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

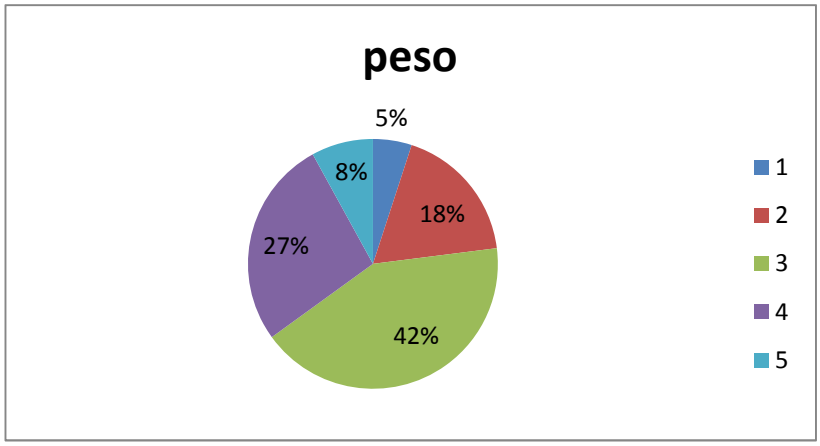
- *Frequenza assoluta* il numero di osservazioni che ricadono su quel dato o classe,
- *Frequenza relativa* il numero compreso tra 0 e 1 che ne deriva dividendo la frequenza assoluta con il numero di osservazioni totali,
- *Frequenza percentuale* è data dalla frequenza relativa moltiplicata per cento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

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

Peso (	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
1. $60 <$ 3	10	0.05	5%
2. $63 <$ 6	36	0.18	18%
3. $66 <$ 9	86	0.42	42%
4. $69 <$ 2	54	0.27	27%
5. $72 <$ 5	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:



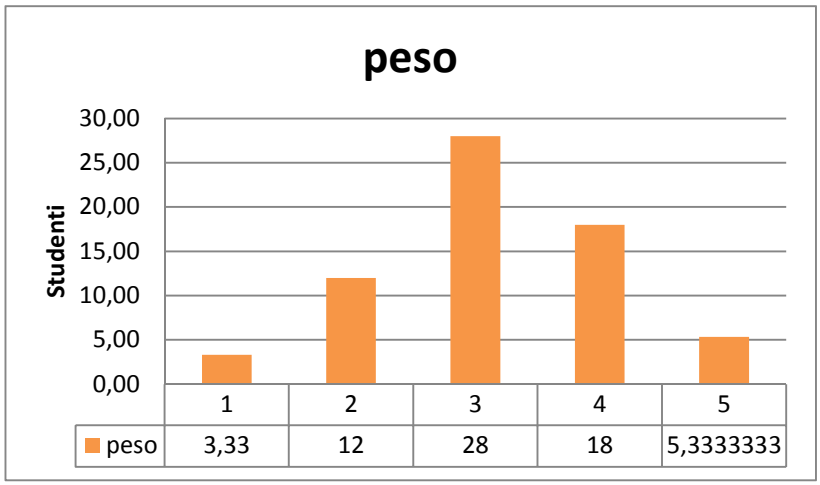
Ogni classe viene rappresentata con un settore circolare, di ampiezza proporzionale alla frequenza della classe. Si usa soprattutto per rappresentare frequenze percentuali. Per costruire le fette si usa la proporzione

$$f_{\%} : 100 = \alpha : 360°$$

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 frequenza assoluta diviso l'ampiezza in modo tale che l'area del rettangolo mi dia la frequenza 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:

$$\underbrace{61.5,\dots,61.5}_{10\text{ volte}},\underbrace{64.5,\dots,64.5}_{36\text{ volte}},\underbrace{67.5,\dots,67.5}_{84\text{ volte}}\dots$$

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,

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

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

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

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 frequenza 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 ottenendo i seguenti risultati:

ltezza in cm	Numero di studenti
$160 <$ 65	8
$165 <$ 70	24
$170 <$ 75	46
$175 <$ 80	82
$180 <$ 85	36
$185 <$ 90	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.

8. STATISTICA INDUTTIVA

8.1 Regressione lineare

In molto esperimenti al ricercatore interessa l'analisi delle variazioni di due o più variabili per evidenziare le eventuali relazioni esistenti tra di loro e predire valori interessanti delle variabili non note sperimentalmente.
Noi ci occupiamo solo di esperimenti con due variabili.
La relazione che viene presa in esame è la dipendenza di una variabile rispetto all'altra

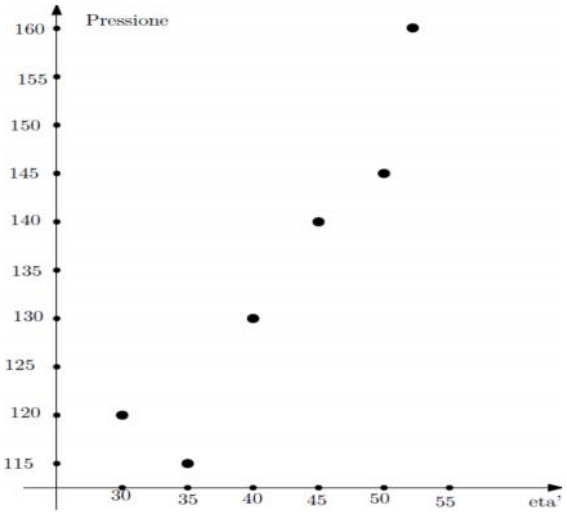
Il rapporto di dipendenza in matematica si indica con *funzione,* in statistica lo indicheremo con la parola *regressione.*
Si indica come indipendente (X), una variabile per cui i livelli possono essere fissati o possono essere fissati sperimentalmente (es. le dosi di una sostanza), oppure possono essere semplicemente rilevati (es. la temperatura).
Si indica come variabile dipendente (Y), una variabile la cui variazione si assume essere la risposta alle variazioni della variabile indipendente.

Esempio Nella seguente tabella sono riportate le età e le pressioni arteriose massime (in $mmHg$ millimetri di mercurio) di 6 soggetti maschili.

ETÀ	30	35	40	45	50	52
PRESSIONE	120	115	130	140	145	160

OBIETTIVO: cercare se esiste una dipendenza del valore della pressione massima dall'età maschile.

Per capire questa dipendenza la prima cosa da fare è rappresentare i dati nel *diagramma di dispersione*: un grafico cartesiano in cui mettiamo in ascissa la variabile indipendente (in questo caso l'età) e in ordinata la variabile dipendente (in questo caso la pressione).
I dati sperimentali sono quindi rappresentati da punti nel grafico di dispersione.
Noi saremo interessati a quelle variabili che dipendono linearmente l'una dall'altra. (graficamente parlando saremo interessati a quei grafici a dispersione in cui i punti sembrano allinearsi su una retta).
Nel nostro esempio:



I dati sembrano effettivamente disporsi su una retta. Esistono diversi tipi di regressione:
uno lineare è quello di regressione basato sulla retta dei minimi quadrati (dal grafico sembra che questo possa essere un buon modello per rappresentare la dipendenza dei dati)

Vedere il grafico non basta però a capire se effettivamente tra i dati esiste una relazione lineare (l'effetto visivo potrebbe confonderci), dobbiamo osservare matematicamente se esiste questa dipendenza.
Per far questo introduciamo il *coefficiente di correlazione* r, che ci permette di capire se esiste o meno una buona relazione lineare tra le variabili.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad \text{con } \overline{x}, \overline{y} \text{ medie dei dati } x_i, y_i$$

Il numeratore può essere espresso più semplicemente :

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

$$= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} - n\overline{x}\,\overline{y} + n\overline{x}\,\overline{y} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$$

Perché:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \quad e \quad \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$$

Dunque:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Il coefficiente di regressione:

- È sempre $-1 \le r \le 1$,
- Se r è vicino a 1 o -1 c'è una buona correlazione lineare,
- Se $r = \pm 1$ la correlazione è perfetta
- Se $r < 0$ la corrispondenza è inversa, all'aumentare della variabile x diminuisce la variabile y
- Se $r = 0$ o si avvicina a 0, significa che i dati non sono in relazione lineare, in questo caso non ha senso approssimare i dati con la retta dei minimi quadrati.

Per i calcoli si può utilizzare la seguente tabella, nel nostro esempio:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 42 \qquad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 135$$

30	-12	144	120	-15	225	3600
35	-7	49	115	-20	400	4025
40	2	4	135	-5	25	5200
45	3	9	140	5	25	6300
50	8	64	145	10	100	7250
52	10	100	160	25	625	8320
		TOT: 370			TOT: 1400	TOT: 34695

Quindi abbiamo che $r = \frac{(34695 - 6 \cdot 42 \cdot 135)}{\sqrt{370 \cdot 1400}} = \frac{675}{719.722} \cong 0.938$

Notiamo che r si avvicina ad 1 ne segue che esiste una buona correlazione lineare tra le variabili.

I dati possono essere rappresentati dalla retta dei minimi quadrati: $y = mx + q$ dove:

$$m = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \quad \text{e} \quad q = \overline{y} - m\overline{x}$$

Nel nostro caso: $m = 1.824$ e $q = 58.38$

$$y = 1.82x + 58.38$$

La retta ci permette di fare delle previsioni, cioè

- Dire, presumibilmente, qual è la pressione massima di un maschio di 47 anni:

$$y = 1.82 \cdot 47 + 58.38 = 143.92$$

- Dire quale presunta età può avere un uomo con pressione massima pari a $150\ mmHg$

$$x = \frac{y - 58.38}{1.82} = 50.34$$

Esercizio