7. STATISTICA DESCRITTIVA

Quando si effettua un'indagine statistica si ha a che fare con un numeroso insieme di oggetti, detto *popolazione* del quale si intende esaminare una o più caratteristiche (matricole di biotecnologia 2015/2016, molecole di un gas, insiemi di batteri,…).
Quando la popolazione è troppo numerosa per essere studiata se ne estrae un *campione casuale* C di dimensione $n \in \mathbb{N}$, ovvero un campione è un sottoinsieme di n individui scelti a caso nella popolazione.
Una volta raccolti i dati di interesse, essi si presentano in forma disordinata e per questo motivo vengono chiamati *dati grezzi.*
La *statistica descrittiva* si occupa di riordinare i dati grezzi in tabelle che siano leggibili e rappresentabili graficamente. Inoltre si occupa di trarre informazioni dei dati così raggruppati (media, moda, mediana, varianza, scarto quadratico medio.)
Possiamo considerare i dati singolarmente oppure possiamo raggrupparli in classi.

Esempi

1) Dati singoli: età di un gruppo di professori: $\{48,49,49,51,54,55,58,58,60,60,60,61,62,62\}$
2) Classi di dati (e= età):
   $47 < e \leq 51$ I classe,
   $51 < e \leq 55$ II classe,
   $55 < e \leq 59$ III classe,
   $59 < e \leq 63$ IV classe.
   Definiamo *ampiezza* di una classe $a < e \leq b$ il numero $b - a$ nel nostro esempio 4.
   Il *valore* con cui si identifica la classe invece è il *valore centrale,* ovvero $\frac{a+b}{2}$

TABELLA DI DISTRIBUZIONE DELLE FREQUENZE
Si tratta di una tabella che riordina e riassume i dati raccolti.

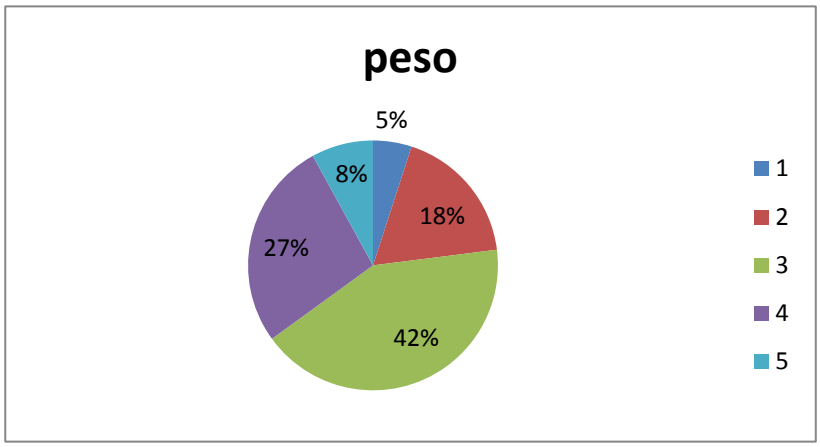
Definizione Chiamiamo

- *Frequenza assoluta* il numero di osservazioni che ricadono su quel dato o classe,
- *Frequenza relativa* il numero compreso tra 0 e 1 che ne deriva dividendo la frequenza assoluta con il numero di osservazioni totali,
- *Frequenza percentuale* è data dalla frequenza relativa moltiplicata per cento e messa quindi in percentuale.

Esempio Si sono rilevati i pesi di 200 studenti maschi di unife ottenendo questi dati :

Peso (x)	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
1. $60 < x \leq 63$	10	0.05	5%
2. $63 < x \leq 66$	36	0.18	18%
3. $66 < x \leq 69$	86	0.42	42%
4. $69 < x \leq 72$	54	0.27	27%
5. $72 < x \leq 75$	16	0.08	8%
TOT	200	1	100%

Solitamente la tabella di frequenze viene rappresentata graficamente mediante due tipologie di grafico, il diagramma a torta e l'istogramma.
Diagramma a torta:



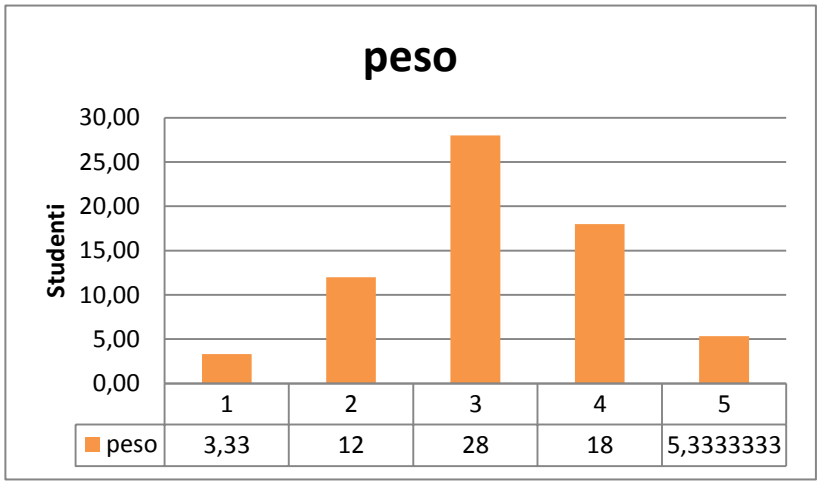
Ogni classe viene rappresentata con un settore circolare, di ampiezza proporzionale alla frequenza della classe. Si usa soprattutto per rappresentare frequenze percentuali. Per costruire le fette si usa la proporzione

$$f_{\%}: 100 = \alpha: 360°$$

Dove con α indichiamo l'angolo del settore.

Questo diagramma è molto usato quando i dati non sono numerici .

Istogramma



Consiste di rettangoli adiacenti aventi per base l'ampiezza della classe ed altezza la frequanza assoluta diviso l'ampiezza in modo tale che l'area del rettangolo mi dia la frequanza assoluta della classe corrispondente.
Si osservi che se le classi sono di ampiezza costante k allora le altezze derivano dalla formula:

$$h = \frac{f_A}{k}$$

Se le classi sono di ampiezza 1 (ovvero abbiamo dati singolo), allora l'altezza corrisponde alla frequenza assoluta.

GRANDEZZE CHE SINTETIZZANO I DATI.

- Indici di posizione centrale: *media, moda, mediana,* ci dico attorno a quale valore si dispongono i dati.
- Indici di dispersione: *varianza, scarto quadratico medio,* ci dicono quanto i dati sono dispersi rispetto al valore centrale.

Definizione La *media aritmetica* dei valori $x_1, \dots, x_n$ è

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

Se invece i dati sono raggruppati in classi, detti $x_1, \dots, x_k$ i valori centrali delle k classi, e dette $f_1, \dots, f_k$ le frequenze assolute delle k classi si ha che:

$$\overline{x} = \frac{x_1 f_1 + \cdots + x_k f_k}{f_1 + \cdots + f_k}$$

E viene chiamata *media campionaria.*

Riprendendo l'esempio precedente si ha che

$$\overline{x} = \frac{61.5 \cdot 10 + 64.5 \cdot 36 + 67.5 \cdot 84 + 70.5 \cdot 54 + 73.5 \cdot 16}{200} = 67.95$$

Definizione Presi i dati e ordinati in ordine crescente, definiamo *mediana* $\tilde{x}$ il dato centrale della lista.

Esempi

1) La mediana dei numeri $3,4,5,5,7$ è $\tilde{x} = 5$,
2) La mediana dei numeri $3,4,5,6,6,7$ è $\tilde{x} = \frac{5+6}{2} = 5.5$

N.B. Se ho un numero pari di dati la mediana è la media aritmetica dei due dati centrali.

3) peso dei 200 studenti unife:

$$\underbrace{61.5, \dots, 61.5}_{16 \text{ volte}}, \underbrace{64.5, \dots, 64.5}_{36 \text{ volte}}, \underbrace{67.5, \dots, 67.5}_{84 \text{ volte}} \dots$$

I due dati centrali si trovano al centesimo e al centunesimo posto e sono entrambi uguali a 67.5 quindi la mediana $\tilde{x} = \frac{67.5+67.5}{2} = 67.5$

Definizione La *moda* è il dato che compare con la frequenza maggiore, non sempre esiste e non sempre è unica.

Esempi Consideriamo le seguenti serie di dati a calcoliamo la moda

1) $2,4,4,7,8$ moda $= 4$ significa che i dati sono unimodali
2) $2,4,5,6,7$ moda $\nexists$ non esiste
3) $2,2,4,4,6,7$ moda $= 2,4$ significa che i dati sono bimodali
4) Presi come dati quelli che ci indicano il peso di 200 studenti unife, abbiamo che la moda è 67.5

Definizione *La varianza* ci dice quanto sono dispersi i dati rispetto al valore centrale.
Se i dati $x_1, \dots, x_n$ sono singoli,

$$s^2 = \frac{(x_1 - \overline{x})^2 + \dots + (x_n - \overline{x})^2}{n - 1}$$

Se i dati sono raggruppati in k classi di valori centrali $x_1, \dots, x_k$

$$s^2 = \frac{f_1(x_1 - \overline{x})^2 + \dots + f_k(x_k - \overline{x})^2}{f_1 + \dots + f_k - 1}$$

Osservazione Si divide per $n - 1$ e non per n perché si è visto sperimentalmente che così si ottengono stime più precise.

Definizione *Lo scarto quadratico medio o deviazione standard:* $s = \sqrt{s^2}$ ovvero è la radice quadrata della varianza.

Esempi Calcolare la media e la varianza dei seguenti gruppi di dati

1) $A = \{10,10,10,10,10\}$ $\overline{x} = 10$ $s^2 = 0$ (per valori costanti la varianza è nulla)
2) $A = \{2,5,10,15,18\}$ $\overline{x} = 10$ $s^2 = 44.5$ (varianza grande)
3) $A = \{8,9,10,11,12\}$ $\overline{x} = 10$ $s^2 = 2.5$ (varianza piccola)
4) 200 studenti unife: $s^2 = 8.6$

Esercizi

1) Si sono registrati i battiti cardiaci al minuto nell'arco di 10 giorni ad una persona. Si sono ottenuti i seguenti dati:

$$\{73,72,73,74,76,76,70,71,72,72,74\}$$

   a) Sistemare i dati nella tabella di distribuzione di frequanza e disegnare l'istogramma delle osservazioni.
   b) Determinare media, moda mediana, varianza e scarto quadratico medio.
   c) Determinare la percentuale dei giorni in cui vengono registrati alla persona un numero di battiti cardiaci al minuto maggiori o uguali a 73.

2) Si sono rilevate le altezze in centimetri di 200 studenti maschi dell'Università di Ferrara ottenedno i seguenti risultati:

$x =$ Altezza in cm	Numero di studenti
$160 < x \leq 165$	8
$165 < x \leq 170$	24
$170 < x \leq 175$	46
$175 < x \leq 180$	82
$180 < x \leq 185$	36
$185 < x \leq 190$	4

   a) Sistemare i dati nella tabella di distribuzione delle frequenze, specificando il valore centrale con cui si identifica ogni classe e disegnare l'istogramma delle osservazioni.
   b) Determinare media, moda mediana, varianza e scarto quadratico medio dell'altezza degli studenti.