Introduzione alla Statistica descrittiva

- È la scienza che studia i *fenomeni collettivi* o di massa. Un fenomeno è detto collettivo o di massa quando è determinato solo attraverso una *molteplicità di osservazioni*.
  - *Esempi:* numero di componenti delle famiglie di una data area geografica, l'eta' dei cittadini di un certo paese, la lunghezza delle foglie di un tipo di pianta,la durata delle lampadine di una certa marca,…
- La statistica insegna a individuare i modi in cui un fenomeno si manifesta, a descriverlo sinteticamente, e a trarne da esso conclusioni più generali di fenomeni più ampi.

Definizioni preliminari

- **Popolazione statistica**: insieme o collettività entro cui si studia il fenomeno
  - *Esempio:* la popolazione statistica relativa alla durata delle lampadine di una certa marca (fabbrica) è costituita da tutte le lampadine prodotte da quella fabbrica.
- **Unità statistica**: ogni elemento della popolazione statistica.
- **Campione statistico**: un qualsiasi insieme di unità statistiche prese da tutta la popolazione. Un campione è dunque una parte della popolazione statistica.
  - *Esempio:* 50 delle lampadine prodotte dalla fabbrica (estratte a caso).

Definizioni preliminari

- **Variabile statistica**: il fenomeno collettivo si presenta secondo modalità diverse nelle varie unità statistiche, perciò lo chiameremo variabile statistica. Il valore assunto dalla variabile statistica in una data unità statistica lo chiameremo **osservazione**.
  - *Esempio:* variabile statistica: durata delle lampadine; osservazione: prima lampadina ha la durata di 230 ore, la seconda lampadina ha durata 300 ore, la terza….
- **Dati statistici** sono costituiti dal numero che esprime quantitativamente una modalità (230 ore) e dal numero delle volte in cui una modalità si presenta nell'indagine (10 volte).
- **Variabile quantitativa**: quando assume valori numerici (durata delle lampadine)
- **Variabile qualitativa**: quando assume valori non numerici (colore dell'iride di una persone).

Fasi di un'indagine statistica

- Individuazione dell'obiettivo da raggiungere, definendo con accuratezza i termini del problema a cui bisogna dare risposta, cioè quali variabili statistiche bisogna osservare.
- Vengono fissati i metodi, i mezzi e i tempi da utilizzare nella raccolta dati. Per quanto riguarda i metodi è fondamentale decidere se l'osservazione viene fatta su tutta la popolazione oppure su un campione.
- Programmazione dell'indagine ed effettiva rilevazione dei dati.
- Sistemazione dei dati raccolti in forma di facile lettura (tabelle e grafici). I dati allo stato grezzo sono riferiti alla singola unità statistica.

Gruppi sanguigni	N. Di persone
0	9
AB	3
A	3
B	5

Fasi di un'indagine statistica

- Rappresentazione grafica dei dati in tabella: rappresentazione più sintetiche e immediate.
- Determinazione di valori che descrivono sinteticamente il fenomeno: misure di tendenza centrale.
- Calcolo delle misure di dispersione che indicano quanto le misure di tendenza (per esempio la media) di discosta dai dati raccolti.
- Determinazione di rapporti statistici o numeri indici: rapporti tra numeri che a volte sono piu' significativi dai valori assoluti. (es. rapporto tra m^2 edificati in una regione e il numero di abitanti della regione)
- Nelle indagini statistiche a campione occorre effettuare delle generalizzazioni (di cui si occupa l'inferenza statistica).

Tabelle statistiche

- Spoglio dei dati e loro sistemazioni in tabelle:
  - **Tabella di variabile qualitativa**: serie.
  - **Tabella di variabile quantitativa**: seriazioni.
- Lo spoglio dei dati raccolti consiste nella loro divisione secondo i diversi valori che la variabile del fenomeno ha assunto.
- Per ogni valore si determina la frequenza assoluta ossia il numero di volte con la quale compare ogni valore.

Distribuzione di frequenze assolute vs distribuzione di intensità: nel primo caso i numeri derivano da una enumerazione (quante volte esce testa lanciando n volte una moneta), nel secondo caso i numeri derivano da una misurazione (temperature di un determinato mese).

Le distribuzioni di frequenza tendono a mostrare la distribuzione reale del fenomeno solo quando è possibile utilizzare un numero sufficientemente elevato di osservazioni.

Tabelle di variabile qualitativa

Distribuzione delle frequenze assolute

Esempio di spoglio in una serie

- Fenomeno: colore dell'iride degli alunni della classe II C della scuola media "a. Vespucci" di torino.
- Dati raccolti:
  - Marroni: 1+1+1+1+1+1+1+1
  - Neri:1+1+1+1
  - Azzurri: 1+1+1+1+1+1
  - Verdi:1+1
  - Grigi:1

Colore iride	N. Di persone
marrone	8
nero	4
azzurro	6
Verde	2
grigio	1

Valori assunti dalla variabile statistica

Frequenze assolute

Tabelle di variabile quantitativa

- Una *variabile quantitativa* si dice *discreta* se può assumere solo un numero finito di valori, oppure una infinità numerabile di valori.
  - Esempio: numero di libri in possesso delle famiglie di un certo territorio.
- Una *variabile quantitativa* si dice *continua* se può assumere uno qualsiasi dei valori di un certo intervallo di numeri reali.
  - Esempio: l'altezza delle persone.

–Esempio: numero di persone di componenti delle famiglie degli alunni di una data scuola.

Tabella delle frequenze assolute

Numero componenti	Frequenze assolute
2	10
3	7
4	15
5	2

Tabelle di variabile quantitativa

Raggruppamento in classi

- *Distribuzione di frequenza in classi*: per variabili continue le osservazioni possono essere molto disperse cioè molte osservazioni sono diverse tra loro. Per essa risulta non molto significativa una tabella come quelle precedenti. In tal caso si raggruppano i dati in fasce di classi.
  - Esempio: l'altezza delle persone.
- Arrotondamenti di dati per inserirli in una delle classi

Altezza (cm) CLASSI	Frequenze assolute	Valori centrali
151-155	4	153
156-160	9	158
161-165	15	163
166-170	7	168
171-175	8	173
176-180	3	178
181-185	3	183
186-190	1	188

Intervallo di classe: 151-155
limiti della classe: 151 e 155
•Inferiore (151)
•Superiore (155)

Anche 150,8 e 154,8 appartengono alla classe,
I limiti reali o confini della classe sono 160,5 e 165,5

Tabelle di variabile quantitativa

- *Confine inferiore*: (limite superiore della classe precedente + limite inferiore della classe)/2
  - Esempio: confine inferiore seconda classe (155+156)/2=155,5
- *Confine superiore*: (limite superiore della classe + limite superiore della classe successiva)/2
  - Esempio: confine inferiore seconda classe (160+161)/2=160,5
- *Ampiezza di una classe* = confine superiore – confine inferiore della classe
  - Esempio: ampiezza seconda classe= 160,5-155,5=5
- *Valore centrale* = semisomma dei limiti della classe
- *Campo di variazione*: differenza tra il valore osservato più grande e quello più piccolo. Questa info serve per decidere l'ampiezza di una classe.
  - Esempio: Supponiamo che il valore min osservato sia 150,9 e che il max sia 189, allora il campo di variazione è 189 -150,8=38,2

Raggruppamento in classi

Altezza (cm) CLASSI	Frequenze assolute
151-155	4
156-160	9
161-165	15
166-170	7
171-175	8
176-180	3
181-185	3
186-190	1

Tabelle di variabile quantitativa

Raggruppamento in classi

Altezza (cm) CLASSI	Frequenze assolute	Confine inferiore	Confine superiore	ampiezza	Valori centrali
151-155	4				
156-160	9	155,5			
161-165	15				
166-170	7				
171-175	8				
176-180	3	180,5			
181-185	3				
186-190	1	185,5			

Confine inferiore: (limite superiore della classe precedente + limite inferiore della classe)/2

Tabelle di variabile quantitativa

Raggruppamento in classi

Altezza (cm) CLASSI	Frequenze assolute	Confine inferiore	Confine superiore	ampiezza	Valori centrali
151-155	4	150,5	155,5		
156-160	9	155,5			
161-165	15	160,5			
166-170	7	165,5	170,5		
171-175	8	170,5			
176-180	3	175,5			
181-185	3	180,5			
186-190	1	186,5	190,5		

•*Confine superiore*: (limite superiore della classe + limite superiore della classe successiva)/2

Tabelle di variabile quantitativa

Raggruppamento in classi

Altezza (cm) CLASSI	Frequenze assolute	Confine inferiore	Confine superiore	ampiezza	Valori centrali
151-155	4	150,5	155,5		
156-160	9	155,5	160,5		
161-165	15	160,5	165,5		
166-170	7	165,5	170,5	5	
171-175	8	170,5	175,5		
176-180	3	175,5	180,5		
181-185	3	180,5	185,5		
186-190	1	185,5	190,5		

Ampiezza di una classe = confine superiore – confine inferiore della classe

Tabelle di variabile quantitativa

Raggruppamento in classi

Altezza (cm) CLASSI	Frequenze assolute	Confine inferiore	Confine superiore	ampiezza	Valori centrali
151-155	4	150,5	155,5	5	
156-160	9	155,5	160,5	5	
161-165	15	160,5	165,5	5	
166-170	7	165,5	170,5	5	168
171-175	8	170,5	175,5	5	
176-180	3	175,5	180,5	5	
181-185	3	180,5	185,5	5	
186-190	1	185,5	190,5	5	

•*Valore centrale* = semisomma dei limiti della classe

Tabelle di variabile quantitativa

Raggruppamento in classi

Altezza (cm) CLASSI	Frequenze assolute	Confine inferiore	Confine superiore	ampiezza	Valori centrali
151-155	4	150,5	155,5	5	153
156-160	9	155,5	160,5	5	158
161-165	15	160,5	165,5	5	163
166-170	7	165,5	170,5	5	168
171-175	8	170,5	175,5	5	173
176-180	3	175,5	180,5	5	178
181-185	3	180,5	185,5	5	183
186-190	1	185,5	190,5	5	188

Esercizio

Completare la seguente tabella

Altezza (cm) CLASSI	Frequenze assolute	Confine inferiore	Confine superiore	ampiezza	Valori centrali
20-22	14				
23-25	45				
26-28	38				
29-31	26				
32-34	17				
35-37	10				

Tabelle di variabile quantitativa

- *Classi con ampiezza diversa*
- *Intervalli aperti a destra o a sinistra*: non si è specificato il limite inferiore o superiore della classe
- Gli intervalli aperti sono usati sia per variabili continue che per quelle discrete
- In genere, a parte per le classi estreme, si usano classi di uguale ampiezza

Altezza (cm) CLASSI	Frequenze assolute
151-155	4
156-160	9
161-165	15
166-170	7
171-175	8
176-190	7

Numero libri	Frequenze assolute
Fino a 15	7
16-30	55
31-50	302
51-100	210
Oltre 100	77

Tabelle di variabile quantitativa

- *Frequenza cumulata dal basso*: si ottiene dalla somma delle frequenze di una classe con tutte le frequenze delle classi che la precedono.
- *Frequenza cumulata dall'alto*: si ottiene dalla somma delle frequenze di una classe con tutte le frequenze delle classi che la susseguono.
- *Frequenza relativa*: rapporto tra la frequenza assoluta e il totale delle frequenze di una distribuzione
- *Frequenza percentuale*: frequenza relativa per 100.
- *Frequenza cumulata relativa*: rapporto fra la frequenza cumulata e il totale delle frequenze assolute.
- *Frequenza cumulata percentuale*: frequenza cumulata relativa per 100

Tabelle di variabile quantitativa

Altezza CLASSI	Frequenze assolute	Frequenze relative	Frequenza percentuale	Frequenze cumulate dal basso	Frequenze cumulate dall'alto	Freq.cum. dal basso relative	Frequenza cumulata percentuale
151-155	4	0.08	8	4	50	0.0231	2.31
156-160	9	0.18	18	13	46	0.0751	7.51
161-165	15	0.30	30	28	37	0.1618	16.18
166-170	7	0.14	14	35	22	0.2023	20.23
171-175	8	0.16	16	43	15	0.2485	24.85
176-190	7	0.14	14	50	7	0.2890	28.90
totale	50	1	100	173	177	1	100

N.B.: le diverse frequenze possono essere calcolate anche per Dati non raggruppati in classi

Esercizio

completare la seguente tabella

CLASSI	Frequenze assolute	Frequenze relative	Frequenza percentuale	Frequenze cumulate dal basso	Frequenze cumulate dall'alto	Freq.cum. dal basso relative	Frequenza cumulata percentuale
0-99	782						
100-499	419						
500-999	85						
1000-2999	23						
3000-6000	12						
totale	1321						

Esercizio

completare la seguente tabella

CLASSI	Frequenze assolute	Frequenze relative	Frequenza percentuale	Frequenze cumulate dal basso	Frequenze cumulate dall'alto	Freq.cum. dal basso relative	Frequenza cumulata percentuale
Fino a 9.5				0			
9.6-14.5				18			
14.6-19.5				46			
19.6-24.5				67			
24.6-29.5				80			
totale				80			

Vedremo come risolvere questi esercizi in Excel.
Esercizi tipo quelli nel paragrafo 14.4 del capitolo 14 (fotocopie).

Tabelle di variabile quantitativa

- Tabella a doppia entrata o a più colonne:spesso capita di studiare due variabili statistiche tra le quali esiste un qualche relazione. In questi casi si usa una tabella a doppia entrata.

Altezza \ peso	43-46	47-50	51-54	55-58	59-62	63-66	67-70
149-153	3	6	2	0	0	0	0
154-158	1	12	16	8	0	0	0
159-163	0	4	19	11	5	1	0
164-168	0	1	2	10	3	0	1
169-174	0	0	1	4	4	3	1
175-180	0	0	0	0	0	0	4

$F_{p,a}$

Esercizio

- Considerata la seguente tabella,
  - calcolare ampiezza delle classi di altezza e di peso
  - I confini delle classi di altezza e di peso
  - I valori centrali delle classi
  - La frequenza degli studenti appartenenti alla classe di altezza 154-158
  - La frequenza degli studenti appartenenti alla classe di peso 51-54
  - Le frequenze cumulate di altezza
  - Le frequenze cumulate di peso

Altezza \ peso	43-46	47-50	51-54	55-58	59-62	63-66	67-70
149-153	3	6	2	0	0	0	0
154-158	1	12	16	8	0	0	0
159-163	0	4	19	11	5	1	0
164-168	0	1	2	10	3	0	1
169-174	0	0	1	4	4	3	1
175-180	0	0	0	0	0	0	4

Rappresentazione grafica

Sono numerose e devono essere scelte in rapporto dei tipi di Dato e della scala di valori utilizzata

Vantaggio: evidenziano a colpo d'occhio le caratteristiche principali di un fenomeno.

Inconveniente: mancanza di precisione e altamente soggettive

Rappresentazioni grafiche

- *Dati quantitativi*:
  - istogrammi
  - poligoni
- *Dati qualitativi*:
  - Diagrammi a rettangoli distanziati
  - Diagrammi a punti
  - Areogrammi (tra cui i diagrammi circolari)
  - Diagrammi a figure (o diagrammi simbolici)

Tabella

Tabella 4. Distribuzione di frequenza assoluta e relativa (in %) dell'altezza di 40 giovani piante.

Classe	X_i	60-79	80-99	100-19	120-39	140-59	160-79	180-99
Freq. Assoluta	n_i	1	3	10	12	7	5	2
Freq. Relativa %	f_i	2,5	7,5	25,0	30,0	17,5	12,5	5,0
Freq. Cumulata	---	2,5	10,0	35,0	65,0	82,5	95,0	100,0

Istogrammi

- *Grafici a barre verticali* nei quali le misure della variabile statistica sono riportate sull'asse orizzontale mentre sull'asse verticale sono indicati il numero assoluto, oppure la frequenza relativa o quella percentuale, con cui compaiono i valori di ogni classe.
- In istogramma è una rappresentazione areale, sono le superfici dei rettangoli ad essere proporzionali alle frequenze corrispondenti.
- L'asse verticale deve sempre mostrare lo zero reale o origine, onde evitare di distorcere le caratteristiche de dati e i rapporti tra essi.

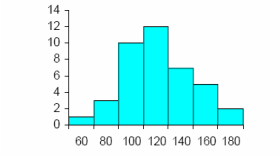


Figura 2. Istogramma dei dati di Tab. 4

(Valore iniz. =60; Valore finale =199; Passo =20; Classi=7)

Istogrammi

- Quando le classi hanno la stessa ampiezza, le basi dei rettangoli sono uguali, quindi saranno le altezze ad essere proporzionali alle frequenze che rappresentano. (ragionare sull'area equivale a ragionare sulle altezze)
- Anche quando le ampiezze delle classi sono diverse bisogna garantire che le superfici dei rettangoli siano sempre proporzionali alle frequenze che rappresentano, ne segue che è necessario ragionare sulle altezze dei rettangoli per garantire la correttezza del diagramma

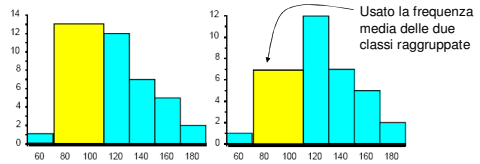


Figura 3. Istogrammi dei dati di Tab. 4

Somma errata di due classi : 2ª e 3ª della figura precedente

Somma corretta di due classi : 2ª e 3ª della figura precedente

Istogrammi

La rappresentazione grafica deve essere in grado di non alterare od interrompere la regolarità della distribuzione

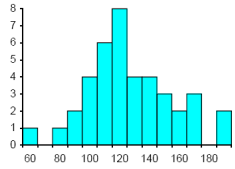


Figura 4. Istogramma dei dati di Tab. 4

(Valore iniziale = 60; Valore finale = 199; Passo = 10; Classi = 14)

(Rappresentazione grafica non adeguata, per eccessiva suddivisione in classi)

Poligoni

Un poligono può essere ottenuto a partire dal relativo istogramma, unendo con una linea spezzata i punti centrali di ogni classe.

La linea spezzata deve essere unita all'asse orizzontale, sia all'inizio sia alla fine, per racchiudere l'area della distribuzione.

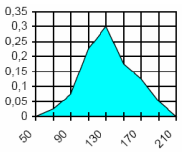


Figura 5. Poligono dei dati di Tab. 4

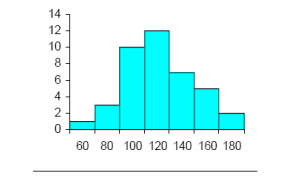


Figura 2. Istogramma dei dati di Tab. 4

(Valore iniz. =60; Valore finale =199; Passo =20; Classi=7)

Poligoni

Le distribuzioni cumulate sono rappresentate sia con istogrammi cumulati sia con **poligoni cumulati**.

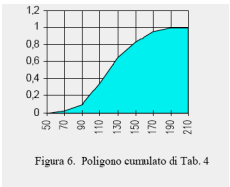


Figura 6. Poligono cumulato di Tab. 4

Rettangoli distanziati e ortogrammi

•I diagrammi a rettangoli distanziati, o grafici a colonne, sono formati da rettangoli con basi uguali ed altezze proporzionali alle intensità (o frequenze) dei vari gruppi considerati.
•A differenza degli istogrammi, sull'asse delle ascisse non vengono riportati misure ordinate ma nomi, etichette o simboli, propri delle classificazioni qualitative.

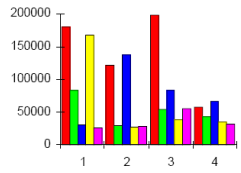


Figura 8. Rettangoli distanziati

Con dati qualitativi o nominali, le basi dei rettangoli sono sempre identiche avendo solo un significato simbolico.

Diagrammi a barre, che rappresentano le frequente, con linee continue più o meno spesse (figura 8).

Rettangoli distanziati e ortogrammi

Gli **ortogrammi o grafici a nastri** sono uguali ai rettangoli distanziati; l'unica differenza è che gli assi sono scambiati, per una lettura più facile

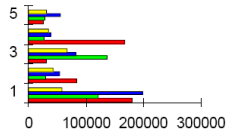


Figura 9. Ortogramma

Anche in questo caso è possibile sostituire ai rettangoli una linea ottenendo **diagrammi a barre o a punti** e l'intensità o frequenza delle varie classi viene letta con una proiezione sull'asse delle ascisse.

Areogrammi

•**Gli areogrammi** sono grafici in cui le frequenze o le quantità di una variabile qualitativa sono rappresentate da superfici di figure piane, come quadrati, rettangoli o, più frequentemente, cerchi oppure loro parti.
•La rappresentazione può essere fatta sia con più figure dello stesso tipo, aventi superfici proporzionali alle frequenze o quantità, sia con un'unica figura suddivisa in parti proporzionali.
•Nel caso dei **diagrammi circolari o a torta**, si divide un cerchio in parti proporzionali alle classi di frequenza.
•Gli areogrammi vengono usati soprattutto per rappresentare **frequenze percentuali**.

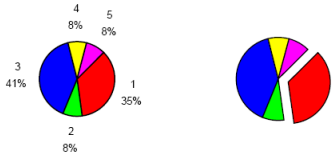


Figura 12. Diagrammi circolari

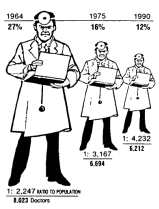
Figure

Con i **diagrammi a figure**, o **diagrammi simbolici** o **pittogrammi**, la frequenza di ogni carattere qualitativo viene rappresentata da una figura, sovente stilizzata, oppure da simboli che ricordano facilmente l'oggetto.

Questi diagrammi a figure **hanno tuttavia il grave inconveniente di prestarsi a trarre in inganno con facilità.**

E' una specie di istogramma costruito con figure, dove **l'altezza della figura deve essere proporzionale alla frequenza, quando le basi sono uguali**.

L'occhio coglie complessivamente non l'altezza di ogni figura ma la superficie che essa occupa, che è il quadrato del valore che si intende rappresentare.



Figure

E' possibile ovviare all'inconveniente, ricorrendo all'artificio di **figure identiche, ripetute tante volte quante sono le proporzioni**.

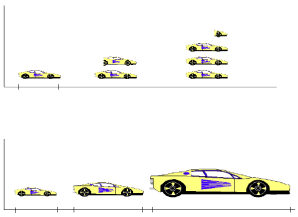


Figura 13. Pittogramma della produzione mensile di auto di 3 case automobilistiche: la prima ha prodotto 100 mila auto, la seconda 180 mila e la terza 320 mila.
La parte superiore della figura fornisce una rappresentazione corretta.
La parte inferiore, fondata sulla proporzione della lunghezza, fornisce una rappresentazione errata: è la superficie coperta dalla figura che deve essere proporzionale, non la lunghezza.

…sintesi dei valori osservati

Per i **caratteri qualitativi**, la tabella e le rappresentazioni grafiche esauriscono quasi completamente gli aspetti descrittivi, quando sia possibile leggere con esattezza le frequenze delle varie classi.

Per i **caratteri quantitativi**, si pone il problema di **sintesi oggettive** che possano essere elaborate matematicamente e quindi che siano **numeriche**, al fine di **un'analisi obiettiva che deve condurre tutti i ricercatori, con gli stessi dati, alle medesime conclusioni**.

Una serie di dati numerici è compiutamente descritta da **3 proprietà principali**:
1) la **tendenza centrale o posizione**;
2) la **dispersione o variabilità**;
3) la **forma**. (che non vedremo)

Misure di tendenza centrale o posizione

- Servono per individuare il valore intorno al quale i dati sono raggruppati.
- La tendenza centrale è la misura più appropriata per sintetizzare l'insieme delle osservazioni, se una distribuzione di dati dovesse essere descritta con un solo valore
- E' la prima indicazione della dimensione del fenomeno.
- Le misure proposte sono essenzialmente tre: **la media, la moda e la mediana**.

Media aritmetica

La **media aritmetica semplice** è la misura di tendenza centrale più comunemente utilizzata.

Quando si parla solo di **media**, si intende la media aritmetica semplice

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

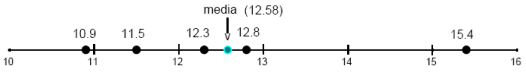
dove:
- x = media del campione
- $\mathbf{x_i}$ = **i**-esima osservazione della variabile X
- **n** = numero di osservazioni del campione
- $\sum_{i=1}^{n}$ = sommatoria di tutti gli **xi** del campione.

Media aritmetica

La media può essere vista come il baricentro della distribuzione campionaria, quando ogni singola osservazione è rappresentata da un peso convenzionale, identico per tutte, lungo l'asse che riporta i valori su una scala di intervalli o di rapporti.

Per dimostrare graficamente che **la media aritmetica corrisponde al punto di bilanciamento o di equilibrio dei dati,** si supponga di avere 5 misure: 10,9 11,5 12,3 12,8 15,4.

$$\bar{x} = \frac{10,9+11,5+12,3+12,8+15,4}{5} = 12,58$$



Esercizio

Numero componenti	Frequenze assolute
2	10
3	7
4	15
5	2

In media, quale è la grandezza delle famiglie considerate?
Di quanti componenti si compone in media una famiglia?

(2*10+3*7+4*15+5*2)/(10+7+15+2)=3.2647 componenti

Abbiamo usato la media ponderata!

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

Media ponderata

La **media aritmetica di distribuzioni di frequenza** raggruppate in classi, detta **media aritmetica ponderata**, è calcolata più rapidamente con

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

dove:
- x = media della distribuzione in classi,
- $\mathbf{x_i}$ = valore medio della **i**-esima classe di intervallo,
- $\mathbf{f_i}$ = numero di osservazioni della classe **i**-esima classe,
- **n** = numero di classi,
- $\sum_{i=1}^{n}$ = sommatoria per tutte le n classi

Media ponderata

ESEMPIO. Da un gruppo di 25 dati, raggruppati nella seguente distribuzione in classi

Classe	$\mathbf{X_i}$	150-159	160-169	170-179	180-189	190-199
Frequenza	$\mathbf{f_i}$	3	5	8	6	3

calcolare la media.

Risposta. Con la formula della **media ponderata**

$$(\text{media})\overline{x} = \frac{(155\cdot 3)+(165\cdot 5)+(175\cdot 8)+(185\cdot 6)+(195\cdot 3)}{3+5+8+6+3} = \frac{4385}{25} = 175,4$$

Esercizi

Calcolare le medie dei sequenti dati:

(a) 10 - 11 - 10 - 7 - 8 - 11 - 10 - 8 - 9 - 11 - 7 - 9

(b)

x_i	-2.2	-0.7	+2	+3.5	+4
f_i	12	23	27	31	21

(c)

Classi	frequenze
10-13	17
14-17	23
18-21	31
22-25	15
26-29	7

Esercizio

Un concorso è articolato in tre prove scritte e un orale, alle quali sono attribuiti rispettivamente i pesi: 1, 2, 4 e 5. Se un candidato ha riportato le votazioni 62, 60 70 e 75 e un altro ha riportato le votazioni 70, 65, 74, 62, quale dei due è primo in graduatoria?

Proprietà della media

- La media è sempre compresa tra il più piccolo e il più grande dei dati
- Chiamiamo **scarto** la differenza x_i-$\bar{x}$ tra un valore x_i e la media $\bar{x}$. La somma algebrica di tutti gli scarti è nulla.
- La somma dei quadrati degli scarti dei dati da un valore *v* è minima quando *v* coincide con la media dei dati.

Moda

- È un valor medio che dipende esclusivamente dalle frequenze f_i dei dati e non dai dati stessi.
- La **moda** (detta più raramente anche **dato prevalente**) è **il valore più frequente di una distribuzione**.
- Caratteristiche:
  - a non è influenzata dalla presenza di nessun valore estremo
  - viene utilizzata solamente a scopi descrittivi, perché **è meno stabile e meno oggettiva delle altre misure di tendenza centrale**.
  - Può differire nella stessa serie di dati, quando si formano classi di distribuzione con ampiezza differente.
  - Per individuare la moda entro una classe di frequenza, non conoscendo come i dati sono distribuiti, si ricorre all'ipotesi della **uniforme ripartizione**.

Esempio

Eta' degli alievi di due classi.

Moda?

17 per entrambe le distribuzioni

Eta'	3 A	3B
16	9	6
17	11	13
18	3	4
19	1	1

Moda

- Oltre alle distribuzioni di frequenza che hanno una sola moda e che si chiamano **distribuzioni unimodali**, si trovano distribuzioni di frequenza che presentano due o più mode; sono denominate **distribuzioni bimodali o plurimodali**.
  - Per esempio, misurando le altezze di un gruppo di giovani in cui la parte maggiore sia formata da femmine e la minore da maschi si ottiene una distribuzione bimodale, con una moda principale ed una secondaria.

Moda

- Esempio di distribuzione bimodale:

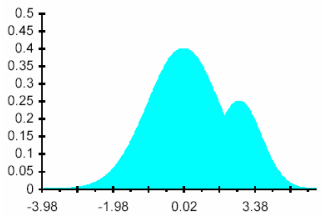


Figura 17. Distribuzione bimodale

Moda

- Esempio di distribuzione bimodale:
  5-6-6-6-6-8-9-9-9-9-15

Moda

Quando la distribuzione dei dati evidenzia due o più mode, si deve sospettare che i dati non siano omogenei, ma formati da altrettanti gruppi con differenti tendenze centrali.

E' pertanto **errato fondare le analisi sulla media generale della distribuzione, poiché non è vera l'assunzione fondamentale che siano dati tratti dallo stesso universo o popolazione con una sola tendenza centrale**.

Moda per dati raggruppati

Requisiti: tutte le classi devono avere la stessa ampiezza

$$\text{Moda} = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} a$$

Dove:
L_1 =confine inferiore della classe avente la massima frequenza, detta classe modale
Δ_1 =differenza tra la frequenza della classe modale e quella della classe precedente
Δ_2 =differenza tra la frequenza della classe modale e quella della classe seguente
a= ampiezza della classe modale

Moda per dati raggruppati

Esempio

classi	Freq.
1-4	3
5-8	16
9-12	7
13-16	2

$$\text{Moda} = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} a$$

Dove:
L_1 =confine inferiore della classe modale = 4.5
Δ_1 =differenza tra la frequenza della classe modale e quella della classe precedente =16-3=13
Δ_2 =differenza tra la frequenza della classe modale e quella della classe seguente= 16-7=9
a= ampiezza della classe modale= 4

Moda= 4.5+(13/(13+9))4=4.5+52/22=6.86

Esercizi

Trovare la moda delle seguenti distribuzioni:

(a) 7-6-7-8-6-7-11-9-8-9-7-5

(b)

x_i	6.12	6.18	7.02	7.05
f_i	57	60	71	62

(c)

Classi di lunghezza	Freq. assolute
25.2-25.4	22
25.5-25.7	58
25.8-26.0	20

Mediana

- **La mediana** è il valore che ***occupa la posizione centrale in un insieme ordinato di dati***.
- E' una **misura robusta**, in quanto **poco influenzata dalla presenza di dati anomali**.
- La sue caratteristiche più importante sono due:
  - è calcolata sul numero di osservazioni; si ricorre al suo uso quando si vuole attenuare l'effetto di valori estremi;
  - in una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.

Mediana

- Per calcolare la mediana di un gruppo di dati, occorre
  1 - disporre i valori in una fila ordinata in modo crescente oppure decrescente e contare il numero totale **n** di dati;
  2 - se il numero (**n**) di dati è dispari, la mediana corrisponde al valore numerico del dato centrale, quello che occupa la posizione **(n+1)/2**;
  3 – se il numero (**n)** di dati è pari, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni **n/2** e **n/2+1**; con poche osservazioni, come mediana viene assunta la media aritmetica di queste due osservazioni intermedie; con molte osservazioni raggruppate in classi, si ricorre talvolta alle proporzioni.

Mediana

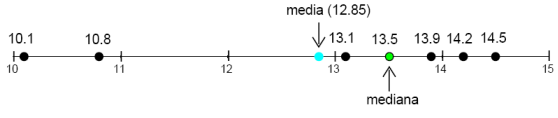
ESEMPIO. Calcolare la mediana nella serie di 7 dati: 13,9 10,1 14,5 13,1 14,2 10,8 18,5.

ESEMPIO. Calcolare la mediana nella serie di 8 dati: 13,9 10,1 13,5 14,5 13,1 14,2 10,8 13,5.

Mediana

ESEMPIO. Calcolare la mediana nella serie di 6 dati: 13,9 10,1 14,5 13,1 14,2 10,8.

Risposta: Il numero di osservazioni è pari e i due valori centrali sono 13,1 e 13,9; la mediana è individuata dalla loro media aritmetica e quindi è uguale a 13,5.



Mediana di dati raggruppati

Per determinare la mediana di questa distribuzione si procede come segue:

(a) Si controlla che le x_i siano in ordine crescente;

(b) si sommano le f_i e si divide per 2, individuando la posizione centrale dei dati. Nel nostro caso (11+17+38+30+45+50)/2=95.5 e quindi la posizione centrale è la 96.

(c) Per trovare quale dato corrisponde alla posizione trovata, si calcolano le frequenze cumulate finche' non si arriva ad una frequenza cumulata maggiore o uguale alla posizione centrale.

x_i	f_i	Freq.cum.
3	11	11
4	17	28
5	38	66
6	30	96
7	45	
8	50	

Mediana di distribuzioni in classi

Vogliamo trovare la mediana di una variabile statistica continua i cui dati sono distribuiti in classi

classi	Freq. assolute	fre.q. cumul.
10-14	2	2
15-19	3	5
20-24	12	17
25-29	10	27

La mediana cade in questa classe dato che 17>27/2

Se supponiamo che i 12 valori nella classe 20-24 siano distribuiti uniformemente all'interno nell'intervallo, avente ampiezza 24.5-19.5=5,
Dobbiamo pensare che essi si trovino a distanza 5/12 l'uno dall'altro. La mediana, ossia il valore che si trova alla posizione (27/2=) 14 è quel valore che nella classe occupa posizione 9=14-5. La mediana quindi è il valore che supera il confine inferiore della classe (19.5) di 9 posizioni ciascuna ampia 5/12

Mediana=19.5+9*5/12=23.25

Mediana di distribuzioni in classi

In formula:

$\text{Mediana}=L_{inf} + (P-F_c)\, a/f$

dove:
1) L_{inf} è il confine inferiore della classe contenente la mediana
2) P-F_c posizione della mediana all'interno della classe (P è la posizione centrale della mediana e F_c è la frequenza cumulata della classe precendente)
3) a è l'ampiezza della classe mediana
4) f frequenza assoluta della classe mediana

Esercizio

- Trovare la mediana della distribuzione in tabella

classi	Freq. Ass.
75-79	5
80-84	7
85-89	14
90-94	7

Quartili

Sono degli indici di della stessa natura della mediana.
Ne sono definiti tre, il primo quartile, il secondo quartile e il terzo quartile.

I quantili ripartiscono la distribuzione in 4 parti di pari frequenza, dove Ogni parte contine la stessa frazione di osservazioni.

Il **primo quartile** è definito come il numero q_1 per il quale il 25% dei dati statistici è minore o uguale a q_1.

Il **secondo quartile** è definito come il numero q_2 per il quale il 50% dei dati statistici è minore o uguale a q_2. Il secondo quartile corrisponde, per definizione alla mediana

Il **terzo quartile** è definito come un numero q_3 per il quale il 75% dei dati statistici è minore o uguale a q_3.

Quartili

Riconsiderando i voti di Anna (30, 30, 28, 27, 26) e Stefano (21, 30, 30, 30, 30), abbiamo:

studente	minimo	q_1	mediana	q_3	massimo
Anna	26	27	28	30	30
Stefano	21	30	30	30	30

Percentili

- Se, invece di dividere I dati statistici in quattro parti, li dividiamo in 100 parti, definiremmo i percentili.
- Sono utilizzati in campo medico
- Esempio: dire che il peso di un neonato ricade nel 35-percentile significa che, in linea di massima, il 35% dei neonati ha un peso inferiore ad esso, ed il rimanente 65% ha un peso superiore.

Misure di dispersione

Misure di dispersione o indici di variabilità

Gli indici statistici di posizione riassumano sinteticamente una lista di dati, ma fanno perdere informazione. Infatti conoscendo l'indice di posizione non sappiamo se i dati sono concentrati intorno ad esso oppure se sono dispersi.

Per quantificare quanto di dati sono distanti tra loro (quindi quanto variano rispetto alla media) sono state definite le misure di dispersione (o variabilità).

Campo di variazione

È il più semplice indice di variazione.

Il **campo di variazione** di una distribuzione è la differenza tra il dato più grande e quello più piccolo della distribuzione:

$$C = x_{max} - x_{min}$$

Questo indice è abbastanza grossolano non dicendo nulla sulla variabilità dei dati intermedi.

Campo di variazione

Esempio 1- il campo di variazione della seguente distribuzione:
25 – 26 – 28 – 29 – 30 – 32

è $C_1 = 32 – 25 = 7$

Esempio 2- il campo di variazione della seguente distribuzione:
25 – 30 – 30 – 31 – 31 – 32

è $C_2 = 32 – 25 = 7$

Guardando le due distribuzioni cosa possiamo concludere?

Scarto semplice medio assoluto dalla media

Un modo di tener conto della variabilità dei dati è quello di misurare quanto ciascun dato x_i si discosti dal valor medio.

Si chiama scarto semplice medio assoluto e si indica con s_m la media aritmetica dei valori assoluti degli scarti $x_i - \bar{x}$. In simboli:

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

Perché usiamo il valore assoluto?

Scarto semplice medio assoluto dalla media

Come diventa questa formula di s_m per dati x_i raggruppati in frequenze f_i ?

Se ad x si sostituisce la mediana dei dati, si ottiene un altro indice di variabilità chiamato scarto semplice medio assoluto dalla mediana.

Questo indice ha la proprietà di essere minimo rispetto agli altri scarti semplici medi assoluti. Perché?

Varianza campionaria

Un altro accorgimento che permette di eliminare i segni degli scarti è quello di elevare al quadrato gli scarti alla media.

Otteniamo in questo modo un nuovo indice di variabilità chiamato varianza.

è una devianza media o devianza rapportata al numero di osservazioni.

La **varianza della popolazione**, il cui simbolo è s^2 , è ottenuta dalla formula:

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

Dove:

- n è il numero di osservazioni
- $\bar{x}$ è la media dei dati osservati
- x_i è l'i-esimo dato statistico osservato

Varianza campionaria

Alcuni statistici definiscono la varianza campionaria come

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

Ovviamente, quando **n** è grande le differenze tra le due formule sono minime; quando **n** è piccolo, le differenze sono sensibili.

Si usa la prima quando il numero delle osservazioni è elevato, la seconda altrimenti. In genere possiamo assumere equivalenti le due espressioni per n>30. (questa osservazione andrebbe fatta oer tutte le formule che seguono)

Varianza campionaria

Applicando le proprietà della sommatoria otteniamo una formula che ci permette di snellire I calcoli della varianza:

$$s_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2) =$$

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

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

Varianza campionaria

Quale è la formula della varianza in caso di dati raggruppati?

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

n = numero di raggruppamenti

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

Deviazione standard o scarto quadratico medio

É la radice quadrata della varianza, in formule:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Dove:

- n è il numero di osservazioni
- $\bar{x}$ è la media dei dati osservati
- x_i è l'i-esimo dato statistico osservato

E' una misura di distanza dalla media e quindi ha sempre un valore positivo.
E' una misura della dispersione della variabile statistica intorno alla media.

Deviazione standard o scarto quadratico medio

E' una misura di distanza dalla media e quindi ha sempre un valore positivo.

E' una misura della dispersione della variabile casuale intorno alla media.

In generale si preferisce usare come indice di variabilità la deviazione standard perche' essa mantiene l'unità di misura dei dati statistici.

Se i dati statistici sono relatici alle altezze di individui espresse in cm, la varianza è espressa in cm^2 mentre la deviazione standard mantiene come unità di misura il cm come I dati osservati.

Esempio

consideriamo i voti di due studenti: Anna (30, 30, 28, 27, 26) e Stefano (21, 30, 30, 30, 30). Entrambi hanno la stessa media dei voti (media=28.2) ma Stefano "sembra" essere più bravo.

Calcolando la deviazione standard otteniamo σ(Anna)= 1.6 σ(Stefano)=3.6.

Cosa significa?

Significa che i voti di Anna sono piu' concentrati (vicini) rispetto a quelli di Stefano.

Convenzioni di notazioni

Alcuni testi usano indicare la varianza con il simbolo σ^2 e la deviazione standard con σ.

Esercizio

Calcolare media, varianza e deviazione standard di : 9, 6, 7, 9, 8, 8

Media= (9+6+7+9+8+8)/6=(9*2+8*2+6+7)/6=47/6=7.833,
Varianza=
deviazione standard=

Esercizi a pagina 299