STATISTICA

Una piccola infarinatura in vista delle prove INVALSI.

INDICE

INTRODUZIONE

Definizione

La **statistica**, in senso moderno, è propriamente l'applicazione dei metodi scientifici alla programmazione della raccolta dei dati, alla loro classificazione, elaborazone, analisi e presentazione e alla inferenza di conclusioni attendibili da essi.

B. Giardina

STATISTICA $\leftrightarrow$ **STATO**

La statistica nasce per esigenze dello Stato (pensa ai censimenti). Successivamente si sviluppa come scienza vera e propria.

TIPOLOGIE

Solitamente la statistica di suddivide in:

- **DESCRITTIVA**: parte della statistica che ha lo scopo di raccogliere ed elaborare i dati per descrivere fenomeni collettivi. É un'indagine eseguita su tutta la popolazione.
- **INFERENZIALE**: ramo che si occupa di eseguire indagini non su tutta la popolazione, ma su un sottoinsieme di essa, su un **campione**. Successivamente cerca di stimare le caratteristiche del fenomeno studiato per tutta la popolazione, a partire dai dati raccolti per il campione.

PROBLEMA: il campione non è scelto a caso, ma deve rispecchiare tutte le caratteristiche della popolazione, in modo da non ottenere risultati fuorvianti (esempio dei cacciatori a Badia). Allora la statistica inferenziale si occupa anche di studiare come formare il campione.

TERMINOLOGIA

Parole nuove che abbiamo imparato:

- **UNITÁ STATISTICA**: è l'oggetto materiale su cui faccio la mia indagine; è l'oggetto su cui vado a studiare il carattere fissato;
- **POPOLAZIONE**: è l'insieme di tutte le unità statistiche;
- **CAMPIONE**: sottoinsieme della popolazione oggetto di studio per la statistica inferenziale;
- **CARATTERE**: caratteristica della popolazione che vado a studiare. Esso può essere:
  - **qualitativo**, cioè esprimibile tramite un attibuto;
  - **quantitativo**, cioè esprimibile tramite un numero;
- **MODALITÁ**: uno dei possibili modi in cui il carattere si manifesta;
- **FREQUENZA**: mi dice quante volte la modalità è ripetuta, il numero di unità statistiche aventi una data modalità.

ESEMPIO

Facciamo un'indagine statistica sul colore dei capelli degli studenti dell'IIS L. Einaudi.

UNITÁ STATISTICA: un singolo studente.
POPOLAZONE: gli studenti dell'istituto, tutti.
CARATTERE: colore dei capelli.
$\Rightarrow$ carattere di tipo **qualitativo**.
MODALITÁ: biondi, castani, mori, rossi, bianchi...

MODALITÁ RACCOLTA DATI

Per raccogliere i dati sono possibili diverse modalità. Alcune possono essere le seguenti:

- Intervista;
- Questionario;
- Sperimentazione;

VISUALIZZAZIONE DEI DATI RACCOLTI

Una volta raccolti,abbiamo bisogno di ordinare i dati, in modo da poter dare una prima letta e provare a dedurne qualcosa.
A questo scopo, essi vengono visualizzati in opportune tabelle, dette **distribuzioni statistiche**.
Allora una distribuzione statistica non è altro che una tabella in cui possiamo vedere tutte le modalità del carattere con le relative frequenze.

Esempio:

Titoli di studio	N. Persone
Senza titolo – Licenza elementare	2.869
Licenza scuola media inferiore	8.648
Licenza scuola media superiore	9.713
Laurea – Corso post-laurea	2.942
Totali	23.992

EXCURSUS SULLA FREQUENZA I

La frequenza, in un'indagine statistica, è una cosa fondamentale. Senza di essa non si potrebbe fare alcuno studio.

Definizione

La **frequenza** di una modalità è un numero intero ($\in \mathbb{N}$) che indica il numero delle unità statistiche sulle quali è stata osservata la fissata modalità.

ATTENZIONE: la definizione data è quelli di frequenza **assoluta**! Esistono anche altri tipi di frequenza. Abbiamo anche la:

- frequenza relativa: cioè il rapporto tra le rispettive frequenze assolute ed la somma totale delle frequenze;
- frequenza percentuale: che è la frequenza relativa, moltiplicata per cento, a cui faccio seguire il %.

EXCURSUS SULLA FREQUENZA II

Esempio: Consideriamo la precedente tabella, quella sui titoli di studio della popolazione e completiamola. I dati presenti sono esattamente le frequenze assolute.

Calcolo delle frequenze relative:

$$f_r(i) = \frac{f_{ass}(i)}{Totale}$$

Tot.: 23.992

$f_r(1) = \frac{2.869}{23.992} = 0,1195815271757252 \cong 0,1196$
$f_r(2) = \frac{8.648}{23.992} = 0,3604534844948316 \cong 0,3605$
$f_r(3) = \frac{9.713}{23.992} = 0,4048432810936979 \cong 0,4048$
$f_r(4) = \frac{2.942}{23.992} = 0,1226242080693565 \cong 0,1226$

Calcolo delle frequenze percentuali:

$$f_{\%}(i) = f_r(i) * 100 \quad \%$$

$f_{\%}(1) = 11,96\%$ $f_{\%}(2) = 36.05\%$ $f_{\%}(3) = 40,48\%$ $f_{\%}(4) = 12,26\%$

RAPPRESENTAZIONE GRAFICA DEI DATI

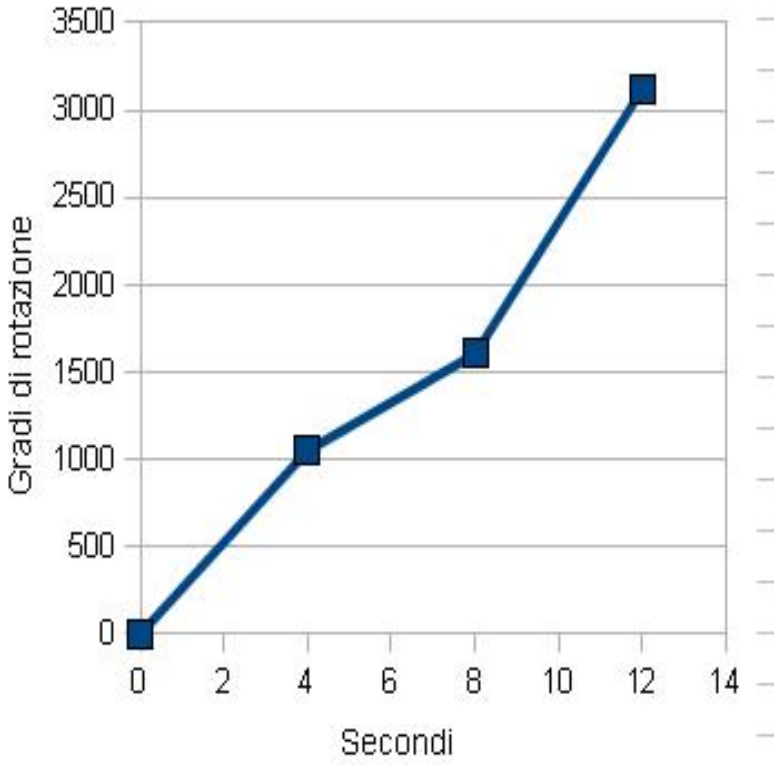
Per visualizzare graficamente i dati raccolti, abbiamo una vasta scelta di grafici, da scegliere in base ai nostri scopi. La rappresentazione grafica è importante per poterne dedurre delle prime informazioni grezze, che successivamente si andrà ad approfondire ed integrare.

Alcuni grafici a nostra disposizione sono:

- DIAGRAMMA CARTESIANO;
- ISTOGRAMMA;
- ORTOGRAMMA;
- DIAGRAMMA A RADAR;
- AEROGRAMMA;
- IDEOGRAMMA;
- CARTOGRAMMA.

ATTENZIONE: tali diagrammi per essere letti necessitano di una **legenda**!!

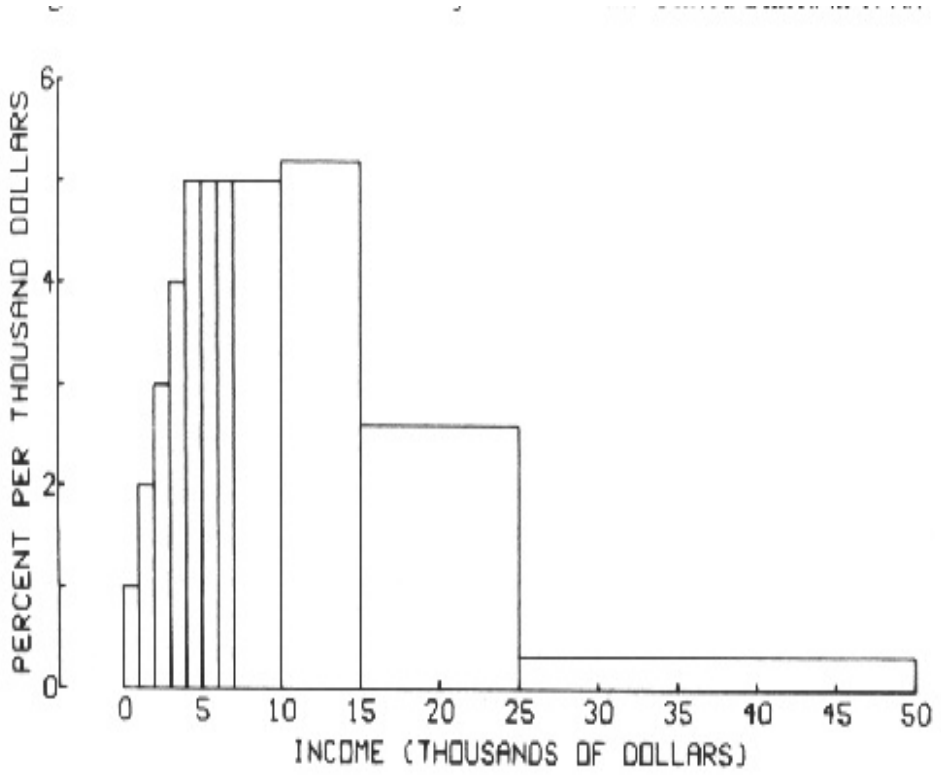
DIAGRAMMA CARTESIANO



Tale rappresentazione necessita del piano cartesiano. Sulle ascisse sono rappresentate le modalità del carattere, mentre sulle ordinate ritroviamo le frequenze assolute.
Le unità di misura dei due assi possono essere diverse.
Si individuano i punti che corrispondono alle righe della distribuzione di frequenza.
Generalmente i punti individuati poi sono uniti tramite una spezzata.

ISTOGRAMMA

Anche questa rappresentazione ha bisogno del piano cartesiano. Sulle ascisse sono rappresentate le modalità, mentre sulle ordinate troviamo dei valori numerici, corrispondenti all'**area** del rettangolo. Infatti in questo grafico si ha che l'**area** dei rettangoli è proporzionale alla frequenza assoluta delle varie modalità.

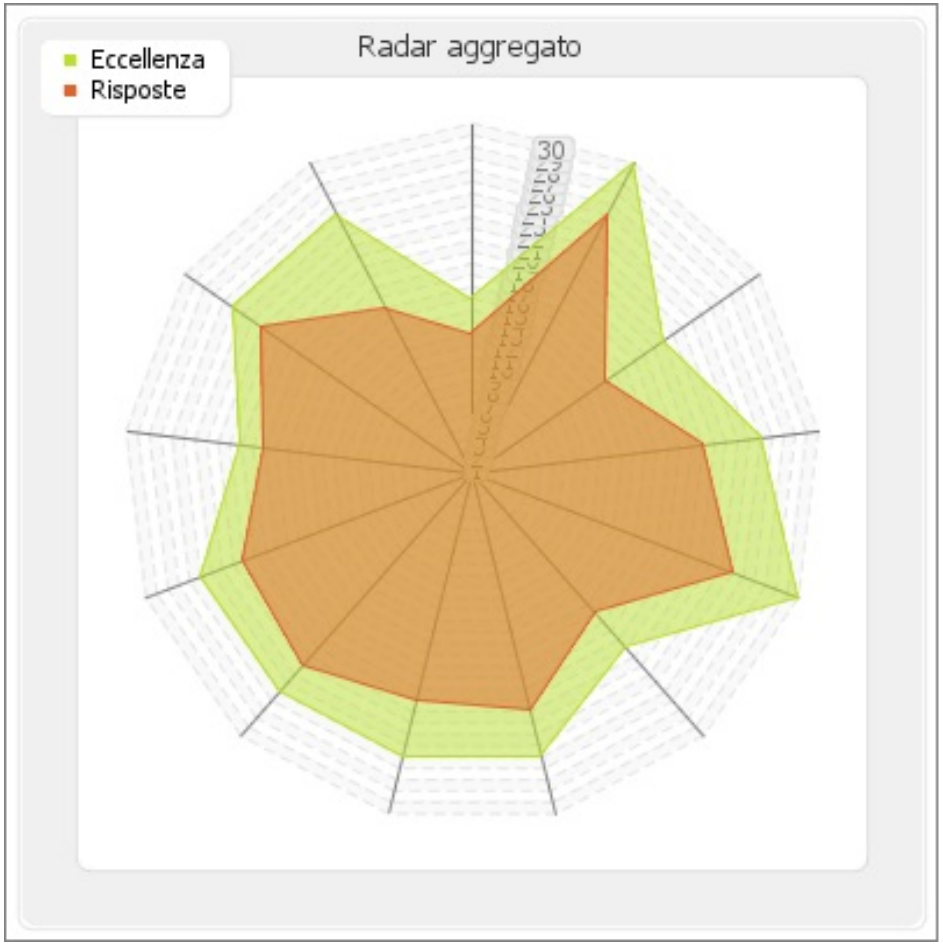


ORTOGRAMMA

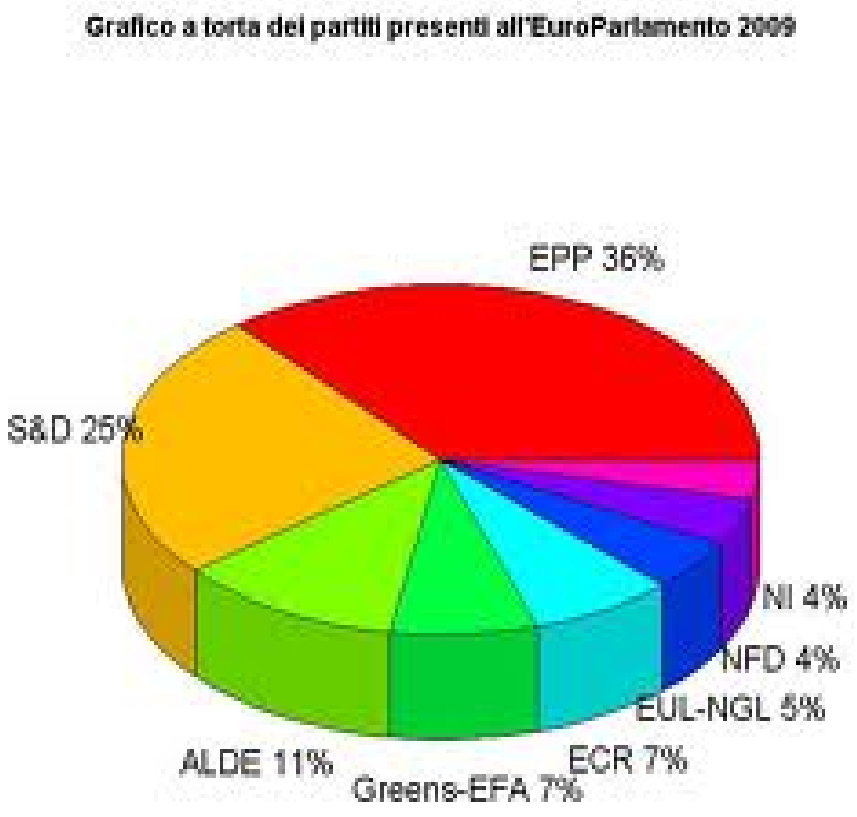
Questo diagramma
è un caso particolare
della tipologia precedente.
Infatti se consideriamo le
basi dei rettangoli tutte uguali,
abbiamo che l'**altezza** è,
in questo caso, proporzionale
alla frequenza assoluta.
Se la base è unitaria,
l'**altezza** è esattamente
uguale alla frequenza assoluta.

DIAGRAMMA A RADAR

Questa tipologia
di diagrammi è utilizzata
soprattutto per rappresentare
particolari indagini statistiche.
Sono ben rappresentabili
fenomeni **ciclici**,
cioò fenomeni che si ripetono
dopo un certo periodo.
Ogni semiretta rappresenta
una modalita, in modo
che gli angoli tra semirette
adiacenti siano uguali.
Sulle semirette si riportano
i dati, che poi vengono
collegati da una spezzata.



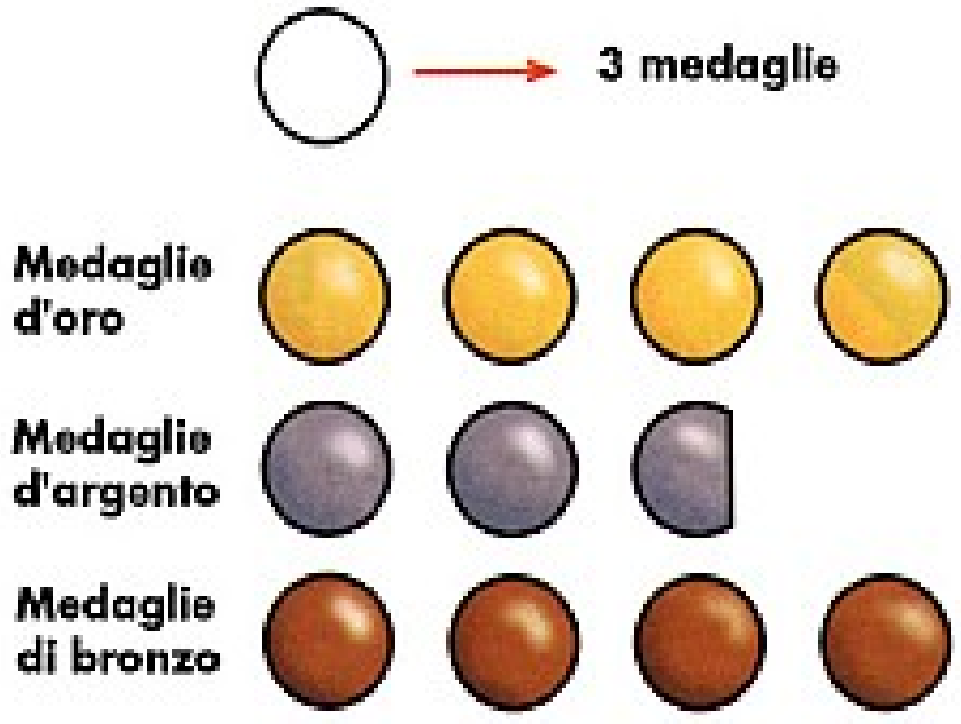
AEROGRAMMA



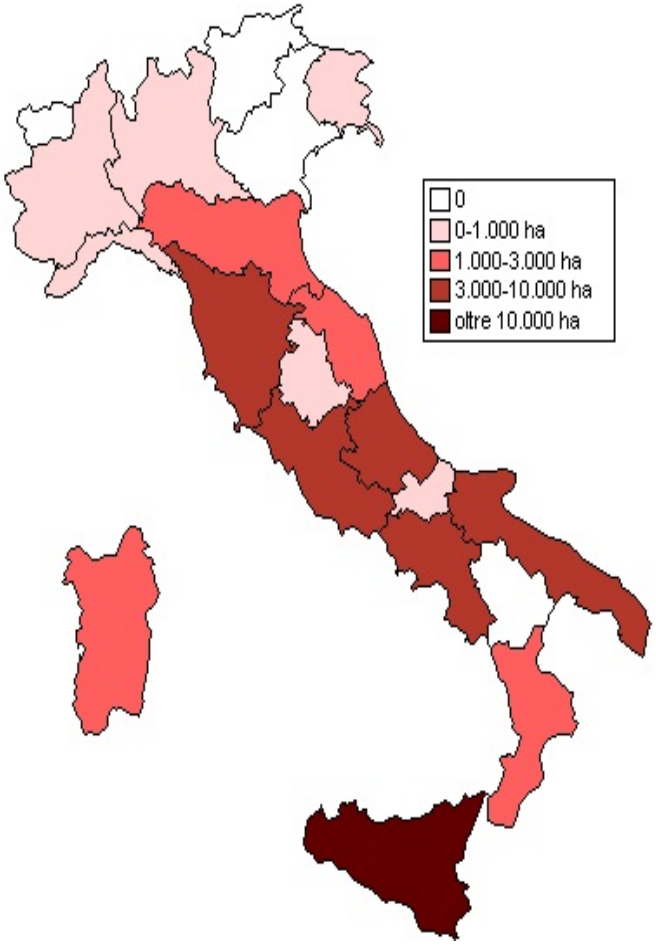
Più comunemente conosciuto come DIAGRAMMA A TORTA, ogni settore circolare rappresenta una modalità. L'angolo al centro è proporzionale alla frequenza assoluta della relativa modalità. Spesso utilizzato da riviste, libri e giornali, è di immediata lettura, soprattutto per indagini destinate ad un pubblico non specializzato.

IDEOGRAMMA

Tipologia di grafico intuitiva. Rappresentazione tramite figure di grandezza diversa, a seconda dell'intensità del fenomeno, oppure mediante tante figure quant'è la frequenza assoluta. Solitamente in quest'ultimo caso, una figura non è un'unica unità, ma rappresenta un certo numero di unità. Utilizzato soprattutto in riviste e libri, in quanto cattura l'attenzione del lettore.



CARTOGRAMMA



Rappresentazione che si serve di una cartina geografica per rappresentare il fenomeno. Vengono utilizzate colorazioni che rappresentano l'intensità del fenomeno. Solitamente un colore più intenso simboleggia una maggiore intensità del fenomeno. Utilizzato generalmente per rappresentare fenomeni legati alla posizione geografica (colture, piovosità, occupazione...).

ESERCIZIO

Consideriamo la solita distribuzione statistica dell'esercizio precedente, quella sul grado di istruzione della popolazione.
Vogliamo costruirne l'aerogramma.
Per prima cosa dobbiamo cercare l'ampiezza dell'angolo al centro corrispondente ad ogni singola modalità. La proporzione risolvente è la seguente:

$$\mathrm{f_{ass}(i)} : \mathrm{tot\,f_{ass}} = x : 360$$

Calcoliamo i gradi relativi ad ogni modalità dell'esempio:
$2.689 : 23.992 = x : 360 \quad x = 40,35$
$8.648 : 23.992 = x : 360 \quad x = 129,76$
$9.713 : 23.992 = x : 360 \quad x = 145,74$
$2.942 : 23.992 = x : 360 \quad x = 44,15$

MEDIE I

La rappresentazione mediante grafico è un primo modo per cercare di ricavare informazioni grezze dai dati.
Un primo strumento che è possibile utilizzare per raffinare le informazioni è la **MEDIA**.

> Si dice **media** un qualsiasi valore compreso fra il minimo ed il massimo
>
> *A. L. Cauchy*

Di medie ce ne sono di molti tipi, e di tali tipi ne esistono anche delle varianti. La scelta del tipo di media da utilizzare dipende dal problema che si sta esaminando.

MEDIE II

Diamo, ora, la definizione di media.
Si consideri la seguente distribuzione statistica:

Modalità	Frequenza
x_1	y_1
x_2	y_2
$\dots$	$\dots$
x_n	y_n

Definizione

Si può chiamare **media** di una distribuzione $x_1, x_2, \dots, x_n$ rispetto ad una funzione $f(x_1, x_2, \dots, x_n)$, quella quantità m che, sostituita alle x_i nella funzione, lascia invariato il risultato

MEDIE III

$\Rightarrow$ per calcolare la media abbiamo bisogno di una funzione f, dipendente dalla distribuzione x_i. In poche parole, la funzione $f(x_1, x_2, \ldots, x_n)$ è la "formuletta" che viene utilizzata per il calcolo di tale valore m.

Principalmente abbiamo questi quattro tipi di media, e di ogni tipo ne esiste la corrispettiva ponderata:

- ARITMETICA;
- GEOMETRICA;
- QUADRATICA;
- ARMONICA;

Noi considereremo solamente la media aritmetica, e la sua ponderata.

MEDIA ARITMETICA

Abbiamo già visto che per calcolare la media di una distribuzione, dobbiamo definire la $f(x_1, x_2, \ldots, x_n)$. Nella media aritmetica, tale funzione la indichiamo con M e, considerando la precedente distribuzione statistica, vale che :

$$M = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Dove n è il numero dei dati della distribuzione.

Mentre la corrispettiva ponderata è:

$$M_p = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{y_1 + y_2 + \cdots + y_n}$$

Dove le y_i sono le frequenze assolute del dato x_i.

ESERCIZIO

Si consideri la seguente situazione scolastica:

Voto	Frequenza
2	1
4	3
5	2
6	7
7	9
8	5
9	5
10	2
Totale	34

Calcoliamo la media ponderata della distribuzione.

$$M_p = \frac{10*2+9*5+8*5+7*9+6*7+5*2+4*3+2*1}{2+5+5+9+7+2+1} = \frac{234}{34} = 6,882352941$$

Calcoliamo la media della distribuzione.

$$M = \frac{10+9+8+7+6+5+4+2}{8} = \frac{51}{8} = 6,375$$

INDICI DI POSIZIONE

Gli indici di posizione sono valori segnaletici. Ciò significa che sono utilizzati per sintetizzare in un unico valore significativo molti valori di uno stesso fenomeno, raccolti durante la rilevazione statistica.
La media è un indice di posizione, in quanto è un unico valore che riassume tutti gli altri.
Altri indici di posizione sono la moda e la mediana.

Definizione

Si dice **MODA** di una distribuzione di frequenze la modalità o il valore della variabile al quale corrisponde la massima frequenza.

$\Rightarrow$ la moda è la modalità che si presenta con maggiore frequenza.

Definizione

Si ordinino le modalità in senso non decrescente. La **MEDIANA** è la modalità che bipartisce la sucessione. É il valore che occupa la *posizione centrale* in una serie di modalità numeriche.

ESERCIZIO I

Data la seguente tabella della distribuzione dei pesi dei neonati alla nascita, calcolare il peso medio con la media aritmetica, aritmetica ponderata, la mediana, la moda.

Pesi in grammi	N. neonati
$1.800 \mapsto 2.200$	10
$2.200 \mapsto 2.600$	32
$2.600 \mapsto 3.000$	120
$3.000 \mapsto 3.400$	254
$3.400 \mapsto 3.800$	134
$3.800 \mapsto 4.200$	40
$4.200 \mapsto 4.600$	10
Totale	600

ESERCIZIO II

Andiamo a calcolare la media aritmetica della distribuzione.
Osserviamo che qui sono presenti delle classi di frequenza, non dei singoli valori ben determinati.
$\Rightarrow$ risolviamo il problema calcolando il valore medio di ogni classe di frequenza, ed utilizzando quel valore per i nostri calcoli:

Valor Medio	Pesi in grammi	N. neonati
2.000	$1.800 \mapsto 2.200$	10
2.400	$2.200 \mapsto 2.600$	32
2.800	$2.600 \mapsto 3.000$	120
3.200	$3.000 \mapsto 3.400$	254
3.600	$3.400 \mapsto 3.800$	134
4.000	$3.800 \mapsto 4.200$	40
4.400	$4.200 \mapsto 4.600$	10
	Totale	600

ESERCIZIO III

Allora otteniamo che:

$$M = \frac{2.000+2.400+2.800+3.200+3.600+4.000+4.400}{7} = \frac{22.400}{7} = 3.200$$

$$M_p = \frac{2.000*10+2.400*32+2.800*120+3.200*254+3.600*134+4.000*40+4.400*10}{10+32+120+254+134+40+10} = \frac{1.932.000}{600} = 3.220$$

Moda: $3.000 \mapsto 3.400$

Mediana: $3.000 \mapsto 3.400$

ESERCIZIO IV

Si osservi che, in questo caso, la moda e la mediana non sono dei valori ma sono delle classi di frequenza.

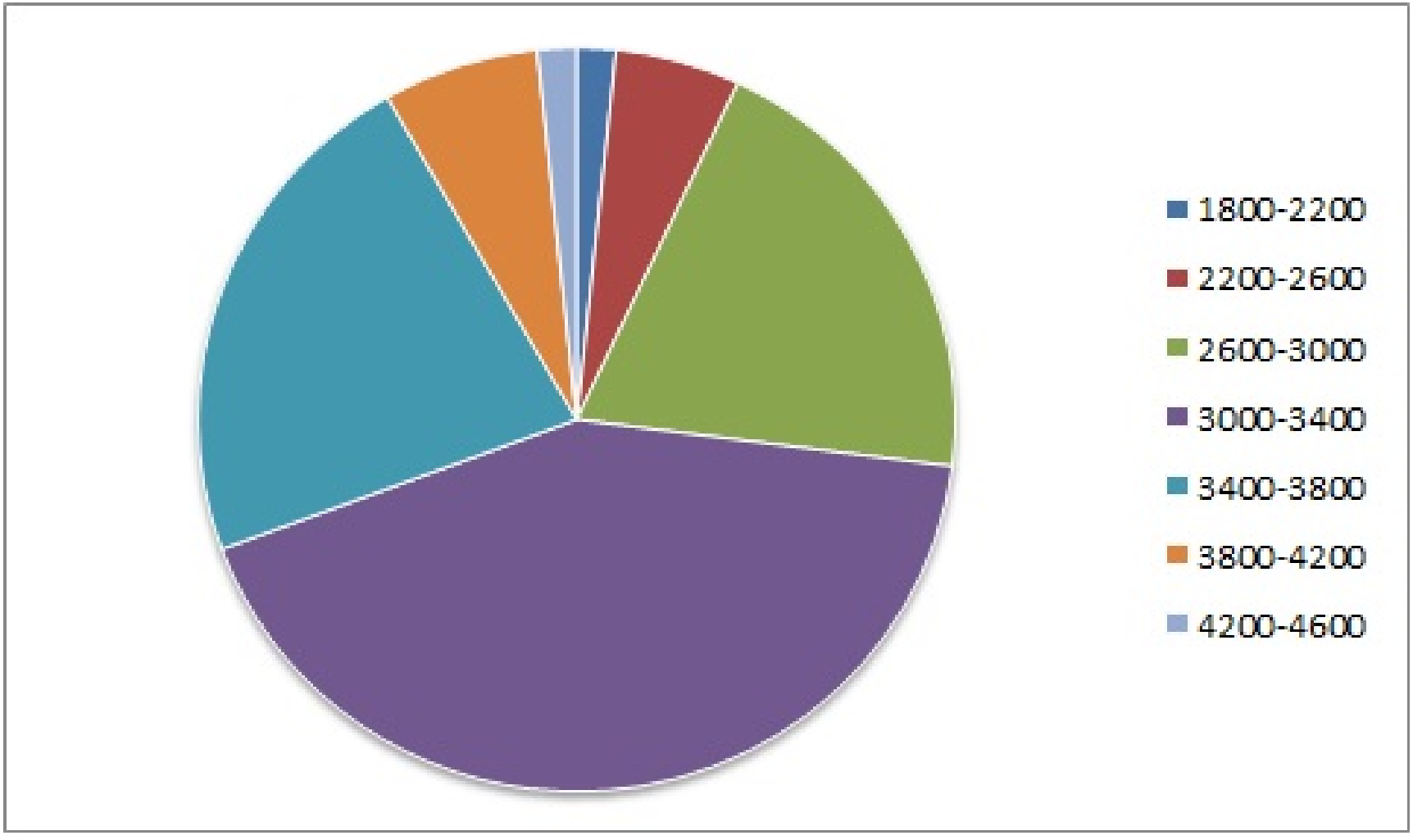
Le classi di frequenza sono molto utilizzate quando si hanno molteplici modalità. Infatti per facilitare lo studio, è possibile scegliere di compattare la distribuzione statistica in classi di frequenza, accorpando assieme tutte quelle modalità che possono avere una qualche caratteristica in comune, oppure nel caso di caratteri quantitativi li raggruppo ad intervalli numerici regolari, come in questo esercizio. La frequenza da attribuire alla classe è la somma delle frequenze delle modalità che rientrano in quella classe.

ESERCIZIO V

Andiamo a completare l'esercizio costruendo il diagramma a torta della distribuzione.
Risolviamo le proporzioni allo scopo di determinare l'angolo al centro di ciascuna classe di frequenza:

$$
\begin{aligned}
10 : 600 = x : 360 \quad & x = 6 \\
32 : 600 = x : 360 \quad & x = 19,2 \\
120 : 600 = x : 360 \quad & x = 72 \\
254 : 600 = x : 360 \quad & x = 154.4 \\
134 : 600 = x : 360 \quad & x = 80.4 \\
40 : 600 = x : 360 \quad & x = 24 \\
10 : 600 = x : 360 \quad & x = 6
\end{aligned}
$$

ESERCIZIO VI



INDICI DI VARIABILITÁ I

Una caratteristica importante dei dati statistici è la **variabilità**.
La maggior parte delle variabili statistiche tendono a disporsi attorno ad un valore medio (che può essere la media aritmetica, la moda o la mediana, per esempio).
Spesso, soprattutto nel caso di distribuzioni molto disomogenee, tali indici di posizione non possono essere considerati indicatori significativi.

$\Rightarrow$ si rende necessario ricorrere allo studio statistico della **dispersione** del fenomeno.

INDICI DI VARIABILITÁ II

Definizione

La **DISPERSIONE** è l'attitudine dei valori di un fenomeno a variare, disponendosi intorno ad un valore medio. La dispersione misura di quanto variano i dati, quanto sono concentrati intorno al valor medio.

Quindi gli **indici di variabilità** sono strumenti utilizzati per misurare la dispersione.
Andiamo a vedere quali sono i principali indici di variabilità.

Definizione

Si definisce **campo di variazione** (o escursione o range), la differenza fra il maggiore ed il minore dei valori rilevati.

INDICI DI VARIABILITÁ III

Definizione

Si definisce **scarto quadratico medio** o **deviazione standard** la radice quadrata della media aritmetica (eventualmente ponderata) degli scarti dalla media elevati al quadrato.

Chiariamo quest'ultimo concetto. Andiamo a definire lo scarto dalla media del valore x_i come la
differenza tra il valore e la media aritmetica: $x_i - M$.
Pertanto, lo scarto quadratico medio, è la radice della media del quadrato degli scarti, in formule:

$$\sigma = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + \cdots + (x_n - M)^2}{n}}$$

questo nel caso faccia la media artimetica del quadrato degli scarti.

INDICI DI VARIABILITÁ IV

Però la definizione mi dice che posso anche fare la media ponderata di tali quantita; indicando con y_i la frequenza dell'i-esima variabile, ho che:

$$\sigma = \sqrt{\frac{(x_1 - M)^2 y_1 + (x_2 - M)^2 y_2 + \cdots + (x_n - M)^2 y_n}{y_1 + y_2 + \cdots + y_n}}$$

Definizione

Si definisce **varianza** il quadrato dello scarto quadratico medio.

Dalla definizione segue che la varianza è σ^2.
Tramite alcuni passaggi, però, otteniamo un'altra formulazione per la varianza:

$$\sigma^2 = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - M^2$$

INDICI DI VARIABILITÁ V

CHE SIGNIFICATO HANNO QUESTI ULTIMI DUE INDICI?

Tali indici valutano la maggiore o minore dispersione dei valori dalla media aritmetica, è il valore medio degli scarti.

Lo scarto quadratico medio è tanto più piccolo quanto più i dati sono prossimi al valore medio, ed è uguale a zero se e solo se i dati sono tutti uguali tra loro.

Lo scarto quadratico medio è un indice della dispersione dei dati molto sensibile per evidenziare dati che si scostano molto dal valore medio.

INDICI DI VARIABILITÁ VI

$\Rightarrow$ se σ è piccolo, i valori della distribuzione sono concentrati attorno alla media aritmetica; se è grande, i valori sono molto dispersi attorno alla media.

La varianza è un indice meno immediato perchè ho un elevamento al quadrato. É un indice molto utilizzato nella statistica inferenziale.