Le fasi di una indagine statistica

La statistica studia i metodi per raccogliere, elaborare ed interpretare i dati relativi ad un fenomeno collettivo. Si dice che un fenomeno è collettivo o di massa quando l'indagine su quel fatto richiede una molteplicità di osservazioni su fenomeni o fatti individuali aventi tutti un carattere comune. Sono fenomeni collettivi la distribuzione del reddito dei cittadini di una nazione, la distribuzione per età dei medesimi o quelli della loro statura.

Si chiama **collettività statistica** o popolazione statistica o **universo statistico** un insieme di elementi considerati omogenei rispetto ad uno o più caratteri. Ogni elemento appartenente alla popolazione statistica prende il nome di **unità statistica**. Nello svolgimento di una qualsiasi indagine statistica distinguiamo le seguenti fasi:

1) **definizione degli obiettivi** 2) **formulazione delle ipotesi, cioè individuazione del fenomeno collettivo che si vuole sottoporre a verifica** 3) **rilevazione dei dati** 4) spoglio dei dati 5) **elaborazione dei dati** 6) **interpretazione dei risultati e divulgazione**.

Quando non è possibile o non conviene effettuare indagini sull'intera popolazione statistica, allora si procede per **campione**. Eseguita la scelta del campione, si proseguirà alla **rilevazione** statistica vera e propria, che consiste nell'annotare su apposite schede, opportunamente approntate e che risultano diverse a seconda dell'indagine che si intende compiere, i dati che si vogliono raccogliere. La raccolta dei dati deve essere il più possibile **esatta e completa**. (*) La mancanza di questi requisiti compromette l'esito dell'indagine. Effettuata la rilevazione si esegue lo **spoglio** delle schede ed i dati raccolti vengono evidenziati per mezzo di: a) **tabulazioni** che consistono nel sistemare in tabelle i risultati dello spoglio b) **tabulazioni** che consistono nel sistemare in tabelle i risultati dello spoglio c) **rappresentazioni grafiche** che visualizzano l'andamento di un dato fenomeno.

Dopo lo spoglio viene effettuata l'**elaborazione dei dati** che consiste nella scelta e nell'applicazione dei procedimenti matematici propri del metodo statistico. Infine abbiamo

(*) Si chiama **rilevamento statistico** l'operazione di raccolta dei dati statistici riguardanti una determinata popolazione di individui in possesso di un medesimo carattere, relativo al fenomeno che si vuole studiare. Un rilevamento statistico può essere saltuario o continuo, parziale o totale, pubblico o privato. Per i rilevamenti statistici sono necessari appositi registri, moduli, questionari, schede...su cui riportare e ordinare tutte le informazioni raccolte sul carattere o i caratteri del fenomeno in esame, per potere, alla fine, costruire le cosiddette **tabelle statistiche** relative

l'**interpretazione dei risultati**, e questa parte del procedimento statistico è interamente affidata alla perizia, al senso critico ed alla capacità interpretativa del ricercatore.

Di ogni unità statistica si studiano i caratteri, la cui scelta dipende dal particolare problema che si vuole esaminare e dalle ipotesi iniziali. Ogni **carattere** si manifesta mediante le sue modalità cioè mediante i diversi modi di apparire, ciascuno dei quali è chiamato **modalità**. I caratteri possono essere **qualitativi** e **quantitativi**, secondo che le modalità con cui si presentano sono espresse mediante attributi, aggettivi, denominazioni, oppure sono espresse mediante valori numerici. Sono caratteri qualitativi, il titolo di studio posseduto da ciascuna unità statistica, il tipo di attività esercitato dai lavoratori di un'azienda, la regione di residenza degli italiani. Sono caratteri quantitativi l'altezza delle persone di una provincia, il numero di studenti inscritti ad un determinato tipo di scuola, la quantità di pioggia caduta in un determinato periodo dell'anno, il reddito pro-capite per le diverse categorie di lavoratori. I **caratteri quantitativi** possono essere **discreti** (quando possono assumere valori appartenenti ad insiemi finiti o infiniti ma numerabili) o **continui** (quando hanno modalità appartenenti ad un intervallo di numeri reali). I caratteri di tipo quantitativo sono chiamati anche variabili statistiche e vengono indicate con lettere maiuscole come **X** o **Y** mentre i caratteri qualitativi prendono il nome di mutabili statistiche e, di solito, vengono indicate con la lettera maiuscola **M.** Il dato statistico esprime il numero totale di unità statistiche che presentano la stessa modalità. Il dato statistico coincide con la frequenza o la intensità di una modalità di un carattere, cioè di una variabile o di un a mutabile statistica.

La tabella mostra la distribuzione per classi di età di 61 dipendenti di un'azienda. Ogni dipendente è una unità statistica. I 61 dipendenti rappresentano la popolazione statistica. L'età di ciascun dipendente è il carattere della popolazione statistica;

Classi di età	Numero dipendenti
18-24	12
24-30	14
30-36	6
36-42	6
42-48	11
48-54	9
54-60	3

Poiché tale carattere è espresso da un numero esso individua una variabile statistica. Le varie classi di età sono le modalità della variabile statistica. Il numero di dipendenti per ciascuna classe di età rappresenta la frequenza assoluta della modalità.

La tabella mostra il mezzo di trasporto utilizzato da 250 studenti per andare a scuola.

Ogni alunno è una unità statistica. I 250 alunni rappresentano la popolazione statistica. Il mezzo di trasporto utilizzato è il carattere della popolazione statistica.

mezzo di trasporto	numero alunni
a piedi	115
autobus	38
motorino	43
bicicletta	26
auto	28

Poiché tale carattere non è espresso da un numero esso individua una mutabile statistica. I vari mezzi di trasporto sono le modalità della mutabile statistica. Il numero di alunni per ogni mezzo di trasporto utilizzato rappresenta la frequenza assoluta della modalità.

La frequenza assoluta della modalità di un carattere

Quando i dati sono molti, ogni singola modalità si presenta più volte. In tali casi si costruisce una tabella con due colonne. Nella prima colonna si scrivono le diverse modalità del carattere. Nella seconda colonna scriviamo le frequenze di quei valori cioè scriviamo il numero di volte in cui quel dato compare nella raccolta. Una tabella di questo tipo è la seguente e prende il nome di distribuzione di frequenze:

modalità	frequenza
x_1	f_1
x_2	f_2
x_3	f_3
…	…
x_k	f_k

k rappresenta il numero delle modalità del carattere considerato, f_3 rappresenta quante volte si è presentata la modalità x_3 della variabile statistica X.

Definizione: Si chiama frequenza assoluta o semplicemente frequenza di una modalità il numero di volte che essa compare nel collettivo statistico osservato.

La somma delle frequenze assolute è uguale al numero N di unità statistiche considerate, cioè:

$$f_1 + f_2 + f_3 + \cdots\cdots + f_k = N$$

La frequenza relativa

Definizione: Si chiama **frequenza relativa** di una modalità il rapporto tra la sua frequenza assoluta e il numero N di unità statistiche del collettivo osservato. $r_k = \frac{f_k}{N}$

La somma delle frequenze relative è uguale ad 1, cioè:

$$r_1+r_2+r_3+\cdots+r_k=\frac{f_1}{N}+\frac{f_2}{N}+\frac{f3}{N}+\cdots+\frac{f_k}{N}=\frac{f_1+f_2+f_3+\cdots f_k}{N}=\frac{N}{N}=1$$

Se il valore viene rapportato a 100 si ha una **frequenza relativa percentuale**. $p_k=r_k\cdot 100=\frac{f_k}{N}\cdot 100$

La somma di tutte le frequenze percentuali è uguale a 100, cioè:

$$p_1+p_2+p_3+\cdots+p_k=\frac{f_1}{N}\cdot 100+\frac{f_2}{N}\cdot 100+\frac{f_3}{N}\cdot 100+\cdots+\frac{f_k}{N}\cdot 100=\frac{f_1+f_2+f_3+\cdots f_k}{N}\cdot 100=\frac{N}{N}\cdot 100=100$$

In molti casi può essere utile conoscere quante sono le unità statistiche che presentano un valore del carattere (variabile o mutabile statistica) minore oppure uguale ad una certa modalità x_k. Conviene introdurre le frequenze cumulate che possono essere **frequenze cumulate assolute**, **frequenze cumulate relative**, **frequenze cumulate percentuali**. Si definiscono frequenze cumulate che corrispondono alla modalità x_k la somma di tutte le frequenze della modalità x_k e di quelle che la precedono. Questo significa che la **frequenza cumulata** rispetto ad una modalità x_k coincide col numero di unità statistiche che presentano una modalità minore o uguale ad x_k. Possiamo avere

$C_k = f_1 + f_2 +\cdots+ f_k$ = frequenze cumulate assolute

$F_k=\frac{f_1}{N}+\frac{f_2}{N}+\cdots+\frac{f_k}{N}=r_1+r_2+\cdots+r_k$ = frequenze cumulate relative

$P_k=100\cdot F_k=\frac{n_1}{N}\cdot 100+\frac{n_2}{N}\cdot 100+\cdots+\frac{n_k}{N}\cdot 100=p_1+p_2+\cdots+p_k$ =

= frequenze cumulate percentuali

Dalla definizione di frequenza cumulata deriva che il valore relativo all'ultima modalità è uguale al totale complessivo N per le frequenze assolute, ad 1 per le frequenze relative, a 100 per quelle percentuali.

Modalità	*Frequenze assolute* f_i	*Frequenze assolute cumulate* F_i	*Frequenze relative cumulate* R_i	*Frequenze percentuali cumulate* $R_i \cdot 100$
5.110 ⊣ 5.115	1	1	0.007	0.7
5.115 ⊣ 5.120	2	3	0.020	2
5.120 ⊣ 5.125	5	8	0.053	5.3
5.125 ⊣ 5.130	13	21	0.140	14.0
5.130 ⊣ 5.135	25	46	0.307	30.7
5.135 ⊣ 5.140	26	72	0.480	48.0
5.140 ⊣ 5.145	22	94	0.627	62.7
5.145 ⊣ 5.150	20	114	0.760	76.0
5.150 ⊣ 5.155	18	132	0.880	88.0
5.155 ⊣ 5.160	12	144	0.960	96.0
5.160 ⊣ 5.165	4	148	0.987	98.7
5.165 ⊣ 5.170	2	150	1	100
TOTALE	150			

Si chiama **distribuzione di frequenze** l'insieme delle coppie ordinate il cui primo elemento corrisponde alla modalità del carattere ed il secondo elemento alla sua frequenza assoluta, relativa, percentuale o cumulata.

Una distribuzione di frequenze è una distribuzione che ha una delle seguenti strutture:

$$X\begin{Bmatrix} x_1 & x_2 & \cdots & x_k \\ f_1 & f_2 & \cdots & f_k \end{Bmatrix}$$ distribuzione di frequenza assoluta

la prima riga rappresenta le modalità del carattere, la seconda le frequenze assolute di ogni singola modalità. Di solito tale struttura ha la forma di una tabella

$$X\begin{Bmatrix} x_1 & x_2 & \cdots & x_k \\ r_1 & r_2 & \cdots & r_k \end{Bmatrix}$$ distribuzione di frequenza relativa

$$X\begin{Bmatrix} x_1 & x_2 & \cdots & x_k \\ F_1 & F_2 & \cdots & F_k \end{Bmatrix}$$ distribuzione di frequenza relativa cumulata

$$X\begin{Bmatrix} x_1 & x_2 & \cdots & x_k \\ P_1 & P_2 & \cdots & P_k \end{Bmatrix}$$ distribuzione di frequenza percentuale cumulata

Le tabelle che riportano nella prima colonna un carattere quantitativo (variabile statistica) vengono dette seriazioni statistiche, quelle che riportano un carattere qualitativo (**mutabile statistica)** vengono dette serie statistiche.

In generale, indicato con X il carattere (variabile o mutabile statistica) che si deve studiare, con x_i le n **modalità** con cui esso si può presentare, con f_i le **frequenze assolute** di tali modalità, con p_i le **frequenze relative**, la distribuzione di frequenza del carattere X si rappresenta una tabella del tipo indicato a fianco.

X	f	p
x_1	f_1	p_1
x_2	f_2	p_2
⋮	⋮	⋮
x_i	f_i	p_i
⋮	⋮	⋮
x_n	f_n	p_n
TOTALI	N	1

La seguente tabella riporta le frequenze assolute, le frequenze assolute cumulate, le frequenze relative cumulate, le frequenze percentuali cumulate dei voti di italiano riportati dagli alunni di un liceo.

voto dello scrutinio finale modalità x_i	**Studenti che hanno riportato tale voto in italiano**					
	frequenze					
	assolute f_i	**assolute cumulate** C_i	**relative** r_i	**relative cumulate** Fi	**percentuali** p_i	**percentuali cumulate** Pi
$x_1 = 3$	$f_1 = 10$	$C_1 = 10$	$r_1 = 0{,}03$	$F_1 = 0{,}03$	$p_1 = 3$	$P_1 = 3$
$x_2 = 4$	$f_2 = 25$	$C_2 = 35$	$r_2 = 0{,}08$	$F_2 = 0{,}11$	$p_2 = 8$	$P_2 = 11$
$x_3 = 5$	$f_3 = 34$	$C_3 = 69$	$r_3 = 0{,}12$	$F_3 = 0{,}23$	$p_3 = 12$	$P_3 = 23$
$x_4 = 6$	$f_4 = 136$	$C_4 = 205$	$r_4 = 0{,}46$	$F_4 = 0{,}68$	$p_4 = 46$	$P_4 = 69$
$x_5 = 7$	$f_5 = 68$	$C_5 = 273$	$r_5 = 0{,}23$	$F_5 = 0{,}91$	$p_4 = 23$	$P_5 = 92$
$x_6 = 8$	$f_6 = 22$	$C_6 = 295$	$r_6 = 0{,}07$	$F_6 = 0{,}98$	$p_6 = 7$	$P_6 = 99$
$x_7 = 9$	$f_7 = 3$	$C_7 = 298 = N$	$r_7 = 0{,}01$	$F_7 = 1$	$p_7 = 1$	$P_7 = 100$
	$N = 298$					

Il numero degli studenti insufficienti è 69 e tale numero corrisponde alla frequenza assoluta cumulata relativa alla modalità 5 .

La variabile statistica X considerata è: voti di italiano riportati dagli alunni di un liceo **unità statistica** = ogni alunno del liceo ; **carattere unità statistica** = voto di italiano nello scrutinio finale ; **modalità del carattere** = 3,4,5,6,7,8,9 (si tratta di un carattere quantitativo)

Il **dato statistico** rispetto alla modalità 3 è 10 , rispetto alla **modalità** 8 è 22 .

In una indagine statistica si vuole analizzare il numero di persone di sesso maschile presenti nelle famiglie italiane. L'indagine viene svolta su un campione di 1000 famiglie ed i risultati sono rappresentati dalla tabella di frequenza della figura. La **variabile** X è il **numero di maschi presenti nelle famiglie italiane** e, poiché essa può assumere solo valori interi positivi, si tratta di una variabile statistica discreta.

Numero dei maschi nelle famiglie	Frequenza assoluta (f_i)	Frequenza relativa (p_i)
0	50	0,05
1	120	0,12
2	300	0,3
3	250	0,25
4	190	0,19
5	60	0,06
6	20	0,02
7	10	0,01
TOTALI	1000	1

In una azienda vengono raccolti i dati relativi al **numero di ore di lavoro mensili** effettuate;tali dati sono facilmente reperibili dai tabulati delle ore lavorative di ogni dipendente. La **distribuzione** che ne risulta è presente nella tabella. La variabile X è ilo mese ed è una variabile di tipo qualitativo; si tratta di una mutabile statistica.

Mese	N. ore lavorate (f_i)	Percentuale nel mese (p_i)
Gennaio	12360	0,0701
Febbraio	15865	0,0900
Marzo	15940	0,0905
Aprile	15758	0,0894
Maggio	16075	0,0912
Giugno	16124	0,0915
Luglio	15635	0,0887
Agosto	4520	0,0257
Settembre	15942	0,0905
Ottobre	16214	0,0920
Novembre	16120	0,0915
Dicembre	15658	0,0889
TOTALI	176211	1

Distribuzioni di frequenza per classi

La suddivisione di una variabile statistica in classi di frequenza scaturisce dallo spoglio di un carattere quantitativo continuo oppure dallo spoglio di un carattere quantitativo discreto con un gran numero di modalità distinte. Un carattere quantitativo è continuo quando può assumere tutti i valori di un certo intervallo $[a,b]$. Spesso si fa ricorso alla suddivisione dell'intervallo $[a,b]$ in classi di ampiezza uniforme, dividendo l'intervallo $[a,b]$ in n intervalli parziali aventi tutti la stessa ampiezza. Tali classi si chiamano classi di intensità e le frequenze relative classi di frequenze. I primi $n-1$ intervalli parziali si considerano chiusi a sinistra ed aperti a destra, mentre l'ultimo intervallo parziale è chiuso sia a destra che a sinistra.

Nel caso di un carattere discreto sarà opportuno non fare coincidere gli estremi superiori di ciascuna classe con quelli inferiori delle classi successive. Converrà che gli estremi inferiori delle classi siano di un'unità in più rispetto a quelli superiori delle classi precedenti. Ecco due esempi di variabili statistiche suddivise in classi di frequenze.

Distribuzione in classi di una popolazione di 10000 individui secondo la statura (**carattere continuo**), suddivisa in classi di frequenza di ampiezza 10 *cm* 10 a partire da 120 *cm*. La prima classe indica i valori compresi tra 120 *cm* incluso e 130 *cm*; l'ultima classe indica i valori compresi tra 190 *cm* incluso e 200 *cm* incluso.	Statura in cm	frequenza
	120⊢ 130	3
	130⊢ 140	14
	140⊢ 150	138
	150⊢ 160	2334
	160⊢ 170	5645
	170⊢ 180	1735
	180⊢ 190	111
	190⊢ 200	2
	Totale	10000

Distribuzione in classi di un gruppo di cliniche private secondo il numero di posti letto	Numero di posti letto	Cliniche
	Fino a 25	23
	26 – 50	158
	51 – 75	134
	76 – 100	101
	101 – 125	43
	126 – 150	50
	151 – 200	27
	201 – 250	10
	251 – 350	11
	Più di 351	3
	Totale	560

La rappresentazione grafica delle distribuzioni di frequenza

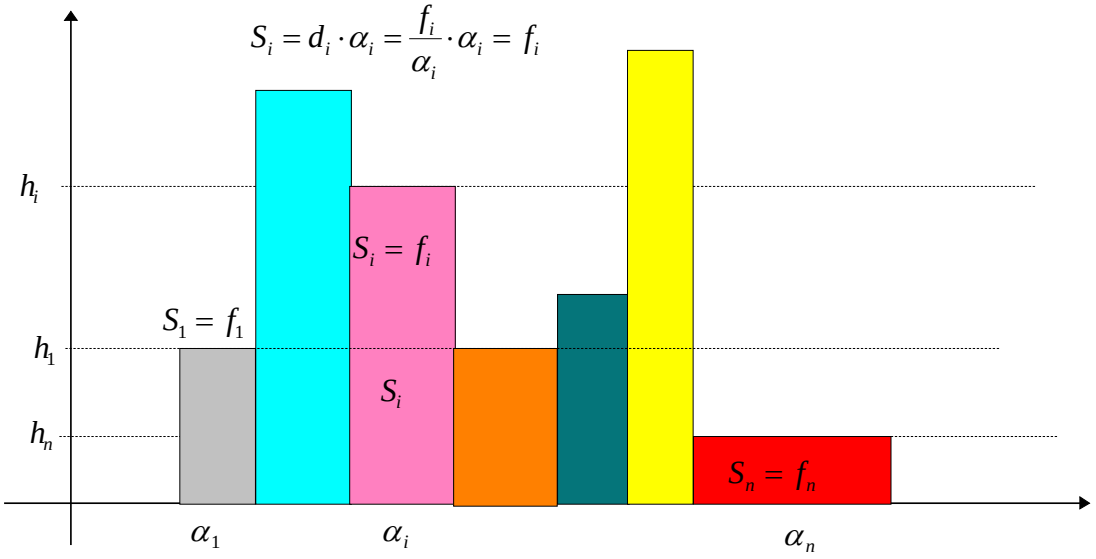
Gli istogrammi

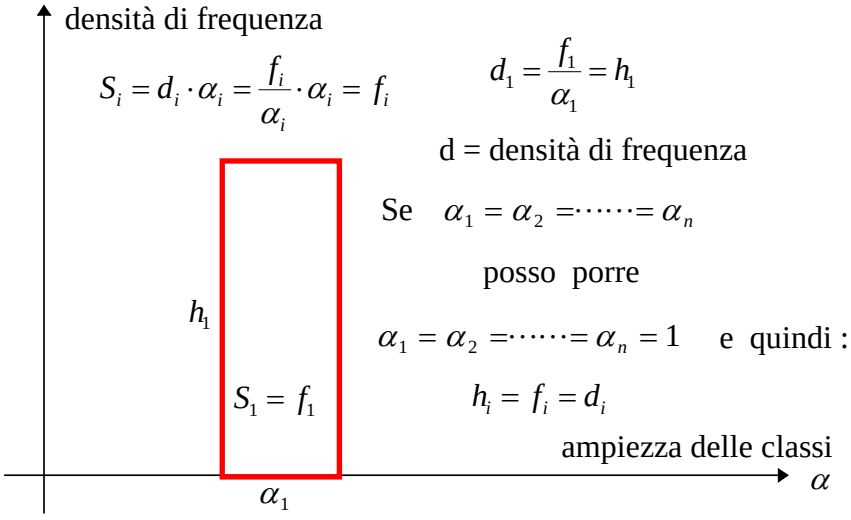
Gli istogrammi vengono utilizzati quando dobbiamo rappresentare graficamente una distribuzione statistica dove le modalità del carattere sono ripartite in n classi di varia ampiezza.

Definiamo densità di frequenza il rapporto tra la frequenza della classe e l'ampiezza della classe stessa. $d_1 = \frac{f_1}{\alpha_1}$ = densità di frequenza della classe di ampiezza α_1

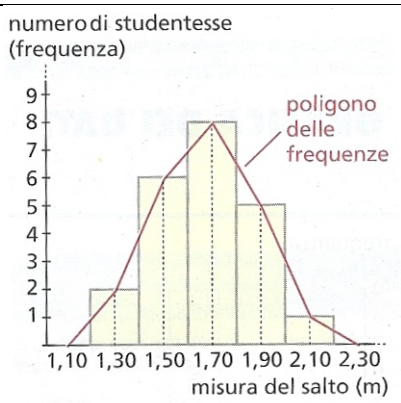
Gli **istogrammi** sono costituiti da una serie di rettangoli contigui che si sviluppano lungo l'asse orizzontale, ed hanno come basi le ampiezze delle classi e come altezze le rispettive densità di frequenza. L'area dell'intero istogramma ci fornisce il totale complessivo N dell'intera popolazione statistica, mentre l'area di ciascun rettangolo ci fornisce la frequenza assoluta di ciascuna classe.

E' facile verificare che, quando le basi sono uguali, l'area della parte di piano delimitata dall'asse orizzontale e dal poligono di frequenza è uguale alla somma delle aree dei rettangoli. Quindi in ascissa riportiamo le ampiezze delle varie classi ed in ordinata i valori delle corrispondenti densità di frequenza. Se tutte le classi hanno la stessa ampiezza possiamo riportare in ordinata le frequenze assolute convenendo di porre numericamente uguale ad uno l'ampiezza di ciascuna classe.





Se in un istogramma si uniscono i punti medi delle basi superiori dei rettangoli si ottiene una spezzata detta **poligono di frequenza**. Il poligono è esteso anche alle classi estreme aventi frequenza nulla.



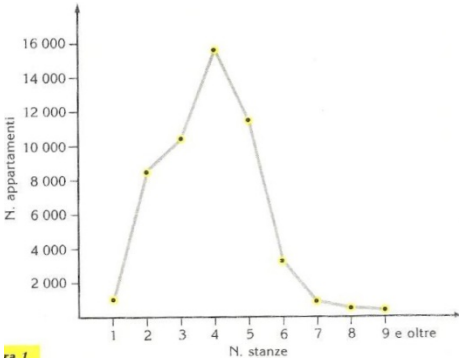
Rappresentazione cartesiana

Si riferisce il piano ad un sistema di assi cartesiani ortogonali. Si riporta sull'asse delle ascisse il carattere e sull'asse delle ordinate la frequenza, cioè il numero che indica quante volte quel carattere si è presentato.

La tabella mostra il risultato di un'indagine sul numero delle abitazioni esistenti in una piccola città e sul relativo numero di stanze.

N. stanze	N. abitazioni
1	950
2	8 450
3	10 352
4	15 624
5	11 430
6	3 250
7	720
8	435
9 e oltre	110

Per rappresentare in un sistema di riferimento cartesiano la nostra tabella, riportiamo in ascissa il numero di stanze di ciascun appartamento ed in ordinata il numero degli appartamenti. Le unità di misura per le ascisse e le ordinate sono necessariamente diverse. Il sistema cartesiano è dimetrico. Nel piano si ottengono dei punti che, congiunti con segmenti di retta, costituiscono una spezzata che è la rappresentazione cartesiana del fenomeno.



Areogramma o diagramma circolare o diagramma a torta

Un diagramma di questo tipo si serve di un cerchio che viene diviso in settori circolari di ampiezza proporzionale alla frequenza (di solito relativa ed espressa in forma percentuale per comodità di calcolo nella individuazione dei vari settori). Questo tipo di grafico è particolarmente utile per rappresentare le frequenze percentuali. Un cerchio viene suddiviso in tanti settori circolari, ognuno dei quali corrisponde ad una frequenza. Gli angoli al centro α_i dei diversi settori hanno ampiezza proporzionale alla frequenza percentuale p_i (o assoluta, o relativa). L'angolo al centro di ogni settore va calcolato applicando la seguente proporzione:

$\alpha_1 : 360° = p_1 : 100 \quad \alpha_1 = \frac{p_1}{100} \cdot 360°$ = angolo al centro del primo settore corrispondente alla prima frequenza percentuale p_1. In maniera simile si opera per le rimanenti frequenze percentuali.

La seguente tabella ci fornisce una distribuzione di frequenze ottenuta analizzando i risultati realizzati da un gruppo di studenti che, nell'ora di educazione fisica, hanno eseguito una prova di salto in lungo da fermo.

classe	frequenza assoluta f_i	frequenza relativa r_i	frequenza percentuale p_i
$120-140$	2	$\frac{2}{22}=0,09$	9%
$1,40-1,60$	6	$\frac{6}{22}=0,27$	27%
$1,60-1,80$	8	$\frac{8}{22}=0,36$	36%
$1,80-2,00$	5	$\frac{5}{22}=0,23$	23%
$2,00-2,20$	1	$\frac{1}{22}=0,05$	5%
	22	1	100

$\alpha_1 : 360° = 27 : 100 \Rightarrow \alpha_1 = \frac{9}{100} \cdot 360° = 32,4°$

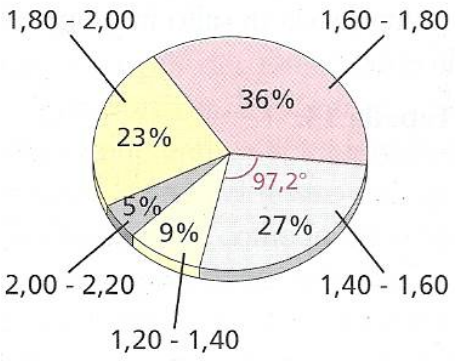
$\alpha_2 : 360° = 27 : 100 \Rightarrow \alpha_2 = \frac{27}{100} \cdot 360° = 97,2°$

$\alpha_3 : 360° = 36 : 100 \Rightarrow \alpha_3 = \frac{36}{100} \cdot 360° = 129,6°$

$$\alpha_4 : 360° = 23 : 100 \Rightarrow \alpha_4 = \frac{23}{100} \cdot 360° = 82{,}8°$$

$$\alpha_5 : 360° = 5 : 100 \Rightarrow \alpha_4 = \frac{5}{100} \cdot 360° = 18°$$

Areogramma di una distribuzione di frequenze ottenuta analizzando i risultati ottenuti da un gruppo di studenti che, nell'ora di educazione fisica, hanno eseguito una prova di salto in lungo da fermo.



Un'indagine statistica riguardante "il tipo di vacanza preferito" su un campione di 250 persone sono scaturite le informazioni esposte nella seguente tabella. Dopo avere calcolato le frequenze percentuali effettua una sua rappresentazione grafica mediante un areogramma.

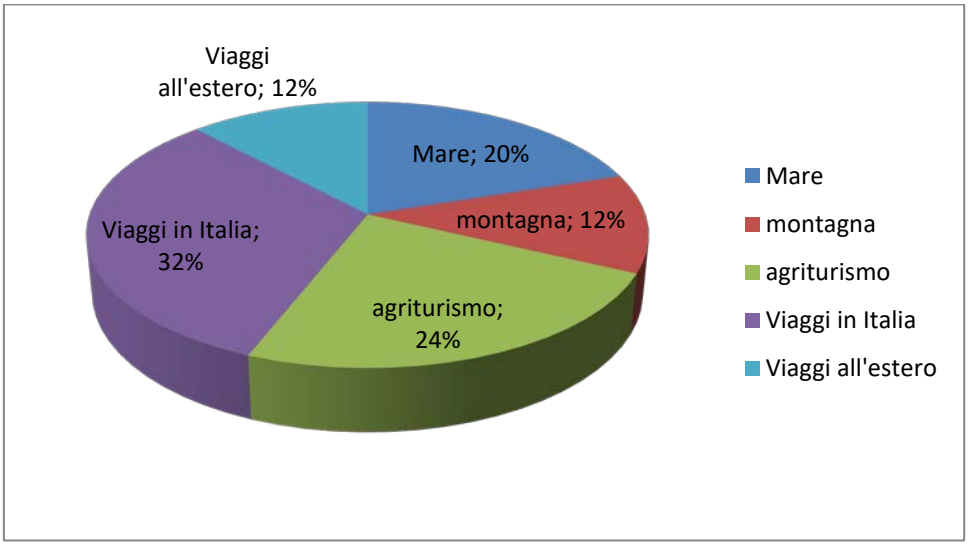
Tipo di vacanza	Mare	Montagna	Agriturismo	Viaggi in Italia	Viaggi all'estero
Frequenza assoluta	50	30	60	80	30

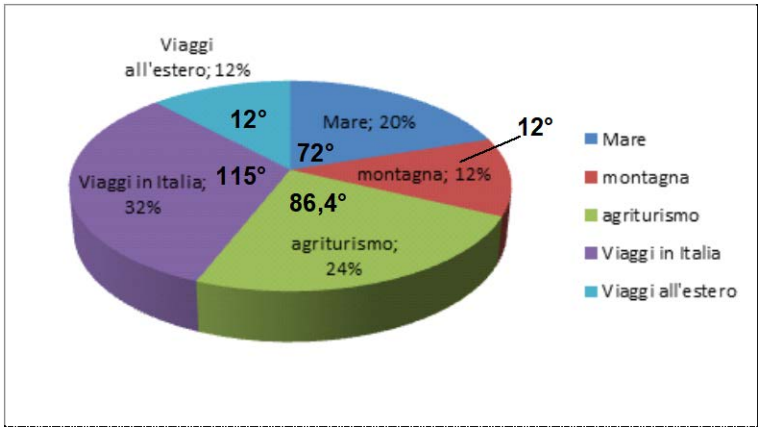
Tipo di vacanza	frequenza assoluta f_i	frequenza relativa r_i	frequenza percentuale p_i
mare	50	$\frac{50}{250} = 0{,}2$	20%
montagna	30	$\frac{30}{250} = 0{,}12$	12%
agriturismo	60	$\frac{60}{25} = 0{,}24$	24%
Viaggi in Italia	80	$\frac{80}{250} = 0{,}32$	32%
Viaggi all'estero	30	$\frac{30}{250} = 0{,}12$	12%
	250	1	100

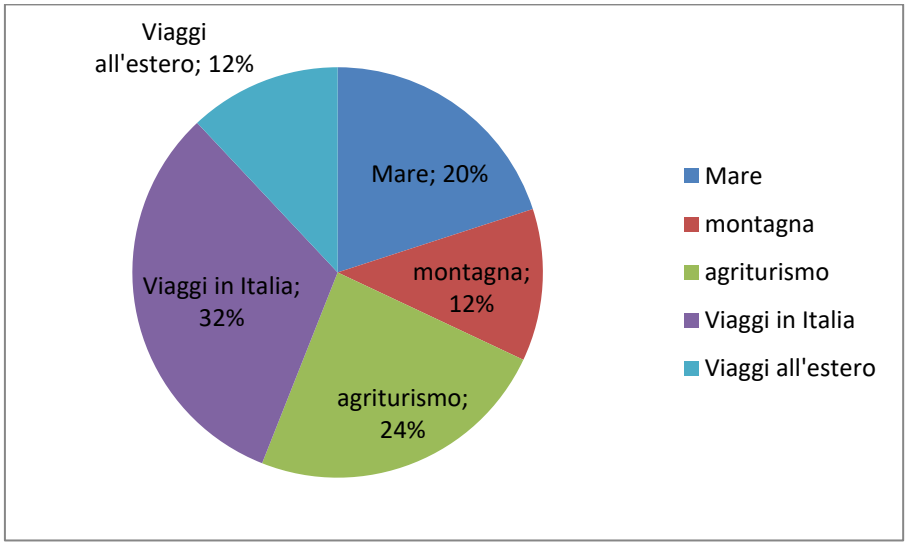
$$\alpha_1 = \frac{p_1}{100} \cdot 360 = \frac{20}{100} \cdot 360 = 72° \qquad \alpha_2 = \frac{p_2}{100} \cdot 360 = \frac{12}{100} \cdot 360 = 43,2°$$

$$\alpha_3 = \frac{p_3}{100} \cdot 360 = \frac{24}{100} \cdot 360 = 86,4° \qquad \alpha_4 = \frac{p_4}{100} \cdot 360 = \frac{32}{100} \cdot 360 = 115,2°$$

$$\alpha_5 = \frac{p_5}{100} \cdot 360 = \frac{12}{100} \cdot 360 = 43,2°$$






Viaggi all'estero; 12%
Mare; 20%
montagna; 12%
agriturismo; 24%
Viaggi in Italia; 32%
Mare
montagna
agriturismo
Viaggi in Italia
Viaggi all'estero

Glossario di statistica

● Universo statistico o popolazione statistica ● Unità statistica ● Carattere statistico ● modalità di un carattere ● frequenze assolute ● frequenze relative ● frequenze percentuali ● frequenze cumulate ● distribuzione di frequenza ● classi di frequenza ● variabile statistica (il carattere è quantitativo) ● mutabile statistica (il carattere è qualitativo) ● rilevazione ● spoglio ● elaborazione ● un carattere di una popolazione statistica è descritto mediante modalità che possono essere di tipo qualitativo o quantitativo

La media aritmetica

La media aritmetica **m** di n numeri $x_1, x_2, \cdots, x_n$ è il **quoziente** fra la loro somma ed il numero n. In simboli abbiamo: $m = \frac{x_1 + x_2 + ... + x_n}{n}$

Esempio: una famiglia ha speso nei successivi 7 giorni di una settimana le seguenti cifre (in euro): 28; 13,50; 10,50; 30; 18; 50; 60. Quanto ha speso mediamente ogni giorno?

$$m = \frac{20 + 13{,}50 + 10{,}50 + 30 + 18 + 50 + 60}{7} = \frac{202}{7} = 28{,}857\,€$$

Se in una distribuzione di frequenze il valore x_1 compare f_1 volte, il valore x_2 compare f_2 volte, e così di seguito, la media aritmetica va calcolata applicando la seguente formula:

$$m = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + ... + f_n \cdot x_n}{f_1 + f_2 + ... + f_n} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + ... + f_n \cdot x_n}{N}$$

La media aritmetica così calcolata prende il nome di media aritmetica ponderata. Essa è denominata così in quanto ogni modalità x_i interviene nel calcolo con un **peso** pari alla sua frequenza.

Esempio: determinare l'età media dei 30 alunni di una classe così ripartiti:

Età, in anni	$x_1 = 14$	$x_2 = 15$	$x_3 = 16$	$x_4 = 17$	$x_5 = 18$	$x_6 = 19$
Numero di alunni	$f_1 = 4$	$f_2 = 12$	$f_3 = 8$	$f_4 = 4$	$f_5 = 1$	$f_6 = 1$

Siamo in presenza di una media aritmetica ponderata, che calcoliamo con la seguente formula:

$$m=\frac{4\cdot 14+12\cdot 15+8\cdot 16+4\cdot 17+1\cdot 18+1\cdot 19}{4+12+8+4+1+1}=\frac{469}{30}=15,6333=15^{a}7^{m}18^{g}$$

Osservazione: Particolare attenzione va posta quando dobbiamo calcolare la media aritmetica ponderata di una distribuzione di frequenze suddivisa in classi. Una scelta conveniente è quella di sostituire ciascuna classe col suo termine centrale, cioè con la semisomma del due estremi della classe considerata. Nella pratica si procede così: a) si calcola il termine centrale di ciascuna classe b) i termini centrali così ottenuti vengono assunti come termini della distribuzione statistica c) si moltiplicano i termini centrali per le rispettive frequenze, si sommano questi prodotti e poi dividiamo la somma ottenuta per la somma totale delle frequenze ottenendo la media aritmetica ponderata richiesta.

Esempio: Calcolare la statura media di un gruppo di giovani come indicato nella seguente tabella:

Classe di statura (in cm)	Frequenza	Valore centrale	Prodotto $x_i f_i$
155-160	0	157,5	0
160-165	20	162,5	3.250
165-170	52	167,5	8.710
170-175	30	172,5	5.175
175-180	12	177,5	2.130
180-185	1	182,5	182,5
Totali	115	–	19.447,5

L'età media richiesta è la media aritmetica ponderata dei valori:

$x_1=157,5;\ x_2=162,5;\ x_3=167,5;\ x_4=172,5;\ x_5=177,5;\ x_6=182,5$ con i pesi rispettivi:

$f_1=0;\ f_2=20;\ f_3=52;\ f_4=30;\ f_5=12;\ f_6=1$

$$m=\frac{0\cdot 157,5+20\cdot 162,5+52\cdot 167,5+30\cdot 172,5+12\cdot 177,5+1\cdot 182,5}{++20+52+30+12+1}=\frac{19447,5}{115}=169,11cm$$

Finora abbiamo calcolato medie aritmetiche di variabili statistiche, cioè di caratteri che si esprimono mediante numeri. Adesso calcoliamo la media aritmetica di una mutabile statistica. Se le modalità della mutabile statistica sono k e le unità statistiche sono N, allora la media aritmetica va calcolata utilizzando la seguente formula: $m=\frac{N}{k}$

La tabella indica la quantità di merce che un'azienda ha venduto nei primi 5 mesi dell'anno. Essendo 5 le modalità del carattere, abbiamo: $m=\frac{10105}{5}=2021$

Mese	Quantità venduta
Gennaio	1.840
Febbraio	2.020
Marzo	1.980
Aprile	2.160
Maggio	2.105
Totale	**10.105**

La mediana

Date n grandezze $x_1, x_2, ..., x_n$ disposte in ordine non decrescente (o non crescente) , cioè tali che $x_1 \le x_2 \le .... \le x_n$ $(x_1 \ge x_2 \ge .... \ge x_n)$ definiamo **mediana** M_e il valore x_k che divide la graduatoria in due parti tali che il numero dei termini che la precedono sia uguale al numero dei termini che la seguono. Se il numero dei dati disponibili è **dispari** la mediana M_e è rappresentata dal termine centrale, cioè quello che occupa il posto $\frac{n+1}{2}$ della successione esaminata. Se n è **pari** non esiste un termine mediano, bensì una coppia di valori mediani. In questo caso si ha un **intervallo mediano** o **zona mediana**, mentre la mediana risulta indeterminata, potendosi assumere come tale un qualsiasi valore dell'intervallo mediano. Nella pratica, tuttavia, è consuetudine adottare come mediana la semisomma dei due termini centrali che occupano rispettivamente i posti $\frac{n}{2}$ ed $\frac{n}{2}+1$.

Esempio: **calcolo della mediana; dati non raggruppati; numero dispari di osservazioni**

Le intensità della variabile statistica (ad esempio i voti riportati all'esame di statistica) siano 21, 25, 27, 28, 30. Poiché il numero delle unità osservate (studenti) è $n = 5$, numero dispari, l'unico posto centrale è dato da $P = \frac{5+1}{2} = 3$. Si tratta del terzo posto centrale della sequenza crescente delle intensità. Ad esso corrisponde l'intensità $M_e = 27$ che è la **mediana** della successione. In effetti, l'intensità mediana $M_e = 27$ separa due gruppi ugualmente numerosi, ciascuno costituito da due osservazioni.

Esempio: **calcolo della mediana; dati non raggruppati; numero pari di osservazioni**

Le intensità della variabile statistica, ad esempio i voti riportati all'esame di Economia Politica, siano 18, 19, 20, 22, 23, 26, 28, 30. Poiché il numero delle unità osservate (studenti) è $n = 8$, numero pari, i due posti centrali sono $P_1=\frac{8}{2}=4$ (quarto posto) e $P_2=\frac{8}{2}+1=4+1=5$ (quinto

posto). Le intensità ad essi corrispondenti sono, rispettivamente, 22 e 23. Può quindi assumersi come **mediana** un qualsiasi valore dell'intervallo [22,23] e in particolare il suo valore centrale:

$$M_e = \frac{22+23}{2} = 22,5$$

Detto valore isola e destra e a sinistra due insiemi ugualmente numerosi, ciascuno costituito da quattro elementi (studenti).

Determinazione della mediana M_e per dati raggruppati e caratteri discontinui

La determinazione della mediana presenta qualche difficoltà quando i termini non sono indicati singolarmente. In questo caso, nota la tabella dei dati raggruppati si costruisce la tabella delle frequenze cumulate. Se $\sum f_i$ è **dispari** si ha un solo posto centrale, se N è **pari** si hanno due posti centrali i quali molto spesso coincidono, in tal caso la **mediana è il loro valore comune** .

Se non coincidono , si può assumere come mediana una qualunque intensità dell'intervallo mediano da essi definito o, più semplicemente , la semisomma dei due valori centrali. Sia data la seguente distribuzione di frequenza di una variabile statistica discreta:

x_i	3	4	5	6	7	8
f_i	11	17	38	30	45	59

Per potere stabilire qual è il dato centrale di questa distribuzione si procede come segue:

a) Si controlla che i dati x_i siano tutti disposti in ordine crescente

b) Si sommano tutte le frequenze assolute f_i e si divide il risultato per due

$$\frac{11 + 17 + 38 + 45 + 50}{2} = 95,5$$

Individuando, così, la posizione centrale (o le posizioni centrali se i dati sono in numero pari). Nel nostro caso la **posizione centrale** è la 96-esima.

c) per trovare quale dato corrisponde alla posizione centrale si calcolano le frequenze cumulate finché non si arriva ad una frequenza cumulata uguale o maggiore della posizione centrale.

x_i	n_i	frequenze cumulate
3	11	11
4	17	28
5	38	66

6	30	96
7	45	
8	50	
	Totale $N=191$	

La **mediana** è **6** perché corrisponde alla frequenza cumulata 96.
La determinazione della mediana si presenta ancora più complicata quando ci si trova in presenza di una distribuzione per classi.
Nella pratica la **serie delle frequenze cumulate**, analogamente a quanto si è visto nel caso di una variabile statistica per singoli valori , ci consente di accertare che la mediana si trova all'interno della prima classe per la quale la frequenza cumulata uguaglia o supera il numero $\frac{N}{2}=\frac{n_1+n_2+....+n_n}{2}$, ossia metà delle osservazioni **N**. Come mediana possiamo scegliere, approssimativamente, il valore centrale della classe mediana.
<<**Pellicole cinematografiche programmate in Italia nel 1976 per classi d'incasso (in milioni di lire**)>>

classi d'incasso x_i	Pellicole f_i	frequenze cumulate C_i
0—2	22	22
2—5	20	42
5—10	18	60
10—20	35	95
20—50	47	142
50—100	52	194
100—200	48	242
200—ω	81	323
Totale N	323	

La **frequenza cumulata** del posto centrale è $C=P=\frac{323+1}{2}=162$ .La pellicola che fa registrare l'**incasso mediano** occupa il posto 162. Essa si trova nella sesta frequenza cumulata ($C_6=194$) , er cui la **classe mediana** è quella che va da 50 a 100 milioni (50—100).

Come valore mediano possiamo scegliere il valore centrale, cioè: $M_e=\frac{50+100}{2}=75$ milioni

La moda

Si dice moda o **valore modale** di una distribuzione di frequenza la modalità o l'**intensità** del carattere, se esiste, cui corrisponde la massima frequenza nella distribuzione.

La **moda** può non esistere (quando tutti i valori hanno la stessa frequenza) e se esiste può non essere unica. Se esiste ed è unica si parla di **distribuzione unimodale**, se invece non è unica la distribuzione è detta **plurimodale**. E' possibile , quindi , imbattersi in **distribuzioni zeromodali** (cioè prive di moda), **unimodali** (con una sola moda), bimodali (con due valori modali) **plurimodali**.

Se un collettivo è distribuito secondo modalità (non raggruppate in classi) di un carattere discreto, l'identificazione della moda è immediata: basta scorrere la colonna delle frequenze e individuare la modalità che presenta la **massima frequenza**.

Se il carattere è presentato in classi, occorre distinguere due situazioni:

1) le classi hanno tutte la stessa ampiezza: la moda cade nella classe di maggiore frequenza

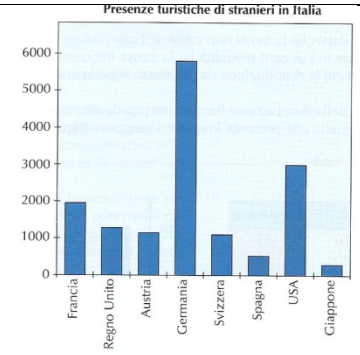
2) le classi hanno ampiezze diverse: la **classe modale** coincide con la classe avente la maggiore **densità di frequenza**, intendendo per **densità di frequenza** di una classe il rapporto fra la frequenza della classe e l'ampiezza della classe stessa.

Quindi nel caso di distribuzioni di frequenza con classi aventi ampiezze diverse, In prima approssimazione possiamo identificare la **moda** con la posizione centrale (semisomma degli estremi) della classe modale.

Esempi

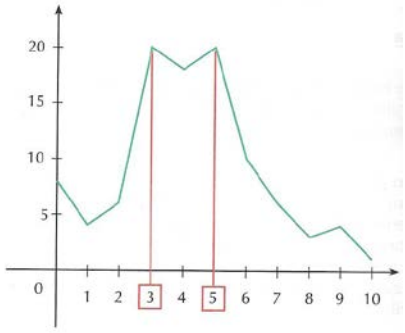
Consideriamo la distribuzione di frequenza che ci propone i dati relativi alle presenze dei turisti stranieri in Italia, espressi in migliaia e riferito all'anno 1994. La modalità che ricorre con frequenza maggiore è quella relativa alla Germania; la modo della distribuzione considerata è la Germania.

Paese	Presenze (in migliaia)
Francia	1950
Regno Unito	1280
Austria	1150
Germania	5800
Svizzera	1100
Spagna	530
USA	3000
Giappone	290



Un'azienda produttrice di elettrodomestici effettua un'indagine sull'affidabilità dei propri prodotti. Su un campione di 100 elettrodomestici. Il numero di riparazioni che questi hanno subito è indicato nella tabella.

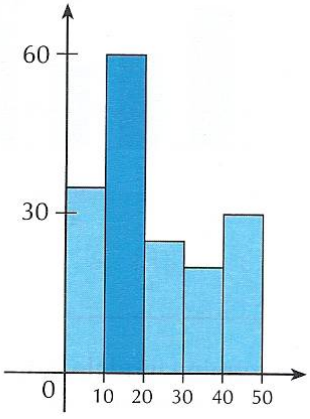
Numero di riparazioni	Frequenza
0	8
1	4
2	6
3	20
4	18
5	20
6	10
7	6
8	3
9	4
10	1
TOTALE	100



Si nota che i valori che presentano la maggiore frequenza sono due, il 3 ed il 5. Questa distribuzione ha due valori modali e per questo motivo è detta bimodale.

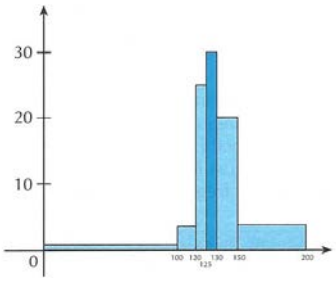
Se le classi della distribuzione hanno tutte uguale ampiezza, la classe modale è quella che presenta la frequenza maggiore.

Classi	Frequenza
0–10	35
10–20	60
20–30	25
30–40	20
40–50	30
TOTALE	170



Se le classi hanno ampiezze diverse, la classe modale è quella che presenta la maggiore densità di frequenza. Approssimativamente la moda coincide con il valore centrale della classe modale.

Classi	Frequenza	Frequenza/ Ampiezza
0–100	80	0,8
100–120	75	3,75
120–125	130	26
125–130	150	30
130–150	400	20
150–200	200	4
TOTALE	1035	



La seguente tabella rappresenta il numero di dipendenti di un'azienda divisi per classi di età. Determinare: 1) l'istogramma 2) l'areogramma 3) la media aritmetica 4) la moda 5) la mediana

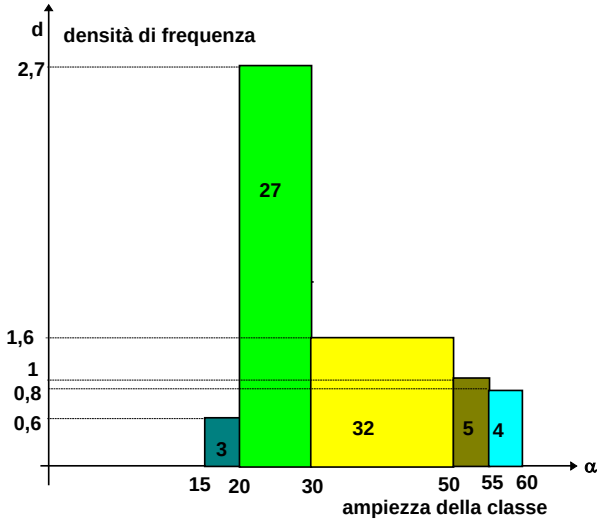
Classi di età	Frequenze f_k	Densità di frequenza $d_k = \frac{f_k}{\alpha_k}$	Modalità del carattere x_k	$f_k \cdot x_k$	Frequenze cumulate
15---20	3	0,6	17,5	52,5	3
20---30	27	2,7	25,5	688,5	30
30---50	32	1,6	40	1280	62
50---55	5	1	52,5	262,5	67
55---60	4	0,8	57,5	230	71
	N = 71			2513,5	

Per disegnare l'istogramma della distribuzione di frequenza basta riportare sull'asse delle ascisse l'ampiezza di ciascuna classe e sull'asse delle ordinate la densità di frequenza di ciascuna classe. L'area di ogni rettangolo ci dà la frequenza di ciascuna classe.

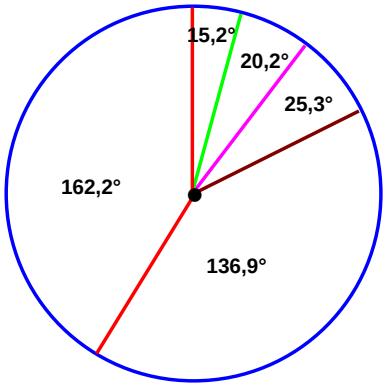
$\alpha_1 = 20-15 = 5$; $\alpha_2 = 30-20 = 10$;

$\alpha_3 = 50-30 = 20$; $\alpha_4 = 55-50 = 5$

$\alpha_5 = 60-55 = 5$



Per disegnare l'areogramma debbo calcolare gli angoli al centro dei 5 settori circolari, in quanto le modalità del carattere sono 5. Utilizzo le seguenti proporzioni:



$3:71 = x:360 \quad x = \frac{3}{71}\cdot 360° = 15,2°$

$27:71 = y:360 \quad y = \frac{27}{71}\cdot 360° = 136,9°$

$32:71 = z:360 \quad z = \frac{32}{71}\cdot 360° = 162,2°$ $\quad 5:71 = t:360 \quad t = \frac{5}{71}\cdot 360° = 25,3°$

$4:71=v:360 \quad v=\frac{4}{71}\cdot 360°=20,2°$

$m=\frac{2513,5}{71}=35,4$ media aritmetica la densità di frequenza massima è $d_{max}=2,7$ la classe modale è: 20 —— 30 ; come moda possiamo prendere, approssimativamente il suo valore centrale $\frac{20+30}{2}=25$ Questo ci consente di affermare che **25 anni è l'età che si presenta con una maggiore frequenza**.

La mediana si trova all'interno della prima classe per la quale la frequenza cumulata uguaglia o supera il numero $\frac{N}{2}=\frac{71}{2}=35,5$, ossia metà delle osservazioni **N**. La prima frequenza cumulata che supera il numero 35,5 è 62 ; ad essa corrisponde la classe mediana 30 —— 50 . Come mediana possiamo scegliere, approssimativamente, il valore centrale della classe mediana, cioè: 40 anni.

Calcolo delle probabilità

La nozione di evento casuale **o** aleatorio è assunta come **primitiva** ed è sinonimo di <<**avvenimento il cui verificarsi dipende dal caso**>>. **Probabilità** è un numero associato al presentarsi di un evento aleatorio e denota l'attendibilità razionale che ha l'evento stesso di verificarsi. Consideriamo un esperimento (o **schema probabilistico**) **E** (ad esempio il lancio di una moneta, il lancio di due dadi, l'estrazione di una pallina da un'urna,...).Col termine prova intendiamo una singola esecuzione di un determinato esperimento. Da questa prova si ottiene un singolo risultato elementare detto evento aleatorio elementare. All'evento associamo, secondo regole da fissare, un numero che esprime la **probabilità** che si verifichi l'evento aleatorio.

Quindi con l'espressione <<**probabilità dell'evento A**>> intendiamo riferirci ad un particolare numero che meglio di altri è in grado di sintetizzare la fiducia che noi riponiamo nella sua realizzazione.

Precisiamo con alcune definizioni quanto finora detto.

- si dice esperimento aleatorio un avvenimento il cui esito non è certo
- si dice evento aleatorio uno dei possibili esisti di un esperimento aleatorio
- si dice probabilità di un evento il numero che esprime una stima approssimata della possibilità che esso si verifichi

Sono esperimenti aleatori il lancio di un dado, l'estrazione di un numero in una lotteria, l'estrazione di una carta da gioco da un mazzo. Sono eventi aleatori:

- nel lancio di una moneta <<esce testa>>, oppure <<esce croce>>.
- nell'estrazione di una carta da gioco, <<esce il re>> oppure <<esce una carta di fiori>>.

Consideriamo l'esperimento casuale del lancio di un dado. Questo esperimento fornisce il seguente **insieme**: $\Omega = \{1,2,3,4,5,6\}$ che evidenzia sei eventi elementari $e_1 = \{1\}$, $e_2 = \{2\}$, $e_3 = \{3\}$, $e_4 = \{4\}$, $e_5 = \{5\}$, $e_6 = \{6\}$ e diversi **eventi complessi**, come l'evento $A = \{2,4,6\}$ = **comparsa di un numero pari** o l'evento $B = \{1,3,5\}$ = **comparsa di un numero dispari**

Definizione classica di probabilità

La **definizione classica di probabilità** enunciata da Laplace afferma quanto segue: **la probabilità $p(E)$ di un evento aleatorio E coincide col rapporto tra il numero f dei casi favorevoli all'evento E ed il numero n dei casi possibili nell'ipotesi che essi siano tutti ugualmente possibili**. In formule abbiamo: $p(E) = \frac{f}{n}$ con $f \le n$ $0 \le p(E) \le 1$

$f = 0 \Rightarrow p(A) = 0 \Rightarrow$ l'evento **A** è impossibile, cioè non può verificarsi mai

$f = n \Rightarrow p(A) = 1 \Rightarrow$ l'evento A è certo, $n = 2f \Rightarrow p(A) = \frac{1}{2} \Rightarrow$ l'evento A si dice equiprobabile $p(A) < \frac{1}{2}$ l'evento A è detto improbabile , $p(A) > \frac{1}{2}$ l'evento A è detto probabile. La definizione classica di probabilità è applicabile quando siamo in grado di definire il numero dei casi possibili ed equiprobabili.

Esempi

Si consideri il lancio simultaneo di due dadi e si determini la probabilità dell'evento complesso "la somma dei punteggi delle due facce è 5"

La tabella indica tutti i casi possibili che sono 36. Ognuno di questi casi rappresenta un evento elementare. L'evento proposto, "la somma dei punteggi delle due facce è 5", è un evento complesso. I casi favorevoli, indicati in rosso, sono 4. $p(E)=\frac{4}{36}=\frac{1}{9}$

6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
	1	2	3	4	5	6

Da un'urna, contenente 5 palline numerate da 1 a 5, si estraggono 2 palline, senza reinserimento. Calcolare la probabilità dell'evento E = "uno dei numeri estratti è il 2".

Dalla seguente tabella deduciamo che i casi possibili sono 20 ed i casi favorevoli sono 8: $p(E)=\frac{8}{20}=\frac{2}{5}$

I casi $(1;1)$, $(2;2)$, $(3;3)$, $(4;4)$, $(5,5)$ non sono possibili in quanto l'estrazione avviene senza reinserimento.

	(1;2)	(1;3)	(1;4)	(1;5)
(2;1)		(2;3)	(2;4)	(2;5)
(3;1)	(3;2)		(3;4)	(3;5)
(4;1)	(4;2)	(4;3)		(4;5)
(5;1)	(5;2)	(5;3)	(5;4)	

Definizione: dato un evento E, si chiama evento contrario di E (si indica col simbolo $\overline{E}$ e si legge evento non E) l'evento che si verifica quando non si verifica l'evento E. L'evento contrario è detto anche evento opposto o evento complementare dell'evento E.

Teorema della probabilità dell'evento contrario

La somma della probabilità di un evento A e di quella dell'evento contrario $\overline{A}$ è uguale ad uno: $p(A) + p(\overline{A}) = 1$

<< **Un dado viene lanciato due volte. Determinare la probabilità che esca il numero 6 in almeno una delle due facce** >>

Gli eventi elementari associati allo schema probabilistico del lancio di due dati sono 36 e sono indicati in figura. Gli eventi elementari favorevoli all'evento E sono 11. Gli eventi elementari possibili sono 36.

6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
	1	2	3	4	5	6

A = esce il numero 6 in almeno una delle due facce $p(A) = \frac{11}{36}$

$\overline{A}$ = il numero 6 non esce in nessuna delle due facce $p(\overline{A}) = \frac{25}{36}$

Quindi: $p(A) = 1 - p(\overline{A}) = 1 - \frac{25}{36} = \frac{11}{36}$

Definizione: dati due eventi che dipendono da uno stesso fenomeno casuale (stesso schema probabilistico), diciamo che essi sono: • **compatibili** se possono verificarsi contemporaneamente (questo si verifica se i due eventi hanno in comune qualche evento elementare dello schema probabilistico) • **incompatibili** se non possono verificarsi contemporaneamente (questo si verifica se i due eventi non hanno in comune qualche evento elementare dello schema probabilistico). Concludendo possiamo affermare che **due eventi** appartenenti allo stesso schema probabilistico sono **incompatibili** se il verificarsi di uno di essi esclude il verificarsi dell'altro. In caso contrario si dicono **compatibili**.

Come **esempio** consideriamo lo schema probabilistico del lancio di un dado. Gli eventi elementari sono i numeri 1, 2, 3, 4, 5, 6. Prendiamo in esame i seguenti eventi complessi:

E_1 = esce un numero dispari E_2 = esce un multiplo di 3 E_3 = esce il numero 4.

Gli **eventi** E_1 ed E_2 sono **compatibili**, in quanto se esce l'evento elementare 3, si verificano entrambi. Gli **eventi** E_1 ed E_3 sono **incompatibili** perché nessuno dei 6 eventi elementari sono contemporaneamente eventi favorevoli agli eventi complessi E_1 ed E_3.

Due **eventi** sono **indipendenti** se il verificarsi di uno di essi non altera la probabilità di verificarsi dell'altro. Quindi l'evento A è **indipendente** dall'evento B se il verificarsi dell'evento B non modifica la probabilità di verificarsi dell'evento A e viceversa. L'estrazione di successive palline da un'urna, dopo che si sia reimmessa nell'urna la precedente, è un modello di studio per **eventi indipendenti**. Due **eventi** sono **dipendenti** se il verificarsi di uno di essi modifica la probabilità di verificarsi dell'altro. L'estrazione di successive palline da un'urna, senza reimmissione nell'urna della pallina estratta, è un modello di studio per **eventi dipendenti**.

Adesso vogliamo considerare **eventi complessi** che hanno una particolare importanza nel calcolo delle probabilità. Abbiamo chiamato **complesso** un qualsiasi **evento** che risulti

combinazioni di altre eventi più semplici, in particolare di eventi elementari nello schema probabilistico considerato.

Definizione: dati gli eventi E_1 ed E_2, si chiama evento composto (o evento prodotto o evento intersezione) l'evento E che si realizza quando si verificano contemporaneamente i due eventi E_1 ed E_2. Per indicare che E è l'evento composto dei due eventi E_1 ed E_2 scriviamo: $E = E_1 \text{ e } E_2$

Col simbolo $p(A/B)$ indichiamo la **probabilità dell'evento A quando l'evento B si è verificato**. Risulta: $p(A/B) = \frac{p(A \text{ e } B)}{p(B)} = \frac{n_{A \cap B}}{n_B}$ dove $n_{A \cap B}$ rappresenta il numero di casi favorevoli all'evento prodotto $A \text{ e } B$ ed n_B il numero di casi favorevoli all'evento B.

Analogamente, la probabilità dell'evento **B** rispetto quando l'evento **A** si è verificato è definita dalla formula (si noti che $A \text{ e } B = B \text{ e } A$)

$$p(B/A) = \frac{p(A \text{ e } B)}{p(A)} = \frac{n_{A \cap B}}{n_A}$$

Dove n_A in dica il numero di casi favorevoli all'evento A.

Esempio: << **Si lancino contemporaneamente due dadi. Qual è la probabilità che la somma dei numeri delle due facce valga 8 (evento A) sapendo che il risultato ottenuto è un numero pari (evento B)?**>>

Evento **A** = la somma dei numeri presenti nelle facce dei due dadi vale 8

Evento **B** = la somma dei numeri presenti nelle facce dei due dadi è un numero pari

6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
	1	2	3	4	5	6

6	7	8	9	10	11	12
5	6	7	8	9	10	11
4	5	6	7	8	9	10
3	4	5	6	7	8	9
2	3	4	5	6	7	8
1	2	3	4	5	6	7
	1	2	3	4	5	6

Noi sappiamo che nel lancio di due dadi esistono 36 risultati possibili (36 eventi elementari). I casi favorevoli all'evento A sono 5: $A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$ Quindi la **probabilità incondizionata** (cioè senza la conoscenza dell'evento **B**) dell'evento **A** è: $p(A) = \frac{5}{36}$

Se si è verificato l'evento **B** i casi possibili sono 18 e non 36 mentre i casi favorevoli sono ancora 5 e, di conseguenza, la probabilità dell'evento A condizionata dall'evento B vale: $p(A/B) = \frac{5}{18}$

$$p(A/B) = \frac{p(A\,e\,B)}{p(B)} = \frac{\frac{5}{36}}{\frac{18}{36}} = \frac{5}{18}$$

Il teorema della probabilità composta per eventi compatibili ed indipendenti

<< **La probabilità di verificarsi di un evento C composto di due eventi A e B compatibili e indipendenti è uguale al prodotto delle probabilità dei singoli eventi** >>

$$p(C) = p(A\,e\,B) = p(A)\cdot p(B) \quad [*]$$

Il teorema della probabilità composta per eventi compatibili e dipendenti

<< **La probabilità di verificarsi di un evento composto C , formato da due eventi A e B compatibili e dipendenti , è data dal prodotto della probabilità che ha il primo evento di verificarsi per la probabilità che ha il secondo nell'ipotesi che il primo si sia verificato**. >>

$$p(C) = p(A\,e\,B) = p(A)\cdot p(B/A) = p(B)\cdot p(A/B)$$

Esempio: **Consideriamo l'estrazione senza reinserimento di due palline da un'urna contenente una pallina verde, due rosse e due blu, come indicato in figura**. **Calcolare la probabilità che la prima pallina estratta sia verde e la seconda blu.**

L'evento del quale dobbiamo calcolare la probabilità è l'evento E = la prima pallina estratta è verde e la seconda è blu. L'evento E lo possiamo immaginare come l'evento composto dei due seguenti eventi elementari: E_1 = la prima pallina estratta è verde E_2 = la seconda pallina estratta è blu. Si tratta di due eventi dipendenti.

$$p(E_1)=\frac{1}{5} \qquad p(E_2/E_1)=\frac{2}{4}=\frac{1}{2} \qquad p(E)=p(E_1\,e\,E_2)=p(E_1)\cdot p(E_2/E_1)=\frac{1}{5}\cdot\frac{1}{2}=\frac{1}{10}$$

Per calcolare la probabilità precedente senza usare la formula della probabilità composta, bisogna considerare le due estrazioni come un fenomeno unico e fare il seguente schema, dal quale deduciamo che i casi possibili sono 20, mentre i casi favorevoli sono 2, precisamente $(5;3)$, $(5;4)$

. $$p(E_1\,e\,E_2) = \frac{p(E_1\,e\,E_2)}{p(E_2)} = \frac{\frac{5}{36}}{\frac{18}{36}} = \frac{2}{20}$$??????

$(1;2)$ $(1;3)$ $(1;4)$ $(1;5)$
$(2;1)$ $(2;3)$ $(2;4)$ $(2;5)$
$(3;1)$ $(3;2)$ $(3;4)$ $(3;5)$
$(4;1)$ $(4;2)$ $(4;3)$ $(4;5)$
$(5;1)$ $(5;2)$ $(5;3)$ $(5;4)$

Esempio: **Consideriamo l'estrazione con reinserimento di due palline da un'urna contenente una pallina verde, due rosse e due blu, come indicato in figura. Calcolare la probabilità che la prima pallina estratta sia verde e la seconda blu.**
Questa volta gli eventi E_1 ed E_2 sono indipendenti, perché l'estrazione avviene con reinserimento.

$$p(E_1)=\frac{1}{5} \quad p(E_2/E_1)=\frac{2}{5} \quad p(E)=p(E_1 \text{ e } E_2)=p(E_1)\cdot p(E_2/E_1)=\frac{1}{5}\cdot\frac{2}{5}=\frac{2}{25}$$

Per calcolare la probabilità precedente senza usare la formula della probabilità composta, bisogna considerare le due estrazioni come un fenomeno unico e fare il seguente schema, dal quale deduciamo che i casi possibili sono 20, mentre i casi favorevoli sono 2, precisamente $(5;3)$, $(5;4)$.

$(1;1)$ $(1;2)$ $(1;3)$ $(1;4)$ $(1;5)$
$(2;1)$ $(2;2)$ $(2;3)$ $(2;4)$ $(2;5)$
$(3;1)$ $(3;2)$ $(3;3)$ $(3;4)$ $(3;5)$
$(4;1)$ $(4;2)$ $(4;3)$ $(4;4)$ $(4;5)$
$(5;1)$ $(5;2)$ $(5;3)$ $(5;4)$ $(5;5)$

Definizione: dati gli eventi E_1 ed E_2, si chiama evento totale (o evento somma o evento unione) l'evento E che si realizza quando si verificano almeno uno dei due eventi E_1 ed E_2, cioè quando si verifica E_1 oppure E_2. Per indicare che E è l'evento totale dei due eventi E_1 ed E_2 scriviamo: $E=E_1 \text{ o } E_2$.

Teorema della probabilità totale per eventi incompatibili

Dati due eventi **A** e **B** **incompatibili** la probabilità dell'evento somma $A \text{ o } B$ è data dalla somma delle probabilità di ciascun evento. $p(A \text{ o } B)=p(A)+p(B)$ $p(A \text{ o } B \text{ o } C)=p(A)+p(B)+p(C)$

Teorema della probabilità totale per eventi compatibili

Dati due eventi **A** e **B** **compatibili** la probabilità dell'evento somma $A \text{ o } B$ è data dalla somma delle probabilità dei singoli eventi diminuita della probabilità dell'evento prodotto:

$$p(A \text{ o } B) = p(A) + p(B) - p(A \text{ e } B)$$

Esempio: Un'urna contiene 15 palline numerate da 1 a 15. Calcolare la probabilità che estraendo una pallina essa rechi: 1) un numero dispari o maggiore di 10 2) un numero minore di 6 o maggiore di 10.

Consideriamo i seguenti eventi: A = esce un numero dispari B = esce un numero maggiore di 10

C = esce un numero minore di 6

Gli eventi dei quali calcolare la probabilità sono: E = A o B D = C o B

Gli eventi ***A e B sono compatibili e dipendenti***, mentre gli eventi **C e D sono compatibili e indipendenti** $p(E)=p(A\,o\,B)=p(A)+p(B)-p(A\,e\,B)=\frac{8}{15}+\frac{5}{15}-\frac{3}{15}=\frac{10}{15}=\frac{2}{3}$

$$p(D)=p(C\,o\,B)=p(C)+p(B)=\frac{5}{15}+\frac{5}{15}=\frac{10}{15}=\frac{2}{3}$$

Definizione frequentista di probabilità

Volendo utilizzare il calcolo delle probabilità nello studio della realtà che ci circonda, l'uso della **probabilità classica** si rivela assai limitato. Ad esempio, la definizione classica di probabilità non ci consente di stabilire quale probabilità ha una persona di 30 anni di essere in vita tra 10 anni. Occorre introdurre altri metodi che ci consentano di determinare la probabilità di più vaste classi di eventi aleatori. La definizione classica di probabilità è applicabile soltanto quanto siamo in grado di definire il numero dei casi possibili ed equiprobabili. Quando questo non è possibile può essere utile fare ricorso alla teoria frequentista di probabilità di un evento. Concetto base per i frequentisti è quello della **frequenza relativa di un evento** intesa come il rapporto fra il numero h di volte in cui l'evento **E** si è verificato ed il numero n delle prove effettuate: $f(E)=\frac{h}{n}$

Evidentemente la frequenza di un certo evento **E** varia al variare delle prove e, addirittura, pur mantenendo fisso il numero delle prove sullo stesso evento **E** e nelle stesse condizioni le frequenze di ogni singolo insieme di prove potranno risultare tra loro diverse. Tuttavia l'esperienza dimostra che, se il numero **n** di prove è abbastanza grande, i valori delle frequenze differiscono di poco l'uno dall'altro. Da questa osservazione si evidenzia che il rapporto $\frac{h}{n}$ assume un valore tendenzialmente costante quanto più grande è n.

LEGGE EMPIRICA DEL CASO o LEGGE DEI GRANDI NUMERI

In una serie di prove, ripetute un gran numero di volte e tutte nelle stesse condizioni, un evento casuale **E** si verifica con una frequenza **f** che varia di poco al variare del numero delle prove e le variazioni, in generale, sono tanto più piccole quanto più grande è il numero delle prove ripetute.

Se n è grande allora: $p(E) \approx f(E)$

Questa probabilità è chiamata **probabilità empirica** di un evento o **probabilità a posteriori** o ***probabilità statistica*** in contrapposizione a quella classica detta **probabilità teorica** o ***probabilità a priori*** o **probabilità matematica**.

In sintesi possiamo affermare quanto segue.

In generale risulta $p(\mathrm{E}) \neq f(\mathrm{E})$, anzi il valore di $f(\mathrm{E})$ dipende dal numero di prove effettuate ed, a parità di prove effettuate, possiamo trovare valori diversi di $f(\mathrm{E})$.

Tuttavia, quando il numero delle prove effettuate è abbastanza grande (teoricamente infinito), il valore di $f(\mathrm{E})$ tende a stabilizzarsi attorno ad un valore ben preciso che si discosta poco dalla probabilità matematica **p(E)** dell'evento **E**.

Osservazioni ripetute hanno portato alla formulazione della seguente legge che, traendo origine dall'esperienza, non è dimostrabile ed è detta, per questo motivo, **legge empirica del caso** o **legge dei grandi numeri**: << su un numero molto grande di prove, effettuate tutte nelle medesime condizioni, la **frequenza** $f(\mathrm{E})$ con la quale si presenta un certo evento E assume generalmente valori molto prossimi a quello della probabilità **p(E)** dello stesso evento e tale approssimazione è tanto migliore quanto più elevato è il numero delle prove effettuate>> Questa definizione di probabilità è dovuta a Richard von Mises (1919).

Definizione di probabilità secondo la teoria soggettivista

Se vogliamo conoscere la probabilità che una squadra di calcio di serie A vinca il campionato in corso non possiamo applicare né la probabilità classica né quella frequentista. Potremmo introdurre la seguente nuova definizione di probabilità: la probabilità p che la squadra vinca il campionato in corso (evento E) è uguale al rapporto $\frac{P}{S}$ che uno scommettitore coerente ritiene equo pagare la somma **P** per riscuotere la somma **S** nel caso che la squadra di calcio vinca il campionato. La probabilità soggettivista di un evento rappresenta il grado di fiducia che un individuo coerente attribuisce al presentarsi di un evento. $p(E) = \frac{P}{S}$

Un individuo si considera coerente nella propria valutazione se è disposto ad accettare indifferentemente il ruolo di **scommettitore** o quello di **controparte**.

La **probabilità in senso soggettivo** di un evento **E** è il rapporto $p(E)$ fra il prezzo P che un individuo coerente è disposto a pagare e la somma S che ha diritto a riscuotere se l'evento E si verifica, perdendo invece la somma P se l'evento non si verifica. In simboli abbiamo: $p(E)=\frac{P}{S}$

Anche in questo caso abbiamo $0\le p(E)\le 1$. Infatti è logico supporre che, se un individuo stima che un evento non ha alcuna possibilità di realizzarsi, ritenga giusto pagare zero euro indipendentemente dalla somma che riceverebbe in cambio. In questo caso avremmo: $p(E)=\frac{0}{S}=0$. Per un evento che uno scommettitore ritiene certo, la controparte non gli potrà dare più della somma che il giocatore intende mettere a disposizione. In questo caso abbiamo: $p(E)=\frac{P}{P}=1$

Osservazione: La teoria soggettivista si sviluppa negli anni venti per opera del filosofo inglese Ramsey (1903-1930) ma solo con De Finetti la **concezione soggettivista della probabilità** diviene una vera e propria teoria matematica.

Gioco equo e speranza matematica

In un gioco d'azzardo il prodotto della vincita per la probabilità di conseguirla dicesi speranza matematica. Se un giocatore ha la probabilità $p(E)$ di vincere la somma V, allora la sua speranza matematica vale: $S=p(E)\cdot V$.

Un gioco si dice equo quando la **speranza matematica del giocatore è uguale alla speranza matematica della controparte**. In simboli abbiamo $p(E)\cdot V=p(\bar{E})\cdot R$ con $p(E)=\frac{R}{V}$

Se si verifica l'evento E il giocatore vince la somma V, se si verifica l'evento contrario $\bar{E}$ la controparte vince la somma R. $p(E)\cdot V$ = speranza matematica del giocatore

$p(\bar{E})\cdot R$ = speranza matematica della controparte