Elementi di statistica

Popolazione statistica e campione casuale

- Si chiama **popolazione statistica** l'insieme di tutti gli elementi che si vogliono studiare (individui, animali, vegetali, cellule, caratteristiche delle collettività ..) e può avere un numero finito o infinito di elementi.

Es: insieme degli abitanti di una città, di una nazione; insieme delle altezze di una popolazione di fascia di età fissata.

- Problema: non sempre si possono raccogliere dati su tutti gli elementi della popolazione, quindi se ne considera un sottoinsieme rappresentativo, il campione.
- Si chiama **campione casuale** una sequenza di elementi scelti a caso dalla popolazione in modo che ogni elemento abbia la stessa probabilità di far parte del campione.

Statistica descrittiva e inferenziale

- **Statistica descrittiva** : se l'indagine è sulla totalità della popolazione, sintesi quantitativa completa del fenomeno studiato (es. censimento)

- **Statistica inferenziale**: studia come e con quale precisione si possono descrivere le caratteristiche di una popolazione se l'indagine viene effettuata su un campione, vi è quindi incertezza.

Variabili statistiche

- Fissata una popolazione si chiamano **variabili statistiche** tutte le caratteristiche che variano al variare dei componenti delle popolazione.
- Le variabili che sono espresse qualitativamente sono dette **attributi** (colore degli occhi, della pelliccia, ecc..); quelle che sono espresse quantitativamente sono dette **misurabili** (temperatura a Cagliari alle 8:00 am)

Dati (informazioni empiriche) e rappresentazioni dei dati

Esempio. Rappresentiamo su un **diagramma di punti,** su **un istogramma (diagramma a blocchi), su un aerogramma** i dati raccolti su una tabella

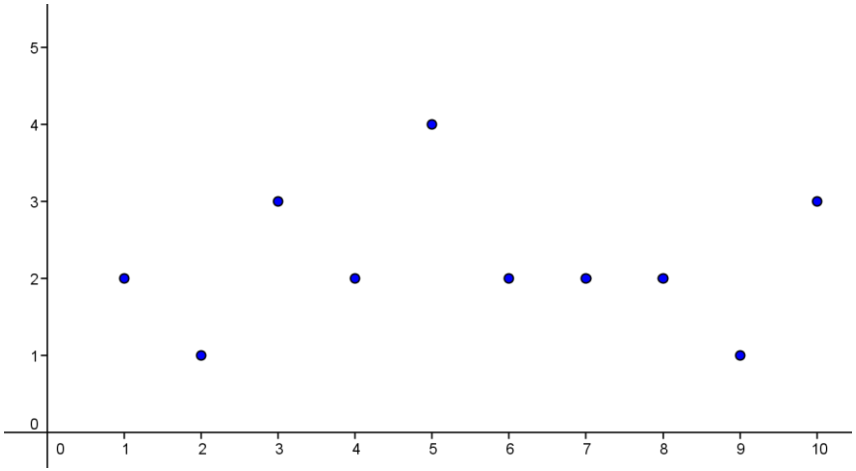
- Campione: 10 esemplari di gatto.
- Varabile misurabile: numero di cuccioli partoriti in un dato periodo.

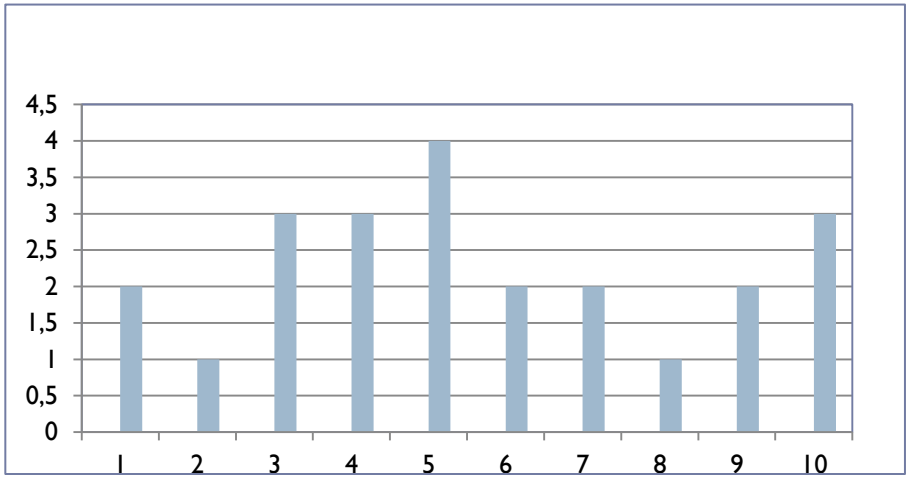
gatto	1	2	3	4	5	6	7	8	9	10
n.cuccioli	2	1	3	3	4	2	2	1	2	3

- **diagramma di punti:** in un sistema di riferimento cartesiano ogni punto è rappresentato dalla coppia ordinata (g_i, n.cuccioli)
- **istogramma (diagramma a canne):** i dati sono rettangoli di base costante (ciascuna rappresenta un individuo) altezza uguale al numero di cuccioli di ciascun individuo.

Dati (informazioni empiriche) e rappresentazioni dei dati

gatto	1	2	3	4	5	6	7	8	9	10
n.cuccioli	2	1	3	3	4	2	2	1	2	3





$(g_i, n.cuccioli)$

$(g_i, n.cuccioli)$

Possiamo organizzare e rappresentare i dati in un diverso istogramma? Quale il senso e il procedimento per la rappresentazione su un **aerogramma ?**

Frequenze assolute e relative

N: **dimensione del campione**, numero totale dei dati raccolti

$X_1, X_2, \ldots, X_N$: i dati, cioè i valori assunti nel campione dalla variabile statistica X

In molti casi i dati sono ripetuti, cioè assumono un numero finito di valori discreti distinti $x_1, x_2, \ldots, x_n$, $n \leq N$. Indichiamo con

F_i : il numero di dati uguali ad x_i , cioè la **frequenza assoluta**

$f_i = \frac{F_i}{N}$: la **frequenza relativa**, dove si prende in considerazione la numerosità del campione

Frequenza assoluta e relativa

F_i: **frequenza assoluta**

$f_i = F_i/N$: **frequenza relativa**

- Esempio: determinare la frequenza assoluta e relativa dei dati:

gatto	1	2	3	4	5	6	7	8	9	10
n.cuccioli	2	1	3	3	4	2	2	1	2	3

Variabile misurabile X: *numero di cuccioli partoriti in un dato periodo.*

Valori assunti dalla variabile X: $X_2 = X_8 = 1$; $X_1 = X_6 = X_7 = X_9 = 2$;
$X_3 = X_4 = X_{10} = 3$; $X_5 = 4$;

I valori distinti che assume X sono 4: $x_1 = 1,\ x_2 = 2,\ x_3 = 3,\ x_4 = 4$

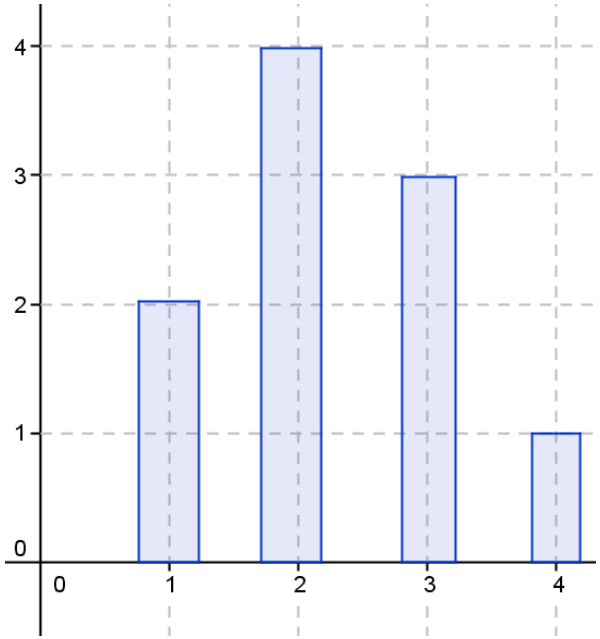
Le corrispondenti frequenze assolute sono : $F_1 = 2, F_2 = 4, F_3 = 3, F_4 = 1$

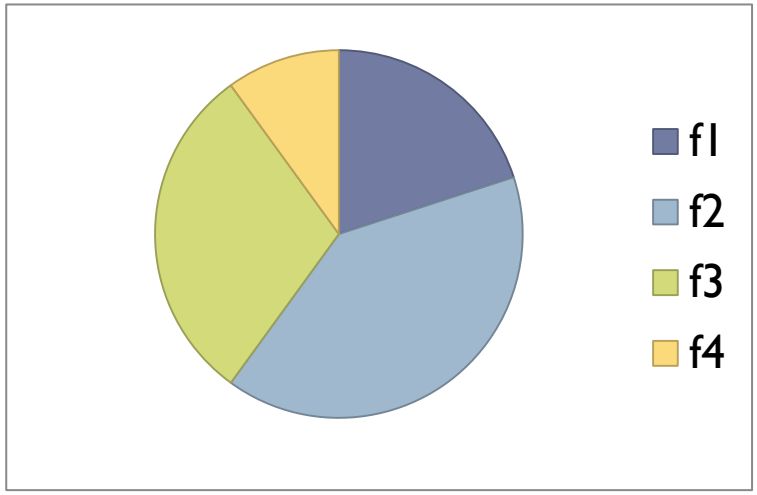
Le corrispondenti frequenze relative sono : $f_1 = \frac{2}{10}, f_2 = \frac{4}{10}, f_3 = \frac{3}{10}, f_4 = \frac{1}{10}$

- Esercizio: scrivere le frequenze relative in percentuale

Istogrammi e aerogrammi

rappresentazione dell'aerogramma e dell'istogramma delle frequenze





Il 20% delle gatte ha 1 cucciolo
Il 40% ha 2 cuccioli
Il 30% ha 3 cuccioli
Il 10% ha 4 cuccioli

I dati saranno generalizzabili a tutta la popolazione dei gatti?

Raccoglimento in classi

Misurando le altezze di 100 ragazze al primo anno di ingegneria sono stati ricavati i seguenti dati:

1 ragazza è alta 150 cm;
2 ragazze sono alte 153 cm;
5 ragazze sono alte 155 cm;
3 ragazze sono alte 156 cm;
2 ragazze sono alte 157 cm;
4 ragazze sono alte 158 cm;
3 ragazze sono alte 159 cm;
5 ragazze sono alte 160 cm;
12 ragazze sono alte 161 cm;
15 ragazze sono alte 162 cm;
8 ragazze sono alte 163 cm;
3 ragazze sono alte 164 cm;
10 ragazze sono alte 165 cm;
10 ragazze sono alte 166 cm;
7 ragazze sono alte 167 cm;
5 ragazze sono alte 168 cm;
5 ragazze sono alte 169 cm;

Come si possono rappresentare questi dati in modo efficiente?

Conviene raggruppare i dati in **classi**.

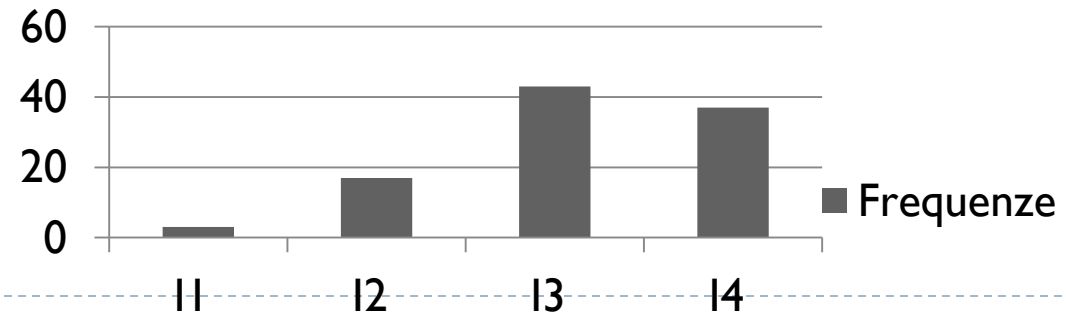
Indici di tendenza centrale: moda e classe modale

- Riassumere e organizzare i dati: indici numerici che riassumono le principali caratteristiche matematiche dei dati.
- Se i dati sono espressi mediante la loro appartenenza a diverse classi (sottoinsiemi), si chiama **classe modale** la classe di frequenza massima, se le classi sono individuate da numeri, il numero che contraddistingue la classe modale prende il nome di **moda.**

Esempio: determinare moda (o classe modale), frequenza relativa e disegnare l'istogramma delle frequenze assolute.

Altezze in centimetri di un campione di ragazze suddivise in classi
$I_1 = [150, 154]\ I_2 = [155, 159]\ I_3 = [160, 164] I_4 = [165, 169]$

frequenza $F_1 = 3, F_2 = 17, F_3 = 43, F_4 = 37$



Classe modale: I_3
che ha frequenza 43

Indici di tendenza centrale: media

Data una variabile statistica X con $X_1, X_2, \ldots, X_N$ i dati numerici relativi ad un campionamento, si chiama **media campionaria** di X la media aritmetica dei dati

$$m_X = \frac{1}{N}\sum_{k=1}^{N} X_k$$

Se i dati distinti sono $x_1, x_2, x_3, \ldots, x_n$ n N e assumono frequenze F_i si può scrivere in modo equivalente

$$m_X = \frac{1}{N}\sum_{i=1}^{n} F_i x_i = \sum_{i=1}^{n} \frac{F_i}{N} x_i = \sum_{i=1}^{n} f_i x_i$$

Si può talvolta preferire la media pesata o ponderata.
Se i dati hanno variazioni esponenziali, l'indice più adatto è **la media geometrica**

Indici di tendenza centrale: media geometrica

Data una variabile statistica X con $X_1, X_2, \dots, X_N$ i valori campionari numerici (positivi) relativi ad un campionamento.

Si chiama **media geometrica**

$$GM_X = \sqrt[N]{X_1 \cdot X_2 \cdot \dots \cdot X_N}$$

Se i dati distinti sono $x_1, x_2, \dots, x_n, n$ N e assumono frequenze assolute F_i si può scrivere in modo equivalente

$$GM_X = \sqrt[N]{x_1^{F_1} x_2^{F_2} \dots x_n^{F_n}}$$

Quando la media non è una buona stima riassuntiva, si può usare la mediana

Indici di tendenza centrale: sintesi

Sia X una variabile statistica, indichiamo con

$X_1, X_2, \dots, X_N$ i dati numerici relativi ad un campionamento

$x_1, x_2, x_3, \dots, x_n, n \leq N$, i dati distinti che vengono assunti con frequenze $F_1, \dots, F_n$

Indici di posizione:

- **Media**:

  aritmetica $m_X = \frac{1}{N}\sum_{k=1}^{N} X_k$ $\qquad m_X = \frac{1}{N}\sum_{i=1}^{n} F_i x_i = \sum_{i=1}^{n} \frac{F_i}{N} x_i = \sum_{i=1}^{n} f_i x_i$

  geometrica $GM_X = \sqrt[N]{X_1 \cdot X_2 \cdot \dots \cdot X_N}$ $\qquad GM_X = \sqrt[N]{x_1^{F_1} x_2^{F_2} \dots x_n^{F_n}}$

- **Moda** (o **classe modale**) : valore numerico (o classe) di frequenza massima
- **Mediana**

Indici di tendenza centrale: mediana

Siano $X_1, X_2, , \ldots, XN$ i valori campionari numerici **ordinati in modo crescente** (non decrescente). La **mediana** è il **valore centrale** (che separa in due parti uguali l'insieme dei dati) che si ottiene con la seguente regola:

Se N è dispari, la mediana è il **valore del dato** che corrisponde all'intero successivo a $\frac{N}{2}$

Se N è pari, è la **media aritmetica dei valori dei dati** al posto $N/2$ e al posto successivo.

Scrivere in formula il valore della mediana per N pari.

Calcolare la mediana nel caso delle tre successioni di dati assegnate

A	3,5	4,2	3,25	4,12
B	3,5	8,2	3,25	4,12
C	35	4,2	3,25	4,12

e verificare che la mediana, contrariamente alla media, risente poco della presenza di dati estremi (o di eventuali errori)

Frequenza relativa cumulata

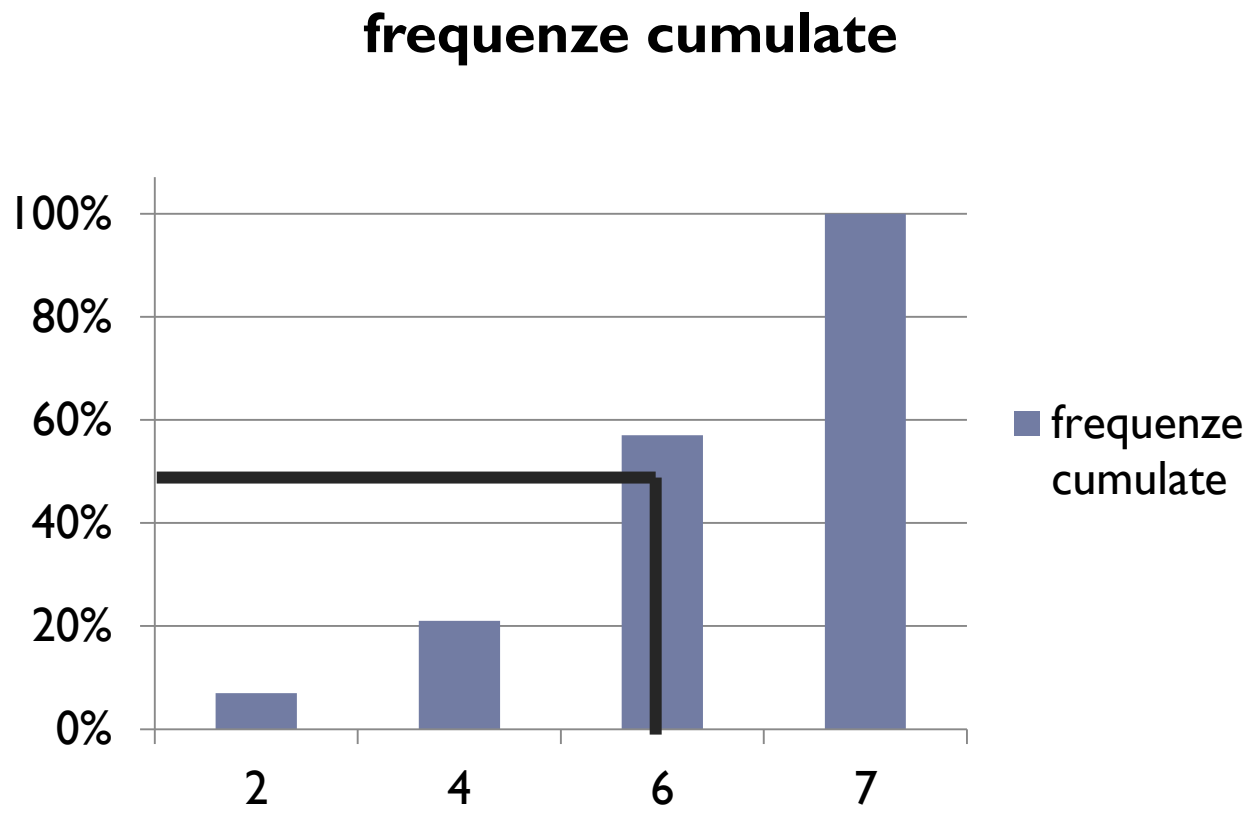
Siano $x_1, x_2, \dots, x_n$ i valori campionari numerici distinti e $F_1, F_2, \dots, F_n$ le rispettive frequenze. Determinare la **frequenza relativa cumulata** di x_i (somma delle frequenze relative dei dati da 1 a n_i) fornisce informazioni sul valore della mediana.

Esempio. A partire dai dati della tabella determinare la media, la mediana e l'istogramma delle frequenze relative e cumulate (fc) espresse in percentuale.

Dati	2	4	6	7
F	1	2	5	6
f	1/14	2/14	5/14	6/14
fc	1/14	3/14	8/14	14/14
f in %				
fc in %				

Media	5.85
Mediana	6

Esempio. A partire dai dati della tabella determinare la media, la mediana e l'istogramma delle frequenze relative e cumulate (fc) espresse in percentuale



La mediana si ha in corrispondenza del 50% e infatti risulta essere 6

Indici di tendenza centrale

Siano $X_1, X_2, , \ldots, XN$ i valori campionari numerici raccolti in N osservazioni e rappresentati nella seguente tabella

1	0	3	2	0	4	5	6	4	0	2
5	2	6	2	9	9	3	4	9	2	1

Esercizio.

Determinare la **media aritmetica**, la **media geometrica**, la **moda**, la **mediana** e rappresentare l'**istogramma** delle **frequenze relative** espresse in percentuale e quello delle **frequenze cumulate**.

Rappresentare media aritmetica e mediana nei due istogrammi.

Varianza e deviazione standard

Sia X una variabile statistica, la **varianza campionaria** di N dati $X_1, X_2, , \ldots, X_N$ aventi media campionaria m_X è il numero

$$s_X^2 = \frac{(X_1 - m_X)^2 + (X_2 - m_X)^2 + \ldots.. + (X_N - m_X)^2}{N-1}$$

che valuta la distanza media al quadrato dei dati dalla media, cioè la loro ***dispersione***

Se i dati assumono un numero finito di valori discreti distinti $x_1, x_2, x_3, \ldots, x_n$ $n \leq N$ e con frequenza assoluta F_i si definisce **varianza campionaria**

La radice quadrata della varianza $s_X^2 = \dfrac{\sum_{i=1}^{n} F_i (x_i - m_X)^2}{N-1}$

$s_X = \sqrt{s_X^2}$ è la **deviazione standard campionaria**

Varianza e deviazione standard

Si eseguono alcune misure di una grandezza X e si rilevano i seguenti risultati con le frequenze indicate sotto

X	0	1.3	1.2	0.3	3.4	0.5	1.6	4.7	0.8	2.9
F	2	6	12	9	19	39	42	39	21	11

Calcolare la media, la varianza campionaria e la deviazione standard

Varianza e deviazione standard

Se i dati a disposizione riguardano un'intera popolazione (non un campione) si usano simboli differenti nelle definizioni degli indici di tendenza centrale.

Indichiamo con $y_1, y_2, \dots, y_M$ i dati relativi a tutti gli individui di una popolazione

la **media** di popolazione è il numero $\mu = \frac{1}{M}\sum_{i=1}^{M} y_i$

la **varianza** di popolazione è il numero $\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}(y_i - \mu)^2$

la **deviazione standard** di popolazione, chiamata anche **scarto quadratico medio** è $\sigma = \sqrt{\sigma^2}$

Per N abbastanza grande la diversità tra la **varianza campionaria** (varianza stimata) e la **varianza** di popolazione (varianza) diventa trascurabile. Analogo risultato si ha per la deviazione standard.

Varianza e deviazione standard

Esempio: Calcoliamo le deviazioni standard dei seguenti dati considerati come dati di un'intera popolazione.

Sia X l'insieme delle altezze degli atleti di una squadra di calcetto

$X = \{176, 181, 168, 176, 172\}$.

Calcolare la media, la varianza e la deviazione standard

Media $\mu = \frac{1}{M}\sum_{i=1}^{M} y_i \qquad \mu = \frac{1}{5}(176+181+168+176+172) = 174.6$

Varianza $\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}(y_i - \mu)^2 \qquad \sigma^2 = \frac{1}{5}\left(1.4^2 + 6.4^2 + (-6.6)^2 + 1.4^2 + (-2.6)^2\right)$

$$\sigma^2 = \frac{1}{5}(1.96 + 40.96 + 43.56 + 1.96 + 6.76) = \frac{95.6}{5} = 19.1$$

Deviazione standard $\sigma = \sqrt{\sigma^2} = \sqrt{19.1} = 4.37$

Varianza e deviazione standard

La **varianza** si può anche calcolare con la formula (di König)

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}(y_i - \mu)^2 \qquad \sigma^2 = \left(\frac{1}{M}\sum_{i=1}^{M} y_i^2\right) - \left(\frac{1}{M}\sum_{i=1}^{M} y_i\right)^2$$

Verificare l'uguaglianza a partire dai dati $X = \{176,181,168,176,172\}$.

N dati considerati come dati di un'intera popolazione $X = \quad X_1, X_2, , \ldots, XN \}$, o di N osservazioni empiriche possono essere considerati come **vettori**.
Gli indici di tendenza centrale, o di dispersione definiti utilizzando gli strumenti dell'algebra dei vettori.
Gli stessi indici possono anche essere definiti in termini probabilistici.
La trattazione in questi diversi ambiti matematici non è oggetto di questa trattazione.