Metodi Statistici per la Bioingegneria

Prof. Alessandra Bertoldo

alessandra.bertoldo@unipd.it
alessandra.bertoldo@dei.unipd.it
Ufficio: terzo piano del DEI/A

VARIABILI ALEATORIE E STATISTICA DESCRITTIVA

La probabilità nel linguaggio corrente

- è probabile che fra poco piova;
- con questo titolo di studio vi sono buone probabilità di trovare lavoro;
- è probabile che l'incendio sia d'origine dolosa;
- ho poche probabilità di superare l'esame.

Utilizziamo frequentemente il termine probabilità quando ci riferiamo a situazioni incerte, a fenomeni che possono o non verificarsi.

Perché parliamo di probabilità?
Conosciamo e incontriamo spesso in natura, in economia, nella nostra vita quotidiana, fenomeni che sembrano avere caratteristiche di casualità. Molto spesso questa casualità è solo apparente e potrebbe essere dovuta a una serie di fattori che, pur essendo deterministici, possono non essere completamente noti o avere una spiegazione troppo complessa.

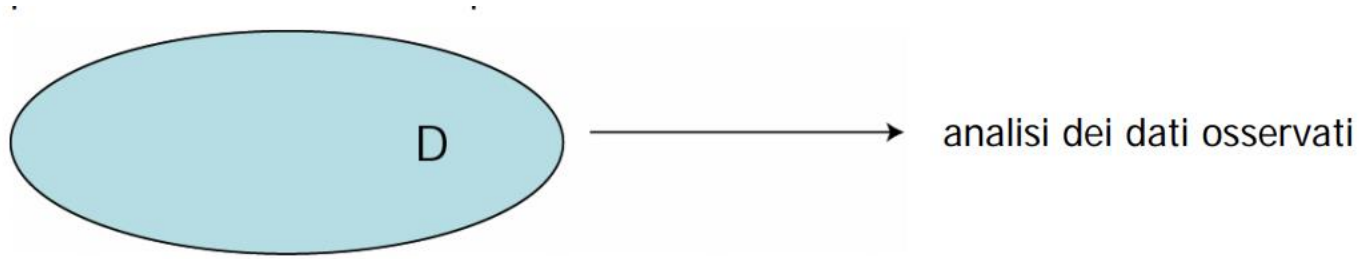
ESEMPIO 1. Nel lancio di una moneta il singolo risultato è incerto; in effetti sono incerte le velocità iniziali di traslazione e di rotazione, la situazione termica ambientale, è incerto il piano d'arrivo, e così via. Teoricamente si potrebbe trasformare il lancio della moneta in un fenomeno dal risultato "praticamente" certo se si tenesse conto di tutte queste condizioni. Però è più comodo adottare un altro punto di vista che rinuncia alla descrizione analitica del fenomeno e attribuisce i diversi esiti dell'esperienza a una variabilità accidentale, intesa come sintesi delle diverse condizioni che si rinunciano a specificare.

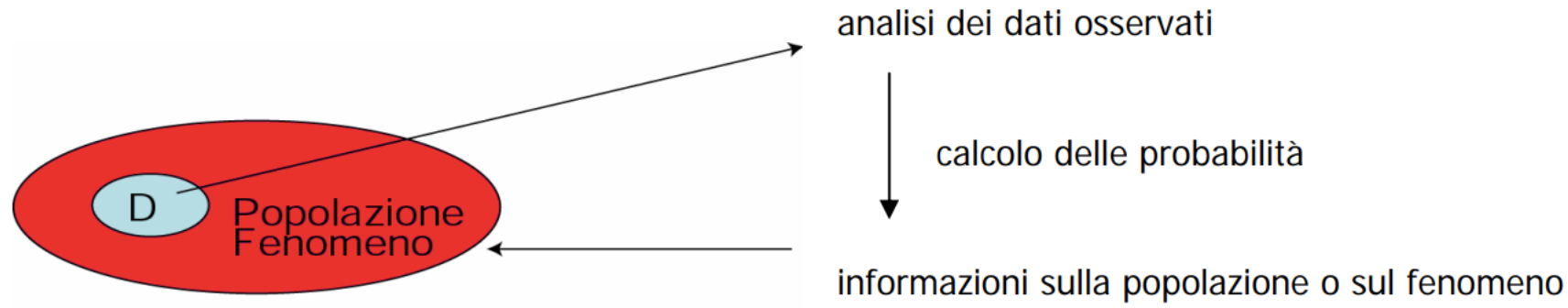
Da: M.P. Rogantin

Possiamo pensare che la probabilità misuri la nostra fiducia che un evento si possa verificare: se siamo sicuri che qualcosa accadrà gli assegneremo probabilità 1 (o equivalentemente 100%); se siamo certi che sia impossibile gli assegneremo valore 0. Per tutti gli altri casi, dovremo capire quale valore compreso fra 0 e 1 sia ragionevole assegnare.

Quindi possiamo avere:

In generale, per lo studio di un fenomeno che manifesta casualità, è necessaria l'osservazione ripetuta dello stesso fenomeno nelle identiche condizioni; per identiche condizioni si intende appunto che i fattori controllabili che influenzano il fenomeno assumano le stesse caratteristiche; tutti i fattori non controllabili potranno essere differenti e saranno quelli che generano la casualità del fenomeno.



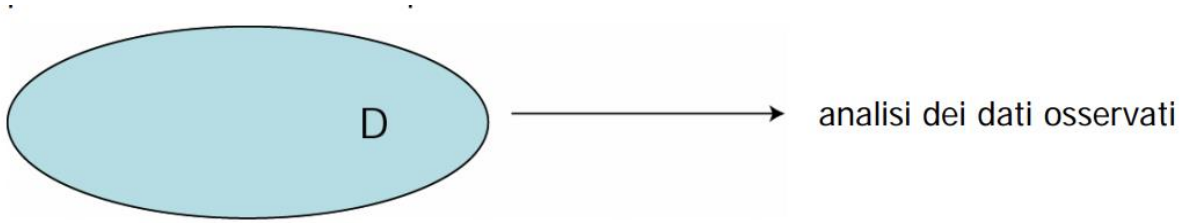


STATISTICA

Dai risultati di un esperimento si determinano alcune caratteristiche della popolazione

Dalle caratteristiche note della popolazione si prevede il risultato di un altro esperimento

STATISTICA DESCRITTIVA



Statistica (= studio delle cose dello Stato) è una disciplina che ha come fine lo studio quantitativo di un particolare fenomeno collettivo (popolazione) in condizioni di fenomeno aleatorio ossia di non completa conoscenza di esso o parte di esso.

La statistica descrittiva è una branca della Statistica e ha come scopo quello di **sintetizzare** i dati attraverso i suoi strumenti grafici (**diagrammi a barre, a torta, istogrammi, boxplot**) e indici (**indicatori statistici, indicatori di posizione come la media, di dispersione, come la varianza e la concentrazione, di correlazione, di forma, come la curtosi e la skewness, ecc.**) che descrivono gli aspetti salienti dei dati osservati, formando così il contenuto statistico.

INDICI DI POSIZIONE: LE MEDIE

Medie analitiche: calcolare tramite operazioni algebriche sui valori del carattere ➔ solo per caratteri quantitativi

Media di posizione: non richiedono operazioni algebriche ➔ anche per caratteri qualitativi

Medie analitiche	Medie di posizione
Media aritmetica	Mediana
Media geometrica	Moda
Media troncata	Percentili

La media aritmetica

- Per una *distribuzione unitaria* di un carattere *quantitativo* di *n* termini, la **media aritmetica** è definita come:

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

Esempio

- Distribuzione del voto in Statistica per un gruppo di 6 studenti:

Unità statistica (*i*)	Voto (x_i)
1	27
2	21
3	28
4	30
5	21
6	27
Totale	**154**

$$\bar{x} = \frac{1}{6} \cdot 154 = 25{,}667$$

- In alcune situazioni si ha una ***distribuzione ponderata*** in cui a ogni modalità viene associato un peso che ne quantifica l'importanza

Unità (*i*)	Modalità (x_i)	Peso (w_i)
1	x_1	w_1
2	x_2	w_2
…	…	…
n	x_n	w_n

- Per calcolare la media aritmetica, il peso viene trattato come una frequenza

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

Media ponderata o pesata

Proprietà della media aritmetica

La somma algebrica della somma degli scarti dalla media è nulla:

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

Esempio

Unità statistica (*i*)	Voto in stat. (x_i)	Scarto
1	27	1
2	21	-5
3	28	2
4	30	4
5	23	-3
6	27	1
Totale	**156**	**0**

$$\bar{x} = \frac{1}{6} \cdot 156 = 26$$

La media minimizza la distanza al quadrato di ogni modalità da una costante

$$\sum_{i=1}^{n}(x_i - c)^2 \quad \text{è minimo per} \quad c = \bar{x}$$

Esempio

Unità statistica (i)	Voto in stat. (x_i)	Scarto $x_i - \bar{x}$	Scarto $(x_i - \bar{x})^2$
1	27	1	1
2	21	-5	25
3	28	2	4
4	30	4	16
5	23	-3	9
6	27	1	1
Totale	**154**	**0**	**56**

$$\bar{x} = \frac{1}{6} \cdot 156 = 26$$

- ***Proprietà 5*** *(di internalità)*: La media è sempre compresa tra il minimo e il massimo della distribuzione

$$x_1 \leq \bar{x} \leq x_k$$

- ***Proprietà 6*** (*invarianza rispetto a trasformazioni lineari*): se a ogni termine della distribuzione viene applicata la trasformazione $aX + b$, allora la media sarà pari a

$$a\bar{x} + b$$

La media geometrica

- Per una *distribuzione unitaria* di un carattere *quantitativo* di *n* termini, la **media geometrica** è definita come:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{1=1}^{n} x_i}$$

- Viene usata per sintetizzare dati che ha senso moltiplicare fra loro o per riassumere distribuzioni che hanno andamento geometrico

- Si applica per determinare un tasso di incremento / decremento medio (prezzi dei prodotti, andamento della popolazione, ecc)

- Ad esempio: quando i dati cambiano velocemente da un anno all'altro vengono chiamati valori geometrici.

- La media aritmetica è normalmente più grande della media geometrica

Esempio

- In un determinato punto vendita si è osservato:

Anni	2005	2006	2007	2008
Vendite (milioni di euro)	207	189	246	298

Si vuole calcolare la variazione media nelle vendite

- Bisogna innanzitutto calcolare le variazioni annue:

Anni	2005	2006	2007	2008
Variazioni $x_i = \frac{V_i}{V_{i-1}}$	--	0,913	1,302	1,21

$$\bar{x}_g = \sqrt[3]{0{,}913 \cdot 1{,}302 \cdot 1{,}211} = 1{,}129$$

- La media aritmetica sarebbe stata invece:

$$\bar{x}_a = (0{,}913 + 1{,}302 + 1{,}211)/3 = 1{,}142$$

Perché la media aritmetica non è adeguata in questi casi?

Supponiamo che V_0 siano le vendite iniziali. Applicando le variazioni x_1, x_2, x_3 otteniamo:

$V_1 = V_0 \cdot x_1$
$V_2 = V_1 \cdot x_2 = V_0 \cdot x_1 \cdot x_2$
$V_3 = V_2 \cdot x_3 = V_0 \cdot x_1 \cdot x_2 \cdot x_3$

La variazione media è quella variazione costante, che applicata di anno in anno, deve restituire il valore corretto delle vendite a fine periodo, dato il valore iniziale.

$$V_3 = V_0 \cdot \bar{x} \cdot \bar{x} \cdot \bar{x} = V_0 \cdot \bar{x}^3$$

Sostituendo la media aritmetica e la media geometrica si ha:

$$V_0 \cdot \bar{x}_a^3 = 207 \cdot 1{,}142^3 = 308{,}03 \qquad V_0 \cdot \bar{x}_g^3 = 207 \cdot 1{,}129^3 = 298$$

Proprietà della media geometrica

- ***Proprietà*** (*invarianza rispetto a cambiamenti di scala*): se a ogni termine della distribuzione viene applicata la trasformazione *aX*, allora la media geometrica sarà pari a

$$a\bar{x}_g$$

- ***Proprietà*** : La media geometrica non è mai superiore alla media aritmetica

$$\bar{x}_g \leq \bar{x}_a$$

per qualsiasi distribuzione

Proprietà della media geometrica

- ***Proprietà*** : Il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi. Quindi, ad esempio, la media geometrica può essere calcolata come

$$\bar{x}_g = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right] \qquad \text{(distribuzione unitaria)}$$

$$\bar{x}_g = \exp\left[\frac{1}{n}\sum_{j=1}^{k} n_j \log(x_j)\right] \qquad \text{(distribuzione di frequenze)}$$

Quando non utilizzare la media geometrica:

- Valori nulli nella distribuzione La presenza di uno zero nella distribuzione annulla il prodotto tra gli elementi e azzera la media.
- Valori negativi nella distribuzione La presenza dei numeri negativi nella distribuzione potrebbe determinare un prodotto negativo sotto radice.

La mediana

Data una distribuzione secondo un carattere *qualitativo ordinato* o *quantitativo*, la **mediana** (M_e) è la modalità del carattere che divide il collettivo in due gruppi di uguale numerosità

Per calcolare la mediana di una *distribuzione unitaria* di un carattere quantitativo di n termini

1. si ordinano le modalità in modo non decrescente:

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

2. se *n* è dispari $\rightarrow$ $M_e = x_{(n+1)/2}$

   se *n* è pari $\rightarrow$ $M_e = \left(x_{n/2} + x_{n/2+1}\right)/2$

(solo per caratteri quantitativi)

Paziente	glicemia (mg/100cc)
p1	100
p2	112
p3	83
p4	103
p5	98
p6	82
p7	90
p8	98
p9	108
p10	121
p11	82
p12	113
p13	101
p14	94
p15	101
p16	93
p17	93
p18	109
p19	108
p20	108
p21	101
p22	83
p23	101
p24	110
p25	99

Paziente	glicemia (mg/100cc)
p1	100
p2	112
p3	83
p4	103
p5	98
p6	82
p7	90
p8	98
p9	108
p10	121
p11	82
p12	113
p13	101
p14	94
p15	101
p16	93
p17	93
p18	109
p19	108
p20	108
p21	101
p22	83
p23	101
p24	110
p25	99

Ordino in modo decrescente (o ascendente)

Paziente	glicemia (mg/100cc)
p10	121
p12	113
p2	112
p24	110
p18	109
p9	108
p19	108
p20	108
p4	103
p13	101
p15	101
p21	101
p23	101
p1	100
p25	99
p5	98
p8	98
p14	94
p16	93
p17	93
p7	90
p3	83
p22	83
p6	82
p11	82

12 pazienti

12 pazienti

N pari, la mediana viene calcolata mediando i due elementi che si trovano in posizione centrale, dopo che i dati sono stati ordinati

N dispari, si sceglie il valore che si trova in posizione centrale

Mediana = 101

Media (media aritmetica) = 99.64

Proprietà della mediana

•***Proprietà*** La mediana minimizza la distanza di ogni modalità da una costante

$$\sum_{i=1}^{n} |x_i - c| \quad \text{è minimo per} \quad c = M_e$$

- ***Proprietà*** *(di internalità)*: La mediana è sempre compresa tra il minimo e il massimo della distribuzione

$$x_1 \leq M_e \leq x_k$$

- ***Proprietà*** (*invarianza rispetto a trasformazioni lineari*): se a ogni termine della distribuzione viene applicata la trasformazione *aX* + *b*, allora la mediana sarà pari a

$$aM_e + b$$

Vantaggio nell’uso della mediana:

non è influenzata dalle osservazioni aberranti o estreme (outliers)

I percentili

- I percentili sono quei valori che dividono la distribuzione in 100 parti uguali.
- I percentili più usati sono il 25° (Q_1), il 50° (M_e) e il 75° (Q_3).

- La mediana è il dato che delimita il primo 50% dei dati (ordinati) dai rimanenti dati,
- se p è un numero tra 0 e 100, il percentile di ordine p (o p° percentile, se p è intero) è il dato che delimita il primo p% dei dati (ordinati) dai rimanenti dati

(è usato anche il concetto di quartile: il 1º quartile è il dato che delimita il primo quarto dei dati dai rimanenti, il 2º è quello che ne delimita i primi due quarti, e coincide con la mediana, e il 3º è quello che ne delimita i primi tre quarti)

Paziente	glicemia (mg/100cc)
p1	100
p2	112
p3	83
p4	103
p5	98
p6	82
p7	90
p8	98
p9	108
p10	121
p11	82
p12	113
p13	101
p14	94
p15	101
p16	93
p17	93
p18	109
p19	108
p20	108
p21	101
p22	83
p23	101
p24	110
p25	99

Paziente	glicemia (mg/100cc)
p1	100
p2	112
p3	83
p4	103
p5	98
p6	82
p7	90
p8	98
p9	108
p10	121
p11	82
p12	113
p13	101
p14	94
p15	101
p16	93
p17	93
p18	109
p19	108
p20	108
p21	101
p22	83
p23	101
p24	110
p25	99

Ordino in modo decrescente (o ascendente)

Paziente	glicemia (mg/100cc)
p10	121
p12	113
p2	112
p24	110
p18	109
p9	108
p19	108
p20	108
p4	103
p13	101
p15	101
p21	101
p23	101
p1	100
p25	99
p5	98
p8	98
p14	94
p16	93
p17	93
p7	90
p3	83
p22	83
p6	82
p11	82

12 pazienti

12 pazienti

Mediana = 101 = 50% percentile

N pari, la mediana viene calcolata mediando i due elementi che si trovano in posizione centrale, dopo che i dati sono stati ordinati

N dispari, si sceglie il valore che si trova in posizione centrale

Paziente	glicemia (mg/100cc)
p1	100
p2	112
p3	83
p4	103
p5	98
p6	82
p7	90
p8	98
p9	108
p10	121
p11	82
p12	113
p13	101
p14	94
p15	101
p16	93
p17	93
p18	109
p19	108
p20	108
p21	101
p22	83
p23	101
p24	110
p25	99

Ordino in modo decrescente (o ascendente)

Paziente	glicemia (mg/100cc)
p10	121
p12	113
p2	112
p24	110
p18	109
p9	108
p19	108
p20	108
p4	103
p13	101
p15	101
p21	101
p23	101
p1	100
p25	99
p5	98
p8	98
p14	94
p16	93
p17	93
p7	90
p3	83
p22	83
p6	82
p11	82

12 pazienti/2 = 6 ***Q3 = 75% percentile = 108***

Mediana = 101 = 50% percentile

12 pazienti ***Q1 = 25% percentile = 93***

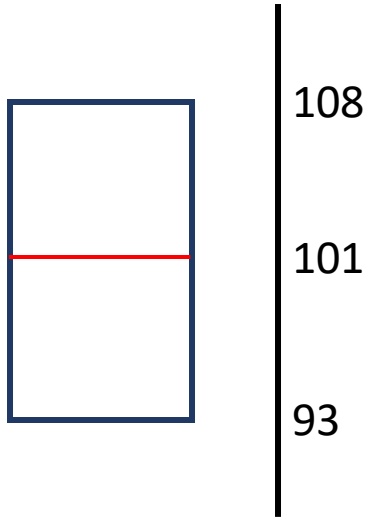
I BOX-PLOT

Una rappresentazione grafica che si basa sulla definizione dei quantili è il box-plot.
Permette di descrivere la variabile in maniera sintetica ed è molto utile per confrontare sottogruppi di dati.

L'idea è quella di individuare con una "scatola" le osservazioni centrali e con dei "baffi" o code uscenti dalla scatola le osservazioni più estreme.

Per costruirlo si disegna una scatola tra i valori Q1 e Q3. Con una linea verticale si individua la mediana (Q2).

Riferendoci ai dati precedenti:



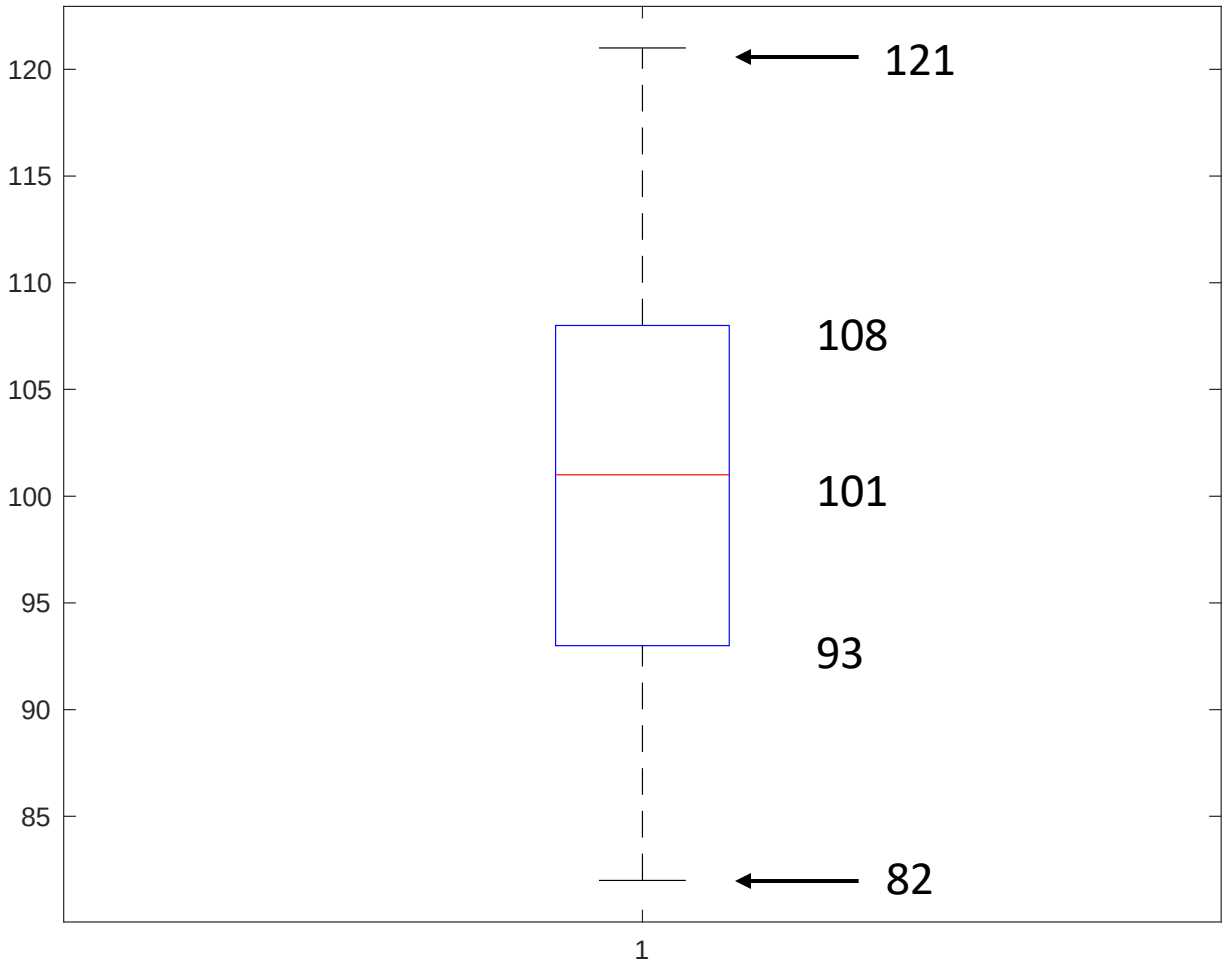
I BOX-PLOT

Si disegnano poi i baffi che sono lunghi al più una volta e mezza la distanza interquartile e terminano in corrispondenza del dato più lontano dalla scatola inferiore a tale valore.
I valori limite per i baffi sono quindi:

Distanza interquartile = 108-93 = 15
Una volta e mezza la distanza interquartile = 22.5
Limite inferiore = 93-22.5 = 69.5
Limite superiore = 108+22.5 = 130.5

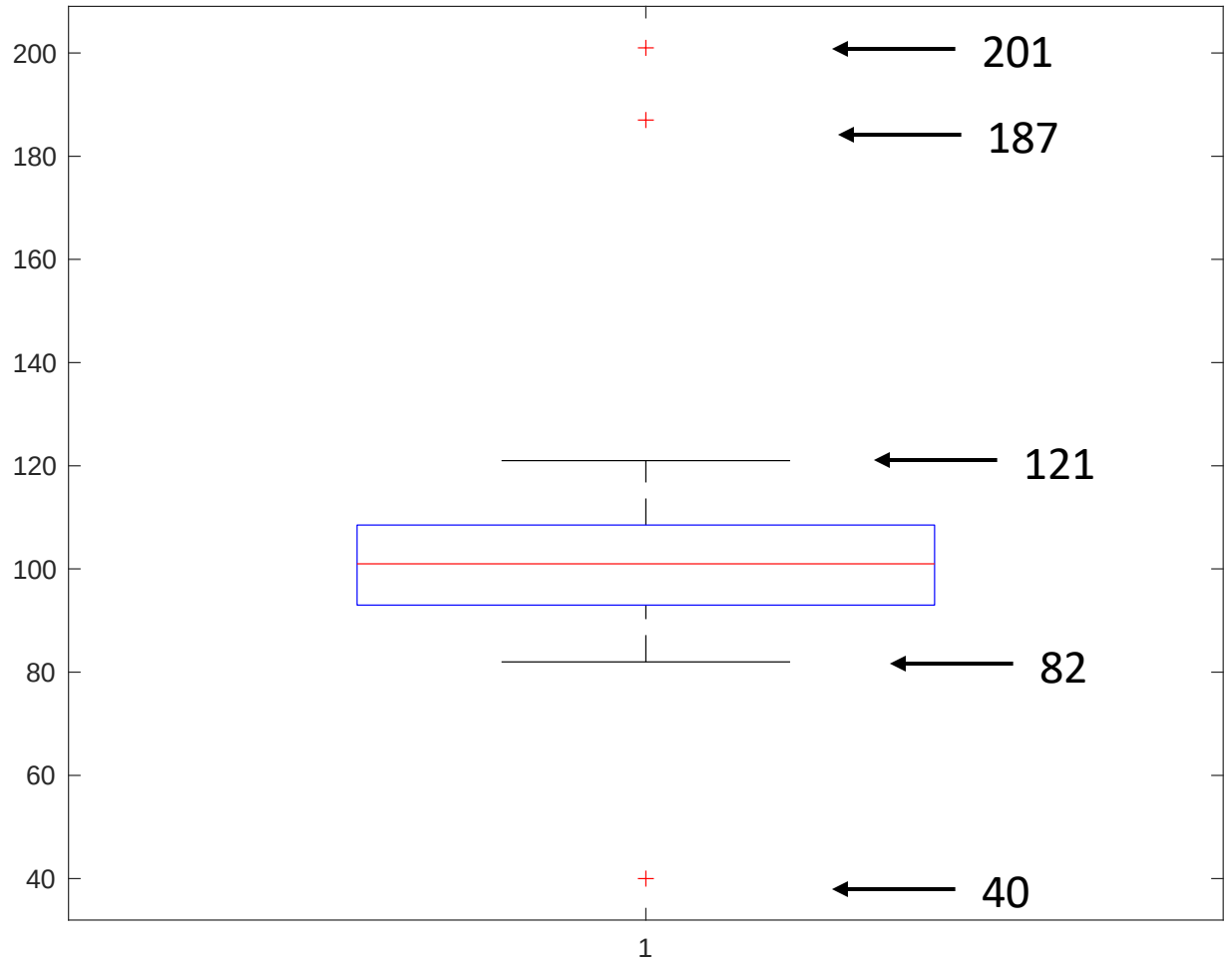
Il valore minimo dei nostri dati di esempio è: 82 e non supera il valore del limite inferiore
Il valore massimo dei nostri dati di esempio è: 121 e non supera il valore del limite superiore

Quindi il box-plot finale è:


120
115
110
105
100
95
90
85
1
121
108
101
93
82

Paziente	glicemia (mg/100cc)
p26	40
p11	82
p6	82
p22	83
p3	83
p7	90
p16	93
p17	93
p14	94
p5	98
p8	98
p25	99
p1	100
p13	101
p15	101
p21	101
p23	101
p4	103
p19	108
p20	108
p9	108
p18	109
p24	110
p2	112
p12	113
p10	121
p28	187
p27	201

I valori che rimangono al di fuori dei limiti si individuano con asterischi, croci, cerchi….



LA MODA

Si chiama **moda** (o moda campionaria) degli n elementi $x_1, \ldots x_n$ l'elemento (o gli elementi) a cui corrisponde la massima frequenza assoluta.

È facile dedurre che la **moda** campionaria non è influenzata da valori estremi e può essere usata anche per dati non numerici, cioè dati qualitativi. Si osservi che la moda campionaria può non esistere o non essere unica.

RIASSUNTO:

media aritmetica
mediana
moda
media geometrica

Esempio 1

Ai 23 alunni di una classe è stato chiesto di indicare il tempo impiegato a raggiungere la scuola. Le risposte sono riportate nella seguente tabella

Alunni	Tempo
A	20
B	12
C	3
D	7
E	5
F	6
G	15
H	5
I	10
L	4
M	7
N	5
O	6
P	9
Q	5
R	6
S	7
T	10
U	7
V	10
Z	5
X	18
Y	2
Totale	184

media (campionaria) = 8, significa che ciascun alunno impiegherebbe 8 minuti per arrivare a scuola se tutti impiegassero lo stesso tempo

moda (campionaria) **= 5**, significa che il maggior numero di alunni impiega 5 minuti per arrivare a scuola

mediana (campionaria) **= 7**, significa che circa la metà degli alunni impiega meno di 7 minuti per raggiungere la scuola e circa la metà ne impiega più di 7.

Esempio 2

A quale misura di tendenza centrale ci riferiamo?

- Il proprietario di una ditta afferma "Lo stipendio mensile nella nostra ditta è 2.700 euro"
- Il sindacato dei lavoratori dice che “lo stipendio medio è di 1.700 euro”.
- L'agente delle tasse dice che “lo stipendio medio è stato di 2.200 euro”.

Queste risposte diverse sono state ottenute tutte dai dati della seguente tabella.

Stipendio mensile	N° di lavoratori
1.300	2
1.700	22
2.200	19
2.600	3
6.500	2
9.400	1
23.000	1

Media aritmetica= lire 2.700

Mediana = lire 2.200

Moda = lire 1.700

- La **media aritmetica** indica che, se il denaro fosse distribuito in modo che ciascuno ricevesse la stessa somma, ciascun dipendente avrebbe avuto 2.700 euro

- La **moda** ci dice che la paga mensile più comune è di 1.700.euro

- La moda si considera spesso come il valore tipico dell'insieme di dati poiché è quello che si presenta più spesso. **Non tiene però conto degli altri valori** e spesso in un insieme di dati vi è **più di un valore** che corrisponde alla definizione di moda.

- La **mediana** indica che circa metà degli addetti percepiscono meno di 2.200.euro, e metà di più.

- La mediana **non è influenzata dai valori estremi** eventualmente presenti ma solo dal fatto che essi siano sotto o sopra il centro dell'insieme dei dati.

ISTOGRAMMA

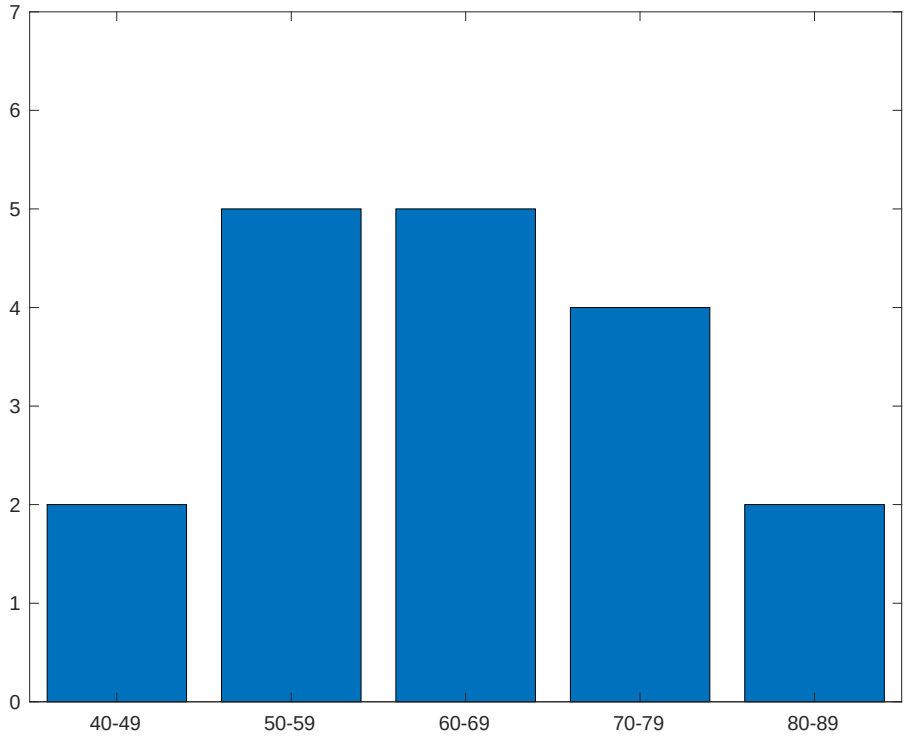
detto anche ortogramma, è un diagramma che fornisce una rappresentazione di un insieme di dati statistici mediante un grafico a barre. Gli istogrammi possono essere rappresentati mediante barre orizzontali (istogramma a barre orizzontali) o verticali (istogramma a barre verticali)

Si suddivide l'intervallo in cui variano i dati in classi (preferibilmente di uguale ampiezza) e si assegna ogni osservazione rilevata alla classe corrispondente. La scelta del numero di classi non è indifferente: troppo poche appiattiscono il grafico fino a renderlo

insignificante; troppe classi introducono tra le barre oscillazioni eccessive, che potrebbero distruggere l'eventuale "regolarita'" dell'istogramma. L'istogramma si disegna come i diagrammi a barre per le variabili qualitative, ma facendo attenzione che i "rettangoli" verticali devono essere adiacenti ed avere come vertici i punti che separano le classi.

pz1	pz2	pz3	pz4	pz5	pz6	pz7	pz8	pz9	pz10	pz11	pz12	pz13	pz14	pz15	pz16	pz17	pz18
47	47	50	50	56	56	57	62	65	65	68	68	70	73	73	73	85	85

peso	Conteggi	frequenze
40-49	2	11,1%
50-59	5	27,5%
60-69	5	27,5%
70-79	4	22,8%
80-89	2	11,1%



INDICI DI VARIABILITA’

L’utilizzo di una media permette di sintetizzare efficacemente l’informazione contenuta in una distribuzione statistica dal punto di vista dell’intensità del valore misurato.

Tuttavia la sintesi può essere eccessiva, nel senso si possono perdere informazioni su altre caratteristiche fondamentali come la variabilità.

Esempio

- Consideriamo le distribuzioni secondo il numero di figli in due collettivi diversi di 25 famiglie ciascuno.

Popolazione 1

N. Figli (x_j)	Frequenze (n_j)
0	1
1	4
2	15
3	3
4	2
Totale	25

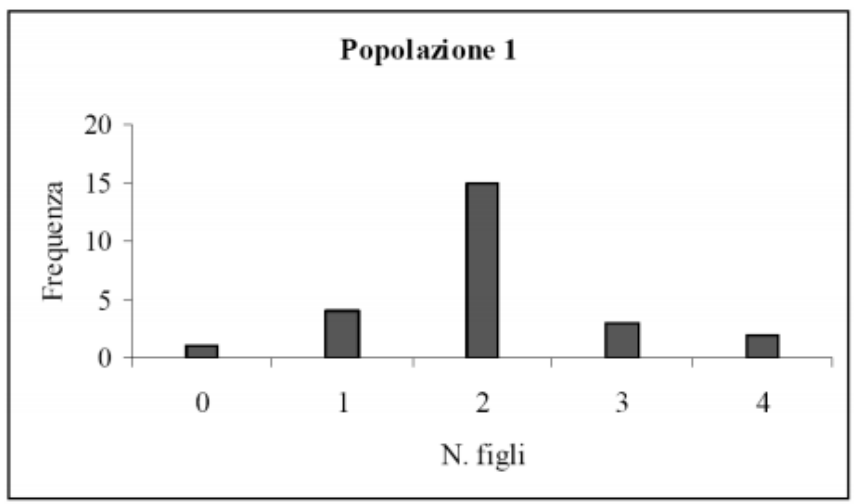
Popolazione 2

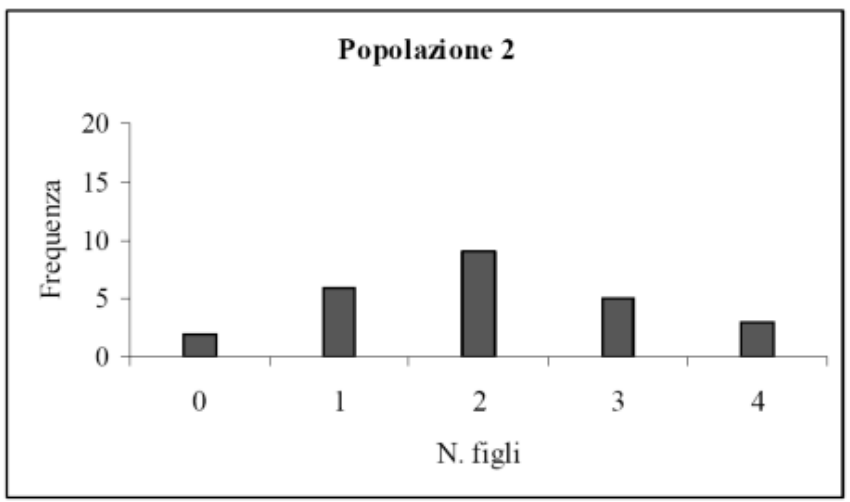
N. Figli (x_j)	Frequenze (n_j)
0	2
1	6
2	9
3	5
4	3
Totale	25

$$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{25}51 = 2{,}04$$

$$\bar{x}_2 = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{25}51 = 2{,}04$$

- Entrambe le distribuzioni hanno media 2,04 ma, come è possibile dedurre dai grafici, sono molto diverse: la prima assume delle modalità molto più concentrate attorno alla media e quindi ha minore variabilità.





INDICI DI VARIABILITA’

- Campo di variazione (range);
- Devianza;
- Varianza;
- Deviazione Standard;
- Coefficiente di variazione (variabilità relativa).

IL CAMPO DI VARIAZIONE O RANGE

DEFINIZIONE: il campo di variazione o range corrisponde alla differenza fra la modalità più piccolo e la modalità più grande della distribuzione di valori

$$R = X_{max} - X_{min}$$

Limiti:

- È un valore troppo influenzato dai valori estremi
- Tiene conto dei due soli valori estremi, trascurando tutti gli altri

Occorre allora un indice di dispersione che consideri tutti i dati (e non solo quelli estremi), confrontando questi con il loro valor medio.

1ª idea →

$$\sum_{i=1}^{n} (x_i - \overline{x})$$

2ª idea →

$$\sum_{i=1}^{n} |x_i - \overline{x}|$$

3ª idea →

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

- Gli indici di variabilità possono essere basati:
  - sullo *scostamento da una media*;
  - sulla *differenze tra statistiche d'ordine*.

Scostamenti da una media	**Differenze tra statistiche d'ordine**
Devianza	Campo di variazione
Varianza	Differenza interquartilica
Deviazione standard	
Coefficiente di variazione	
Scostamento semplice medio dalla mediana	

Devianza e varianza

- Per una distribuzione di *n* valori quantitativi, la **devianza** è definita come:

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

- La **varianza** è normalmente preferita alla devianza e si ottiene come:

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

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

➔ Varianza campionaria corretta

Deviazione standard e coefficiente di variazione

- La ***deviazione standard*** (o ***scostamento quadratico medio***) è l'indice di variabilità più utilizzato in quanto è espresso nella stessa unità di misura del carattere. Si ottiene come:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \left(\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

- Nel caso in cui la distribuzione abbia media aritmetica positiva, il ***coefficiente di variazione*** si calcola come (normalmente in percentuale):

$$CV = \frac{\sigma}{\bar{x}}100$$

Proprietà

- **Proprietà 1**: gli indici D, σ^2 e σ sono sempre non negativi e assumono il valore minimo (0) se e solo se tutte le modalità della distribuzione sono uguali tra loro.

- **Proprietà 2**: la devianza può essere calcolata come (*formula semplificata*)

$$D = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad \text{(distribuzione unitaria)}$$

$$D = \sum_{j=1}^{k} x_j^2 n_j - n\bar{x}^2 \qquad \text{(distribuzione di frequenze)}$$

che ha vantaggi nel calcolo anche della varianza e della deviazione standard

- **Proprietà 3**: se a ogni termine della distribuzione viene applicata la trasformazione $aX + b$, allora gli indici di variabilità cambieranno nel modo seguente:

Devianza	-------->	$a^2 D$
Varianza	-------->	$a^2 \sigma^2$
Deviazione standard	-------->	$\lvert a \rvert \sigma$

Scostamenti semplici medi

- Per una *distribuzione unitaria* di un *carattere quantitativo*, lo ***scostamento semplice medio dalla media aritmetica*** è definito come

$$S_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

- Lo ***scostamento semplice medio dalla mediana*** si ottiene sostituendo la mediana alla media aritmetica:

$$S_{Me} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Me\right|$$

Come calcolare la media e la varianza con dati raggruppati:

Se i dati statistici non sono grezzi ma sono raggruppati in una tabella di frequenze, la formula per calcolare la media e la varianza sono:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j f_j$$

$$\sigma^2 = \sum_{j=1}^{k} (x_j - \bar{x})^2 f_j$$

Esempio:

anni	studenti
15	4
16	12
17	7
totale	23

k =3
f1 = 4, f2 = 12, f3 = 7
x1 = 15, x2 = 16, x3 = 17
n = 23

$$\bar{x} = \frac{371}{23} = 16.13\ anni$$

$$\sigma^2 = \sum_{j=1}^{k} (x_j - \bar{x})^2 f_j = 10.6\ anni^2$$

ERRORE STANDARD (della media aritmetica)

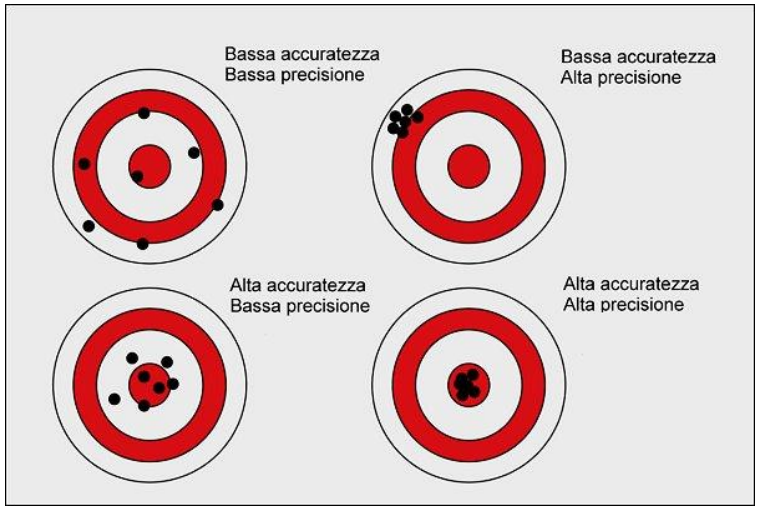
In generale è impossibile acquisire i valori di ogni singolo individuo (o elemento) di una popolazione. Si ricorre quindi allo studio di una campione di n individui (elementi) sperando che il campione sia rappresentativo della popolazione (ad esempio: la media aritmetica del campione è uguale alla media aritmetica della popolazione).

Tutti i metodi statistici sono costruiti sull'assunto che gli individui testati nel campione rappresentino un *campione casuale* della popolazione non interamente osservata.

La media (aritmetica) e deviazione standard (ad esempio) calcolate da un campione n scelto con procedura casuale sono pensate essere stime della media e deviazione standard dell'intera popolazione.

Per quantificare l'*accuratezza* di queste stime, possiamo calcolare i loro *errori standard*.

Si possono calcolare gli errori standard per ogni indice visto, ma ci focalizzeremo sull'errore standard della media.



L'*errore standard* della media ➔ quantifica il grado di certezza con il quale la media, calcolata da un campione casuale, stima la vera media della popolazione.

La formula precisa per il calcolo è:

$$\sigma_{\bar{x}} = \frac{standard\ deviation\ della\ popolazione}{\sqrt{n}}$$

La stima migliore di questo valore preciso è:

$$\hat{\sigma}_{\bar{x}} = \frac{standard\ deviation\ del\ campione}{\sqrt{n}}$$