3
Vergine	5
Bilancia	1
Scorpione	4
Sagittario	3
Capricorno	4
Acquario	3
Pesci	3
Totale	38

Esempio di protocollo elementare

Collettivo in esame → 38 clienti di un'altra cartomante

Carattere osservato → A=Segno zodiacale

Protocollo elementare

{Scorpione, Sagittario, Leone, Toro, Sagittario, Capricorno, Scorpione, Vergine, Ariete, Sagittario, Pesci, Capricorno, Sagittario, Vergine, Ariete, Toro, Leone, Capricorno, Ariete, Ariete, Toro, Pesci, Sagittario, Capricorno, Scorpione, Ariete, Leone, Sagittario, Sagittario, Scorpione, Vergine, Toro, Ariete, Ariete, Scorpione, Vergine, Scorpione, Vergine}

Esempio di distribuzione di frequenza

Segno zodiacale	Numero di clienti
Ariete	7
Toro	4
Leone	3
Vergine	5
Scorpione	6
Sagittario	7
Capricorno	4
Pesci	2
Totale	38

Esempio di protocollo elementare

Collettivo in esame → 98 alunni maschi di 11 anni di età

Carattere osservato → X=tempo in secondi per percorrere 30m piani

Protocollo elementare

{5,16	5,32	5,62	5,81	5,20	6,06	5,78	5,44	5,83	5,65
5,35	5,54	6,41	5,15	5,25	5,82	5,13	4,91	5,52	5,33
5,20	5,65	6,34	5,99	5,13	5,99	5,65	5,48	5,45	6,46
6,24	5,15	5,71	5,78	5,51	5,45	5,65	5,18	6,41	5,41
5,50	5,37	6,13	5,37	6,06	5,46	5,12	5,68	5,25	5,17
5,44	5,64	5,12	5,75	6,04	5,14	6,39	5,94	5,63	5,08
5,92	5,67	5,24	4,93	6,01	5,90	6,44	6,46	6,13	5,61
6,08	5,55	5,03	5,86	5,47	4,97	6,13	5,57	4,92	5,51
5,43	5,11	5,59	5,30	6,38	5,23	5,19	6,19	6,28	5,52
5,19	4,95	5,31	5,54	5,79	4,93	5,17	5,75}		

Esempio di distribuzione di frequenza

Tempo (in secondi)	Numero di alunni
4,90 |— 5,10	8
5,10 |— 5,30	20
5,30 |— 5,50	16
5,50 |— 5,70	20
5,70 |— 5,90	10
5,90 |— 6,10	10
6,10 |— 6,30	6
6,30 o più	8
Totale	98

Esempio di distribuzione di frequenza

Numero di abitanti	Numero di comuni
0 —| 1000	1956
1000 —| 2000	1706
2000 —| 5000	2224
5000 —| 10000	1164
10000 —| 20000	589
20000 —| 50000	324
50000 —| 100000	90
100000 —| 250000	34
250000 —| 500000	6
Oltre 500000	6
Totale	8102

Requisiti di una distribuzione di frequenza

Nella scelta del metodo classificatorio si devono tener presente:

- Requisito dell'esaustività: Ogni unità statistica deve appartenere a una delle classi, cioè deve poter essere classificata

- Requisito della disgiuntività: ogni unità statistica non può appartenere contemporaneamente a due classi distinte

Simbologia

Mutabile

Modalità di A	Frequenze
a_1	n_1
a_i	n_i
a_k	n_k
	n

In sintesi

$\{a_i; n_i\}$ $i=1,\ldots,k$

Simbologia

Variabile discreta

Modalità di X	Frequenze
x_1	n_1
x_i	n_i
x_k	n_k
	n

In sintesi

$\{x_i; n_i\}$ $i=1,\dots,k$

Simbologia

Variabile continua

Modalità di X	Frequenze
x_0-x_1	n_1
x_{i-1}-x_i	n_i
x_{k-1}-x_k	n_k
	n

In sintesi

$\{x_{i-1}\text{-}x_i;\ n_i\}\ i=1,\ldots,k$

Esercizio

In una distribuzione statistica (secondo l'età in anni compiuti), si viola il requisito dell'esaustività se si ricorre alla seguente classificazione?

0 |-| 6
7 |-| 14
15 |-| 18
19 |-| 29
30 |-| 45
46 |-| 65
66 e oltre

Ampiezza di un intervallo

L'intervallo x_{i-1}|-x_i, come l'intervallo x_{i-1}-|x_i ha ampiezza w_i uguale alla differenze tra i suoi estremi

$$\boxed{w_i = x_i - x_{i-1}}$$

Esercizio

Determinare l'ampiezza degli intervalli di questa classificazione

0	|-|	6
7	|-|	14
15	|-|	18
19	|-|	29
30	|-|	45
46	|-|	65
66 e oltre		

0	|-	7	7
7	|-	15	8
15	|-	19	4
19	|-	30	11
30	|-	46	16
46	|-	66	20
66 e oltre			

$$\boxed{w_i = x_i - x_{i-1}}$$

Valore centrale di un intervallo

Il valore centrale $\hat{x}_i$ dell'intervallo di estremi x_{i-1}, x_i è dato dalla semisomma degli estremi stessi

$$\boxed{\hat{x}_i = \frac{x_{i-1} + x_i}{2}}$$

Esercizio

Determinare il valore centrale degli intervalli di questa classificazione

0	|-	7	⟵	3,5
7	|-	15	⟵	11
15	|-	19	⟵	17
19	|-	30	⟵	24,5
30	|-	46	⟵	38
46	|-	66	⟵	56
66 e oltre				

$$\boxed{\hat{x}_i = \frac{x_{i-1} + x_i}{2}}$$

Quantità complessiva del carattere

La quantità complessiva del carattere quantitativo X portato dalle n_i unità della i-esima classe è pari a

$$\boxed{c_i = x_i n_i}$$

E' una quantità che si incontrerà parlando di valori medi, ma in alcuni casi ha anche un significato fenomenico

Esempio

Numero di componenti x_i	Numero di famiglie n_i	Popolaz. resid. in famiglie $c_i=x_in_i$
1	4106	
2	4857	
3	4666	
4	4648	
5	1542	
6	368	
7	78	
8 o più	40	
Totale	20305	

Quantità complessiva del carattere

In altri tipi di distribuzione, invece, c_i non ha alcun significato.

X=età in anni compiuti

$\{x_{i-1}-x_i;\ n_i\}$ $i=1,\ldots,k$ è la distribuzione per classi corrispondente

In tal caso c_i corrisponde al numero di anni vissuti dagli n_i individui di età compresa tra x_{i-1} e x_i anni

Frequenza relative

Date n unità statistiche di cui n_i presentano la i-esima modalità, si definisce frequenza relativa i-esima il rapporto

$$\boxed{f_i = n_i / n}$$

La frequenza percentuale è il prodotto

$$\boxed{p_i = f_i \cdot 100}$$

Simbologia

X	Frequenza assoluta	Frequenza relativa	Frequenza relativa percentuale
x_1	n_1	f_1	p_1
x_2	n_2	f_2	p_2
...	...	...	...
x_i	n_i	f_i	p_i
...	...	...	...
x_k	n_k	f_k	p_k
Totale	n	1	100

$$\boxed{\sum_{i=1}^{k} f_i = 1}$$

$$\boxed{\sum_{i=1}^{k} p_i = 100}$$

Esempio sulle frequenze relative

Le tabelle riportano dati relativi a persone di 6 anni o più che praticano sport con continuità (in migliaia) classificate in base al sesso, nel 1988 e nel 1997

Anno 1988

Sesso	n_i
M	8245
F	3962
Totale	12207

Anno 1997

Sesso	n_i
M	6071
F	3664
Totale	9735

Confronto tramite frequenze relative

Anno 1988

Sesso	f_i	$p_i = f_i\ 100$
M	0,675	67,5
F	0,325	32,5
Totale	1,000	100,00

Anno 1997

Sesso	f_i	$p_i = f_i\ 100$
M	0,624	62,4
F	0,376	37,6
Totale	1,000	100,0

Osservazioni sulle frequenze relative

- La frequenza relativa indica il peso di una classe sul totale delle unità statistiche
- Le frequenze relative (e percentuali) permettono i confronti tra distribuzioni del medesimo carattere riferite a collettivi di diversa numerosità

Osservazioni sulle frequenze relative

- La distribuzione delle frequenze relative (e percentuali) non indica il numero n di unità del collettivo cui essa è riferita

- Da una distribuzione di frequenza è sempre possibile ricavare la distribuzione delle frequenze relative (e percentuali), qualunque sia il carattere considerato

Osservazioni sulle frequenze relative

- Dalla distribuzione delle frequenze relative posso ricavare la distribuzione di frequenza solo se conosco il valore di n

- Il numero di cifre decimali da considerare quando si calcolano le frequenze relative deve essere:
  - il minimo possibile
  - tale che ogni cifra dia informazioni significative (inutili 10 decimali)

Quadratura dei risultati

a_i	n_i	f_i	
a_1	580	0,21	
a_2	959	0,35	
a_3	1236	~~0,45~~	0,44
Totale	2775	$\boxed{1,01}$	

L'arrotondamento a due cifre fa in modo che la somma degli f_i non sia uguale a zero.

Per far quadrare le frequenze relative, in questo caso, riduco la più grande perché così commetto il minore errore relativo

Esercizio sulle frequenze relative

Superficie abitabile	Frequenze percentuali	Frequenze assolute
95 —| 125	11	
125 —| 155	28	$n_i=p_i/100 \cdot n=f_i \cdot n$
155 —| 185	42	
185 —| 250	19	
Totale	100	1228

Sapendo che il collettivo considerato è composto da 1228 appartamenti, ricavare la distribuzione delle frequenze assolute.

Vero o Falso?

- La frequenza assoluta indica il numero di unità di un collettivo su cui è stato osservato un certo carattere

  FALSO

- La distribuzione di frequenza mantiene la stessa informazione del protocollo elementare

  FALSO

Vero o Falso?

- Per confrontare la distribuzione di un carattere su due collettivi statistici di numerosità diversa è opportuno confrontare le frequenza relative o percentuali

VERO

- Frequenze relative o percentuali forniscono la stessa informazione

VERO

Problema

X	n_i
18	3
19	5
20	7
21	5
22	10
23	6
24	10
25	6
26	11
27	6
28	5
29	10
30	4
Totale	88

X=Voti in Statistica
n=88 studenti

Bisogna mandare ai corsi di recupero gli studenti che hanno preso meno di 22.

Quanti sono?

La distribuzione di frequenze cumulate

Data la distribuzione $\{x_i;n_i\}$ della variabile X, la distribuzione **non decrescente** delle frequenze cumulate, indicata con $\{X \le x_i;N_i\}$, si ottiene enumerando le unità per le quali $X \le x_i$ per i=1, 2, …, k.

$$N_i = \sum_{h=1}^{i} n_h \quad \Rightarrow \quad N_{i-1} \le N_i \text{ per } i=1, 2, \ldots, k$$

La distribuzione **non crescente**, è indicata con $\{X \ge x_i;N'_i\}$

$$N'_i = \sum_{h=i}^{k} n_h \quad \Rightarrow \quad N'_{i-1} \ge N'_i \text{ per } i=1, 2, \ldots, k$$

Problema

Qual è la quota di studenti con meno di 22?

X	n_i	f_i
18	3	0,034
19	5	0,057
20	7	0,080
21	5	0,057
22	10	0,114
23	6	0,068
24	10	0,114
25	6	0,068
26	11	0,125
27	6	0,068
28	5	0,057
29	10	0,114
30	4	0,045
Totale	88	1,000

Le frequenze relative cumulate

Vengono indicate con F_i e P_i rispettivamente le distribuzioni **relativa cumulata non decrescente** e **percentuale cumulata non crescente**

$$F_i = \sum_{h=1}^{i} f_h$$

$$P_i = \sum_{h=1}^{i} f_h \cdot 100$$

Con F_i' e P_i' si indicano le distribuzioni relative cumulate **non crescenti**

$$F_i' = \sum_{h=i}^{k} f_h$$

$$P_i' = \sum_{h=i}^{k} f_h \cdot 100$$

Problema

Qual è la quota percentuale di atleti che fanno meno di 5,30 sui 30 metri?
E di quelli che fanno più di 5,90?

X=tempo 30m	n_i	p_i
4,90 |— 5,10	8	8,2
5,10 |— 5,30	20	20,4
5,30 |— 5,50	16	16,3
5,50 |— 5,70	20	20,4
5,70 |— 5,90	10	10,2
5,90 |— 6,10	10	10,2
6,10 |— 6,30	6	6,1
6,30 o più	8	8,2
Totale	98	100,0

Le rappresentazioni grafiche

La trasformazione delle distribuzioni di frequenza da forma tabellare a grafici ha senso se vengono resi di più facile lettura i dati.

Aspetti da valutare di una rappresentazione grafica

- Accuratezza
- Semplicità
- Chiarezza
- Aspetto
- Struttura

Diagramma a nastri

Colore	Maschi	Femmine
Marrone	48	36
Nero	31	37
Azzurro	15	22
Verde	12	8
Totale	106	103

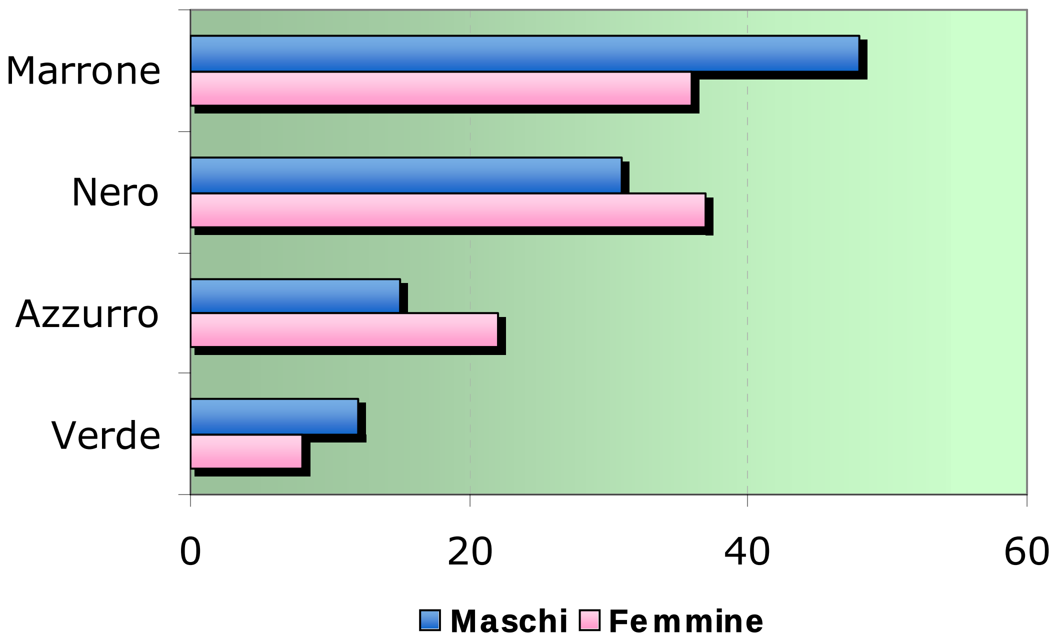


Diagramma a barre

Voti	Studenti
18	3
19	5
20	7
21	5
22	10
23	6
24	10
25	6
26	11
27	6
28	5
29	10
30	4
Totale	88

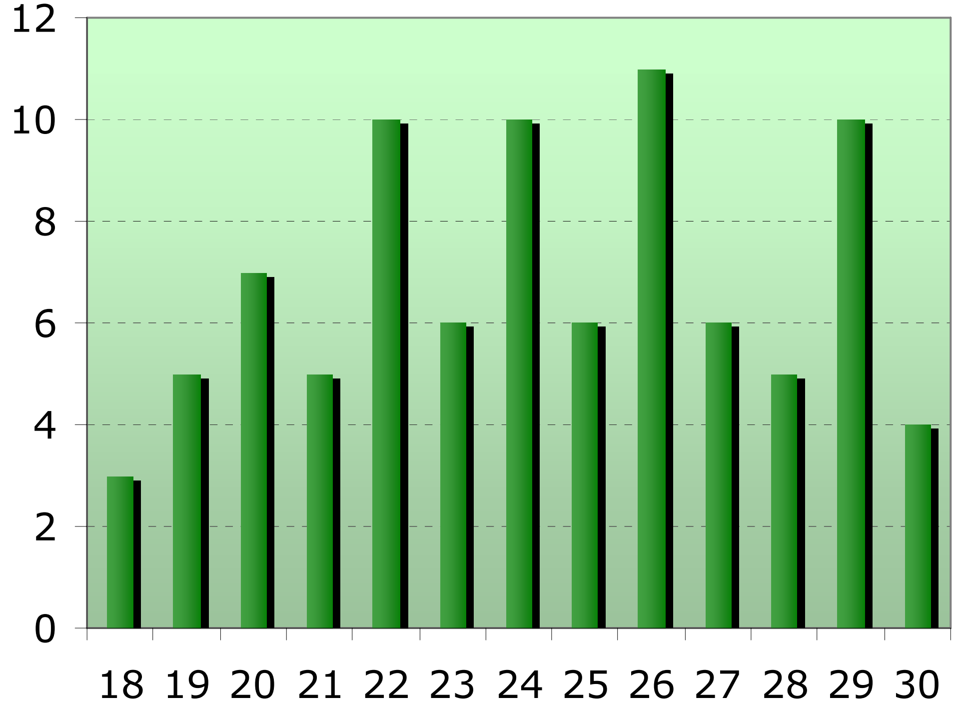
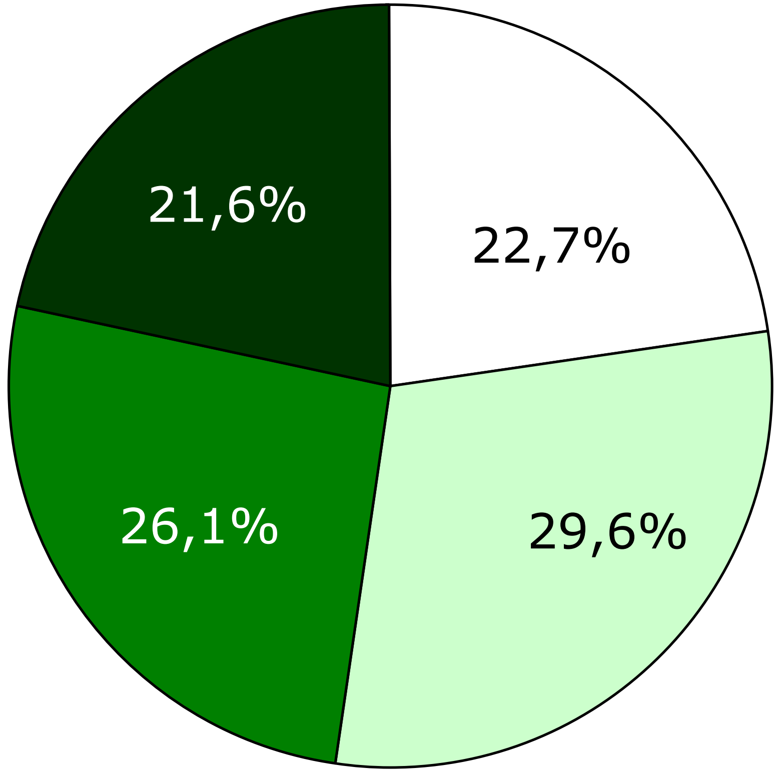


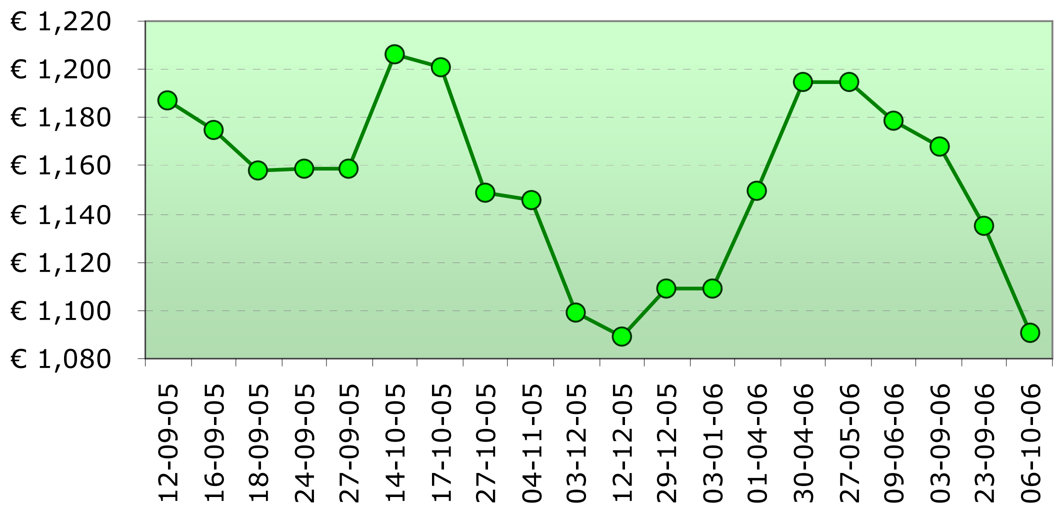
Diagramma a torta

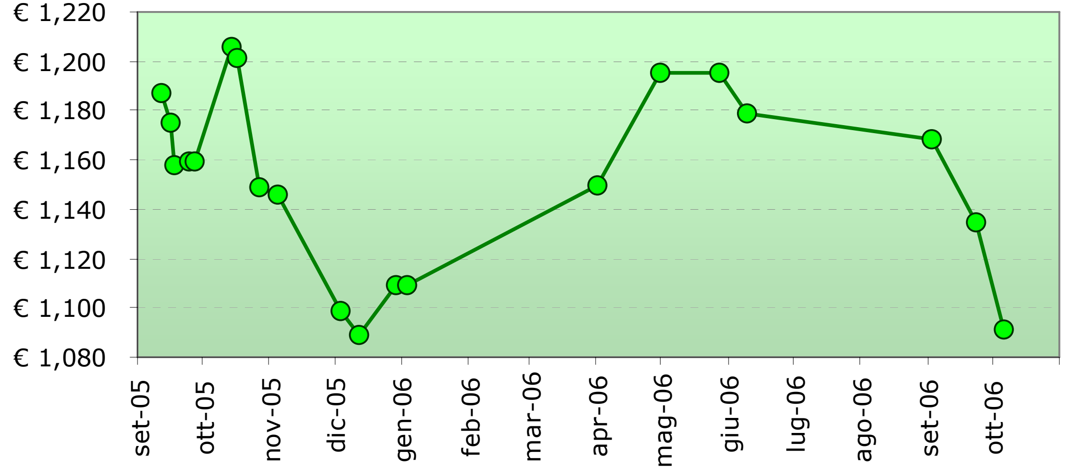
Voti	Studenti
18|-|21	20
22|-|24	26
25|-|27	23
28|-|30	19
Totale	88



Grafici lineari e diagrammi cartesiani

Data	Prezzo
12-09-05	€ 1,187
16-09-05	€ 1,175
18-09-05	€ 1,158
24-09-05	€ 1,159
27-09-05	€ 1,159
14-10-05	€ 1,206
17-10-05	€ 1,201
27-10-05	€ 1,149
04-11-05	€ 1,146
03-12-05	€ 1,099
12-12-05	€ 1,089
29-12-05	€ 1,109
03-01-06	€ 1,109
01-04-06	€ 1,150
30-04-06	€ 1,195
27-05-06	€ 1,195
09-06-06	€ 1,179
03-09-06	€ 1,168
23-09-06	€ 1,135
06-10-06	€ 1,091





Istogramma

- Nel caso di distribuzioni di un carattere quantitativo continuo con classi di ampiezza diversa è possibile ottenere una rappresentazione più efficace tramite gli istogrammi
- L’istogramma è un diagramma cartesiano basato sulla nozione di densità di frequenza

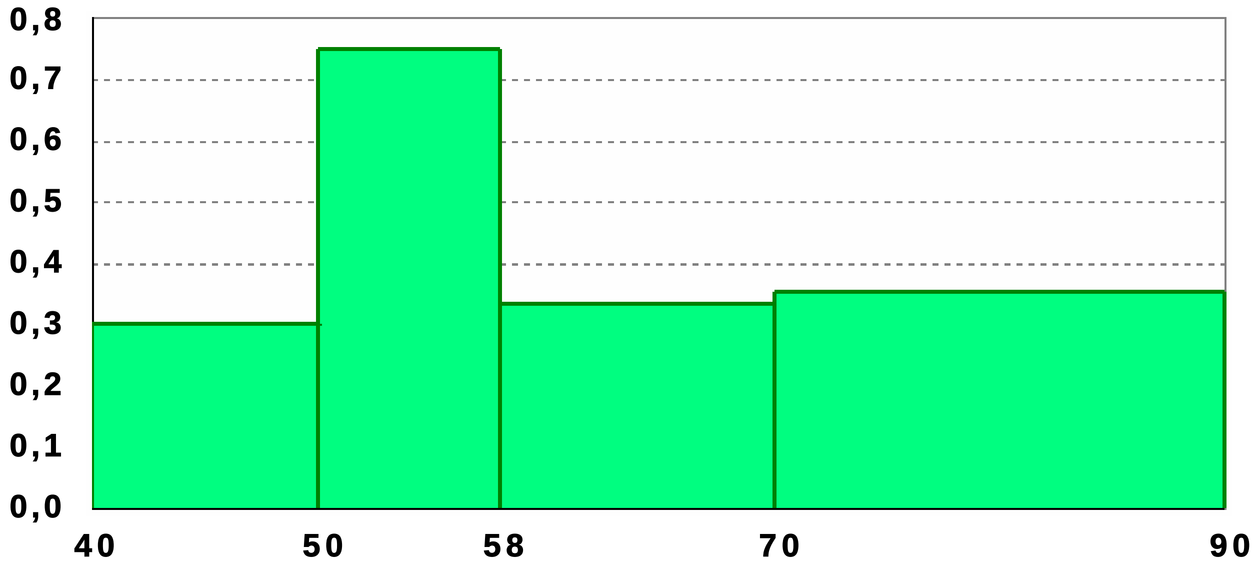
Densità di frequenza

Data la distribuzione $\{x_{i-1}\text{-}x_i; n_i\}$, si chiama densità di frequenza della i-esima classe la frequenza media nell'intervallo unitario contenuto in $x_{i-1}\text{-}x_i$, vale a dire il rapporto h_i della frequenza di tale classe rispetto l'ampiezza dell'intervallo corrispondente.

$$\boxed{h_i = n_i / (x_i - x_{i-1}) = n_i / w_i} \quad \text{per } i=1, \ldots, k$$

Istogramma

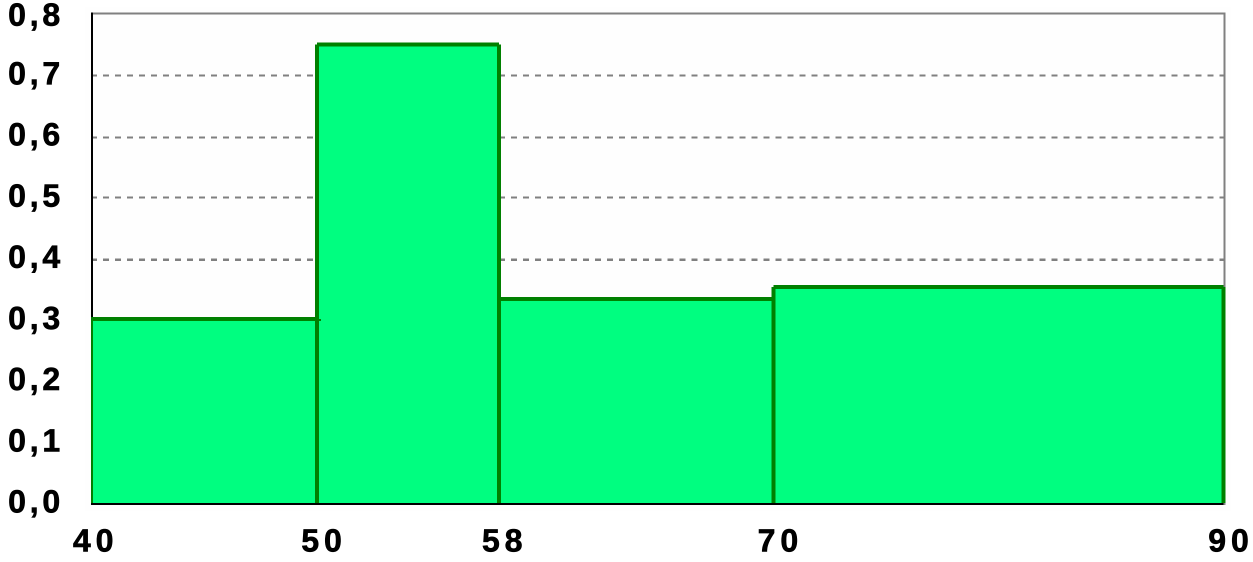
X=Peso in kg	n_i	w_i	h_i
40 —| 50	3	10	0,30
50 —| 58	6	8	0,75
58 —| 70	4	12	0,33
70 —| 90	7	20	0,35



Istogramma

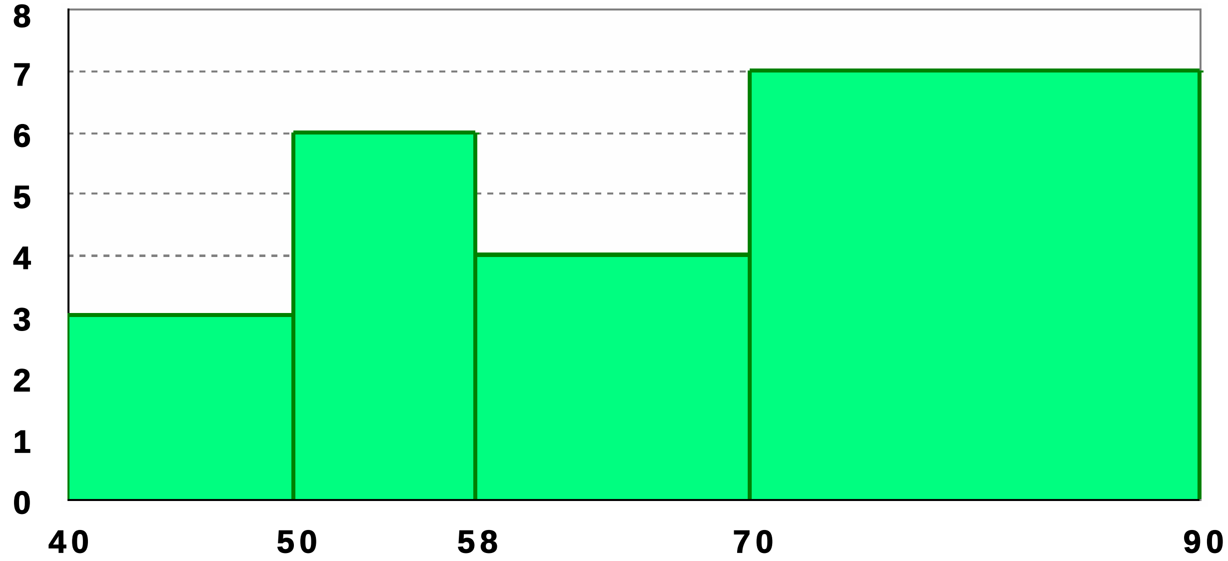
X=Peso in kg	h_i
40 —| 50	0,30
50 —| 58	0,75
58 —| 70	0,33
70 —| 90	0,35

Corretto →



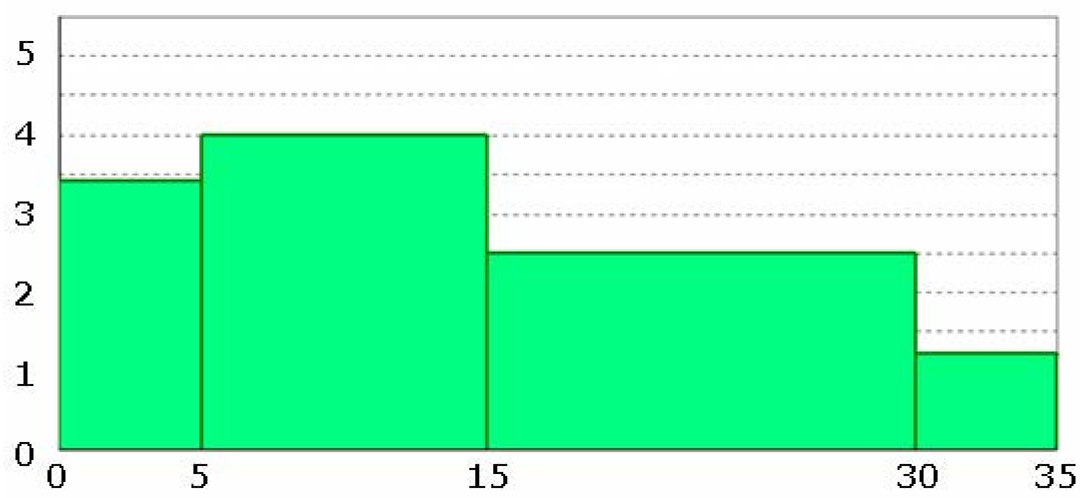
X=Peso in kg	n_i
40 —| 50	3
50 —| 58	6
58 —| 70	4
70 —| 90	7

Sbagliato →



Istogramma

Classi di età	p_i
0 |— 5	17,0
5 |— 15	40,0
15 |— 30	37,0
30 |— 35	6,0
Totale	100,0



Indici statistici di posizione

- Occorrono misure sintetiche che consentano il passaggio da una pluralità di informazioni (le modalità e le rispettive frequenze) a una sola grandezza
- Obiettivo di una misura di posizione è quello di sintetizzare in una singola modalità l'intera distribuzione di frequenza per effettuare confronti nel tempo, nello spazio o tra circostanze differenti

Indici statistici di posizione

- Medie analitiche
  - Media aritmetica
  - Media geometrica
  - Media armonica
- Medie di posizione o Medie lasche
  - Moda
  - Mediana
  - Quantili

Media aritmetica

La media aritmetica di un insieme di n valori osservati $x_1, x_2, \ldots, x_n$ di un carattere quantitativo X è pari alla somma dei valori osservati divisa per il loro numero:

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

Calcolo della media aritmetica

Formula generale per il calcolo della media aritmetica

$$\boxed{\bar{x} = \frac{c}{n}}$$

c = Ammontare totale del carattere

n = Numero di unità statistiche

Media aritmetica
per un protocollo elementare

Collettivo in esame → 88 individui iscritti al corso di Statistica

Carattere osservato → Voto conseguito all'esame di Statistica

Protocollo elementare

{29, 29, 24, 20, 22, 28, 19, 19, 21, 26, 20, 24, 21, 19, 25, 25,
23, 28, 22, 29, 26, 23, 28, 30, 20, 27, 22, 27, 20, 24, 25, 18,
26, 29, 29, 23, 23, 24, 22, 25, 27, 26, 23, 18, 19, 26, 22, 25,
20, 26, 22, 24, 20, 22, 21, 29, 30, 19, 24, 24, 26, 26, 29, 30,
29, 25, 28, 26, 22, 27, 27, 29, 26, 26, 22, 27, 24, 29, 30, 20,
24, 24, 21, 18, 22, 28, 23, 21}

$$\bar{x} = \frac{c}{n}$$

$c = 29+29+\ldots+23+21 = 2140$

$n = 88$

$$\boxed{\bar{x} = 24{,}32}$$

Media aritmetica per un protocollo elementare

Formalmente, disponendo del protocollo elementare la media aritmetica si calcola con la seguente formula:

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

Esempio

Tempi di percorrenza in minuti, del tratto casa-lavoro, rilevati per 12 giorni utilizzando due diversi mezzi di trasporto

Viaggio	Auto	Metro
1	23	22
2	32	24
3	44	22
4	21	33
5	36	26
6	30	31
7	28	24
8	33	28
9	45	32
10	34	31
11	29	37
12	31	24

Con quale mezzo si fa prima?

Tempo minimo: 21 (Auto)

Tempo massimo: 44 (Auto)

Tempo medio Auto: 32,17

Tempo medio Metro: $\boxed{27,83}$

Media aritmetica
per una distribuzione di frequenza

X	n_i
18	3
19	5
20	7
21	5
22	10
23	6
24	10
25	6
26	11
27	6
28	5
29	10
30	4
Totale	88

$$\bar{x} = \frac{c}{n}$$

$$c = \sum_{i=1}^{k} c_i = \sum_{i=1}^{k} x_i n_i$$

$$c = 2140 \qquad n = 88$$

$$\bar{x} = \frac{2140}{88} = 24{,}32$$

Media aritmetica per una distribuzione di frequenza

Formalmente, disponendo della distribuzione di frequenza di un carattere quantitativo discreto con le modalità non espresse in intervalli, la media aritmetica si calcola con la seguente formula:

$$\boxed{\bar{x} = \frac{c}{n} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}}$$

Esempio

Numero di figli a carico rilevati su un campione di famiglie. Determinare il numero medio di figli a carico per famiglia.

X= n° figli a carico	n_i
0	5
1	5
2	3
3	3
4	4

$$\bar{x} = \frac{c}{n}$$

$$c = 36$$
$$n = 20$$

$$\bar{x} = \frac{36}{20} = 1{,}8$$

Esempio di media per una distribuzione con classi intervallari

Tempo di percorrenza sui 30 metri di un campione di 98 atleti. Determinare il tempo medio di percorrenza sui 30 metri.

X=tempo 30m	n_i
4,90 |— 5,10	8
5,10 |— 5,30	20
5,30 |— 5,50	16
5,50 |— 5,70	20
5,70 |— 5,90	10
5,90 |— 6,10	10
6,10 |— 6,30	6
6,30 o più	8

$$\bar{x} = \frac{c}{n}$$

$$c = 550$$
$$n = 98$$

$$\bar{x} = \frac{550}{98} = 5,61$$

Media aritmetica per una distribuzione di frequenza con classi intervallari

Formalmente, disponendo della distribuzione di frequenza di un carattere quantitativo con le modalità espresse in intervalli, la media aritmetica si calcola con la seguente formula:

$$\bar{x} = \frac{c}{n} = \frac{\sum_{i=1}^{k} \hat{x}_i n_i}{\sum_{i=1}^{k} n_i}$$

dove $\hat{x}_i$ è il valore centrale dell’intervallo

Esempio di media per una distribuzione con classi intervallari

Peso in chilogrammi rilevato su un collettivo di individui. Determinare il peso medio degli individui del collettivo.

X=Peso in kg	n_i
40 —| 50	3
50 —| 58	6
58 —| 70	4
70 —| 90	7

$$\bar{x} = \frac{c}{n}$$

$$c = 1275$$
$$n = 20$$

$$\bar{x} = \frac{1275}{20} = 63{,}75$$

Esempio di media per una distribuzione con classi intervallari

Distribuzione dei comuni per classi di ampiezza demografica in Italia nel 2002. Determinare la popolazione media per comune.

Classi di ampiezza	Comuni
0 —| 1.000	1.981
1.000 —| 3.000	2.654
3.000 —| 5.000	1.191
5.000 —| 10.000	1.154
10.000 —| 20.000	649
20.000 —| 40.000	289
40.000 —| 80.000	121
80.000 —| 250.000	50
Oltre 250.000	12
Totale	8.101

$$\boxed{\bar{x} = \frac{c}{n}}$$

$$\boxed{\bar{x} = \frac{70.132.500}{8.101} = 8.657}$$

Esempio di media per una distribuzione con classi intervallari

Classi di ampiezza	Comuni	c_i	Popolazione
0 —| 1000	1.981	990.500	1.107.695
1000 —| 3000	2.654	5.308.000	4.877.693
3000 —| 5000	1.191	4.764.000	4.607.838
5000 —| 10000	1.154	8.655.000	8.088.910
10000 —| 20000	649	9.735.000	8.870.625
20000 —| 40000	289	8.670.000	8.090.304
40000 —| 80000	121	7.260.000	6.657.657
80000 —| 250000	50	8.250.000	6.199.631
Oltre 250000	12	16.500.000	8.820.717
Totale	8.101	70.132.500	57.321.070

$$\bar{x}^* = 8.657$$

$$\bar{x} = \frac{57.321.070}{8.101} = 7.076$$

Esempio

La tabella riporta il tempo necessario a produrre tre diversi tipi di copertone per auto. Determinare il tempo medio di produzione.

Tipo	Tempo	Pezzi prodotti
A	45	400
B	15	2.500
C	18	1.500
Totale		4.400

$$\bar{x} = \frac{c}{n}$$

$$c = 82.500$$
$$n = 4.400$$

$$\bar{x} = \frac{82.500}{4.400} = 18,75$$

Proprietà della media aritmetica

- Identità di somma
- Nullità della somma algebrica degli scarti dalla media aritmetica
- La somma dei quadrati degli scarti dalla media aritmetica è un minimo
- Equivarianza rispetto a trasformazioni lineari
- La media è compresa tra la più piccola e la più grande modalità della distribuzione
- Associatività

Identità di somma

Il valore costante $\bar{x}$, attribuito a ciascuna delle n unità, è tale da riprodurre, nell'insieme, l'ammontare totale del carattere.

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

$$\boxed{\sum_{i=1}^{m} x_i n_i = n\bar{x}}$$

Nullità della somma algebrica degli scarti

La somma delle differenze tra i valori x_j e la loro media aritmetica $\bar{x}$ è pari a zero.

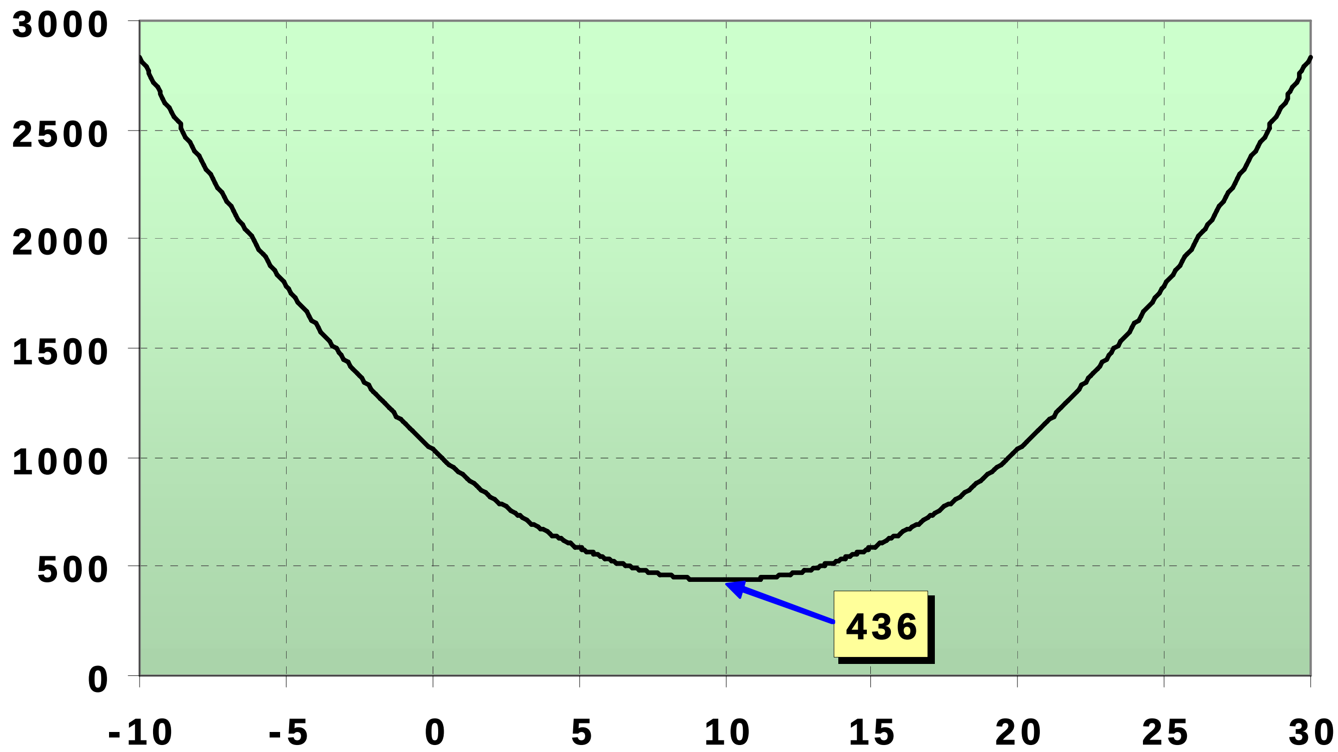
$$\boxed{\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0}$$

La somma dei quadrati degli scarti dalla media aritmetica è un minimo

La somma degli scarti al quadrato dei valori x_i da una costante *c* è minima quando *c* è uguale alla media aritmetica

$$\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = \min$$

Grafico con l’andamento della somma degli scarti al quadrato da *c* al variare di *c*



Sono dati i valori 2,3,5,7,18 e 25 (media 10).
Per *c* pari alla media aritmetica, si ha il minimo.

Equivarianza rispetto a trasformazioni lineari

Se si opera una trasformazione lineare della variabile X, tale che l'unità di misura viene moltiplicata per una costante *a* e l'origine spostata di *b* unità, allora la media aritmetica dei valori trasformati ax_1+b, ax_2+b,… , ax_n+b coincide con il valore trasformato della media aritmetica dei valori originali.

$$\boxed{M(aX+b)=aM(X)+b}$$

$$M(aX+b)=\frac{\sum_{i=1}^{n}(ax_i+b)}{n}=\frac{a\sum_{i=1}^{n}x_i+nb}{n}=aM(X)+b$$

Esempio

Anno	Celsius	Fahrenheit
1983	18,4	65,12
1984	16,8	62,24
1985	16,9	62,42
1986	17,7	63,86
1987	16,8	62,24
Media	17,32	62,176

$$X^F=32+1,8X^C$$

La temperatura media in Fahrenheit può essere ottenuta sia come media delle cinque temperature Fahrenheit, sia operando la trasformazione della temperatura media Celsius.

La media è compresa tra la più piccola e la più grande modalità della distribuzione

Indicate con x_m e x_M rispettivamente la più piccola e la più grande modalità della distribuzione e con M(X) la media del carattere, tale proprietà è indicata nel seguente modo

$$\boxed{x_m \leq M(X) \leq x_M}$$

Associatività

La media aritmetica generale $\bar{x}$ è una media aritmetica ponderata delle medie dei sottoinsiemi con pesi uguali alle loro numerosità

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \ldots + \bar{x}_G n_G}{n_1 + n_2 + \ldots + n_G} = \frac{1}{n} \sum_{i=1}^{G} \bar{x}_i n_i$$

Esempio

Collettivo di 10 unità:{2,3,3,4,5,7,8,11,13,15}

Media=7,1

Sottoinsieme 1: {2,3,3,4}

$\text{Media}_1=12/4=3$

Sottoinsieme 2: {5,7,8,11,13,15}

$\text{Media}_2=59/6=9{,}83$

$$\text{Media}=(\text{Media}_1\cdot n_1+\text{Media}_2\cdot n_2)/(n_1+n_2)=$$

$$=(3\cdot 4+9{,}83\cdot 6)/10=(12+59)/10=71/10=7{,}1$$

Esempio

Calcolare l'età media dell'intera popolazione.

Sesso	Frequenza	Età media
Maschi	37.500	45
Femmine	28.100	49

Influenza dei valori estremi sulle medie

Consideriamo i seguenti 8 valori
{3, 5, 5, 6, 8, 8, 9, 150}

La media aritmetica è 24,25

7 degli 8 valori sono inferiori alla media!!

La media aritmetica tiene “in eccessiva considerazione” il valore estremo molto alto

Moda

La moda m_t è la modalità della distribuzione che si presenta con la massima frequenza (assoluta, relativa o percentuale)

Qual è il valore modale?

Titolo di studio	Freq %
Analfabeti	3,1
Alfabeti	18,2
Licenza elementare	40,6
Licenza media	23,8
Diploma	11,5
Laurea	2,8

Moda di una mutabile sconnessa

Gruppi di corsi di laurea	Laureati
Scientifico	11.749
Medico	10.481
Ingegneria	12.056
Agrario	2.607
Economico	13.881
Politico-sociale	4.696
Giuridico	14.276
Letterario	17.466
Totale	87.212

Qual è il valore modale?

Moda di una mutabile ordinata

Titolo di studio	Freq %
Analfabeti	3,1
Alfabeti	18,2
Licenza elementare	40,6
Licenza media	23,8
Diploma	11,5
Laurea	2,8

Qual è il valore modale?

Moda di una variabile discreta

Voti	Studenti
18	3
19	5
20	7
21	5
22	10
23	6
24	10
25	6
26	11
27	6
28	5
29	10
30	4
Totale	88

Qual è il valore modale?

Classe modale

Qual è la classe modale?

X	Freq
0|—10	100
10|—50	150
50|—100	550
100|—200	1.000
Totale	1.800

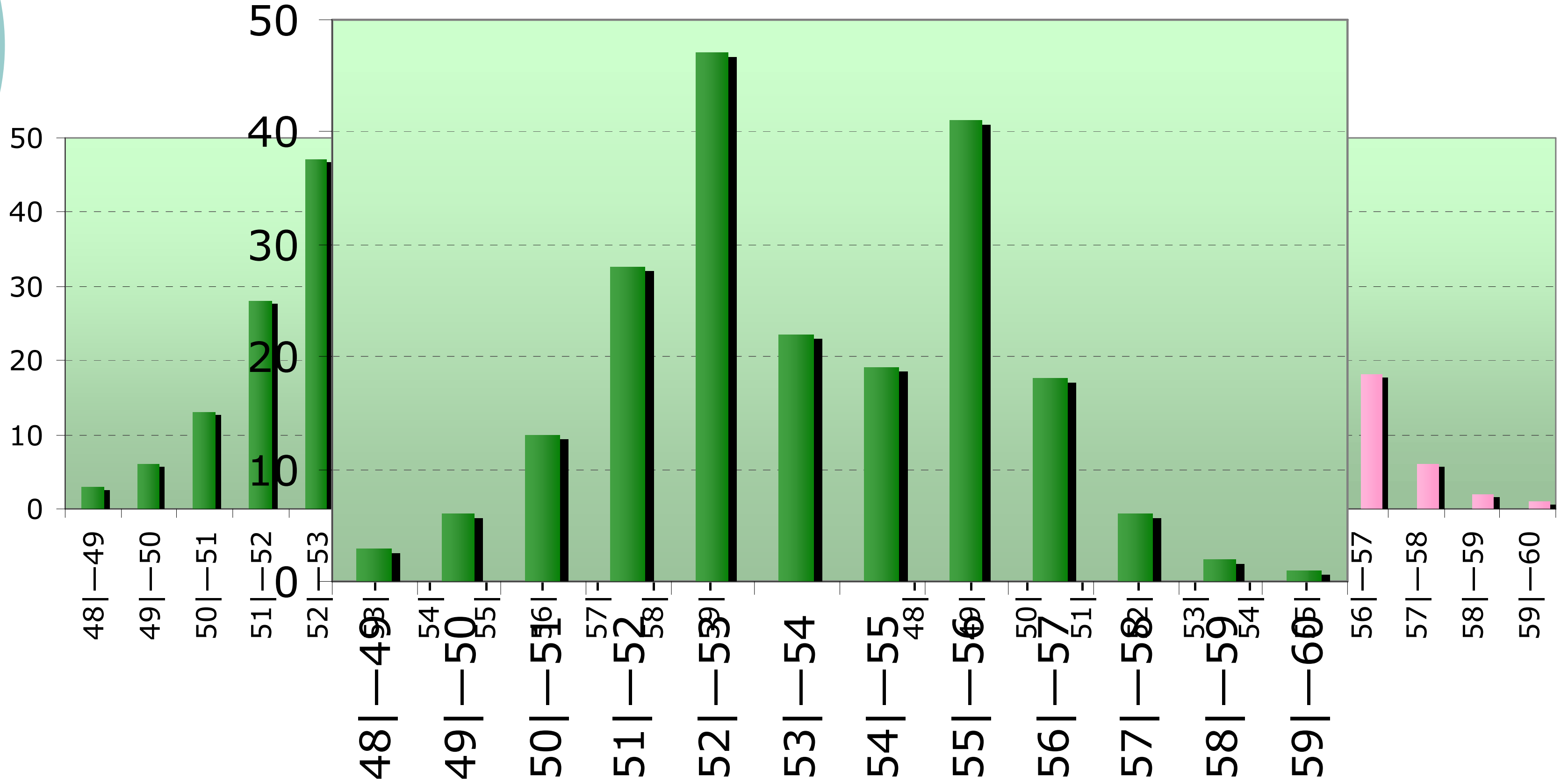
Se la distribuzione possiede classi di diversa ampiezza occorre, dividere le frequenze delle classi per la loro ampiezza e confrontare tali quozienti: la più elevata densità di frequenza individuerà la classe modale

Moda in una distribuzione bimodale

x_i	n_i
x_1	20
x_2	15
x_3	20
x_4	17
x_5	13
Totale	85

Se la distribuzione possiede due mode si dice bimodale.
Nella tabella le due mode sono rappresentate da x_1 e x_3

Distribuzione bimodale



Proprietà della moda

Prendiamo in considerazione un carattere quantitativo X.

Calcolando tutti gli scarti tra i valori individuali e un loro valor medio, la moda è, fra tutti i valori medi, quello che da luogo al maggior numero di differenze uguali a zero.

Mediana

La mediana m_C è il valore che occupa il posto centrale nella successione ordinata delle n osservazioni individuali.

E’ definita per mutabili ordinabili rettilineari e per caratteri quantitativi.

Esempio

Individuo	Giudizio
1	Insufficiente
2	Sufficiente
3	Buono
4	Sufficiente
5	Discreto
6	Ottimo
7	Ottimo

Qual è la mediana?

Calcolo della mediana

- Ordinare le unità
- Individuare la posizione centrale
  - Se n è dispari: (n+1)/2
  - Se n è pari esistono due posizioni centrali: n/2 e (n/2)+1
- Osservare la modalità presentata dall'unità centrale:
  - Se n è dispari la mediana è $m_c=x_{(n+1)/2}$
  - Se n è pari abbiamo due modalità corrispondenti alle due unità centrali: $x_{n/2}$ e $x_{(n/2)+1}$

Mediana per *n* pari e *n* dispari

Capi d'abbigliamento venduti in 5 negozi: {18, 27, 15, 11, 6}. Qual è la mediana m_c?

n=5 (dispari) → posizione centrale=(5+1)/2=3

Capi venduti	6	11	15	18	27

$\boxed{m_c=15}$

Sensibilità delle medie ai valori anomali

Età di 10 pazienti di un pediatra:

{5, 6, 3, 7, 3, 6, 8, 2, 5, 60}

Media aritmetica = 10,5

Età	2	3	3	5	5	6	6	7	8	60

Mediana = (5+6)/2 = 5,5

Esempio su dati qualitativi

Grado istruzione	Paesi Bassi	Portogallo
Primaria	1.139.955	1.203.119
Secondaria	1.466.956	489.696
Universitaria	384.407	101.308
Totale	2.991.318	1.794.123

Paesi Bassi
(2.991.318/2)=1.495.659
(2.991.318/2)+1=1.495.660

Portogallo
(1.794.123+1)/2=897.062

Grado istruzione	Paesi Bassi	Portogallo
Primaria	1.139.955	1.203.119
Almeno Secondaria	2.606.911	1.692.815
Almeno Universitaria	2.991.318	1.794.123

Esempio su dati quantitativi

Numero di figli a carico rilevati su un campione di famiglie. Determinare il numero mediano di figli a carico per famiglia.

X= n° figli a carico	n_i
0	5
1	5
2	4
3	3
4	4
Totale	21

N dispari → Posizione centrale = (21+1)/2 = $\boxed{11}$

Valore mediano = $\boxed{2}$

Altro esempio su dati quantitativi

Numero di figli a carico rilevati su un campione di famiglie. Determinare il numero mediano di figli a carico per famiglia.

X= n° figli a carico	n_i
0	5
1	5
2	3
3	3
4	4
Totale	20

N pari → Posizioni centrali $\boxed{10 \text{ e } 11}$

Valore mediano = (1+2)/2 = $\boxed{1{,}5}$

Proprietà della mediana

Per un carattere quantitativo X, la somma degli scarti in valore assoluto dei valori x_i da una costante c è minima quando c è uguale alla mediana

$$\sum_{i=1}^{n} \left| x_i - c \right| \quad \text{è minima per } c = m_c$$

Esempio

Dati i seguenti valori {2,4,6,8,8,10,11,15,20}, la mediana è 8.

Gli scarti in valore assoluto dalla mediana sono: {6,4,2,0,0,2,3,7,12} la cui somma è 36.

Con qualunque altra costante, diversa dalla mediana, otteniamo un valore più grande.

Prendiamo in considerazione la media aritmetica pari a 9,33

La somma degli scarti in valore assoluto dalla media è data da 7,33+5,33+3,33+1,33+1,33+ +0,67+1,67+5,67+10,67=37,33

Difetto della moda

Dati la seguente distribuzione calcolare moda, media aritmetica e mediana.

X	n_i
1	4
2	2
3	2
4	3
5	2
6	3
7	3
8	3
9	3
10	2
Totale	27

Moda = 1

Media = 146/27 = 5,41

Mediana (N dispari) = 6

Media e Mediana riescono a sintetizzare meglio questa distribuzione

Percentili e quartili

Definiamo percentili quei valori che dividono la popolazione in cento parti di uguale numerosità.

La mediana è il 50-esimo percentile.

Il 25-esimo e il 75-esimo percentile sono detti anche primo (Q_1) e terzo quartile (Q_3).

Calcolo dei quartili

Il primo quartile Q_1 è il valore tale che il 25% delle osservazioni è più piccolo di Q_1 e il 75% più grande di Q_1.

Q_1 = osservazione di posto $(n+1)/4$ nella lista ordinata

Il terzo quartile Q_3 è il valore tale che il 75% delle osservazioni è più piccolo di Q_3 e il 25% più grande di Q_3.

Q_3 = osservazione di posto $3(n+1)/4$ nella lista ordinata

Vero o Falso?

- La media aritmetica non può mai essere negativa

  FALSO

- La media aritmetica può essere calcolata su tutti i tipi di carattere

  FALSO

- La moda è la modalità con la frequenza maggiore

  VERO

Vero o Falso?

- I quartili dividono in tre parti il collettivo

  FALSO

- La mediana non si può calcolare per caratteri qualitativi

  FALSO

- Se tutti gli individui presentano lo stesso valore del carattere, allora media, moda e mediana coincidono

  VERO