Statistica descrittiva (ESEMPI)

UNA DEFINIZIONE GENERALE ED INFORMALE DI STATISTICA È UN INSIEME DI STRUMENTI CHE PERMETTONO DI ELABORARE I METODI PIÙ IDONEI PER ANALIZZARE FENOMENI COME QUELLI ECONOMICI PER SCOPI DESCRITTIVI/PREVISIONALI/DI CONFRONTO.
NATURALMENTE PER FARE QUESTO C'È BISOGNO DI INFORMAZIONI SUI FENOMENI STESSI.
SE QUESTE INFORMAZIONI SONO RACCOLTE MEDIANTE DETERMINATE REGOLE SI PARLA DI RIVELAZIONI STATISTICHE.
PER DESCRIVERE UN FENOMENO IN MANIERA PIÙ VICINA ALLA REALTÀ, LE RILEVAZIONI STATISTICHE, CHE SIANO SULLA POPOLAZIONE (AD ESEMPIO IL CENSIMENTO) O SU UN CAMPIONE (AD ESEMPIO UN SONDAGGIO), DANNO LUOGO A GROSSI QUANTITATIVI DI DATI.
DI CONSEGUENZA AL FINE DI DESCRIVERE AL MEGLIO UN FENOMENO PER POTERLO PREVEDERE O CONFRONTARE È NECESSARIA UNA BUONA ORGANIZZAZIONE DEI DATI.
PER FARE CIÒ LA STATISTICA METTE A DISPOSIZIONE GLI STRUMENTI CHE PERMETTONO DI ORGANIZZARE E DESCRIVERE GRANDI MOLI DI DATI MEDIANTE UNA LORO OPPORTUNA SINTESI.
QUANDO I DATI SONO QUANTITATIVI, CIOÈ NUMERABILI, TALI STRUMENTI SONO I COSIDDETTI INDICATORI STATISTICI:

1- INDICI DI POSIZIONE, PER DARE UNA POSIZIONE AI DATI LOCALIZZANDONE IL CENTRO.
2- INDICI DI VARIABILITÀ, PER MISURARE LA DISPERSIONE DEI DATI RISPETTO AL LORO CENTRO.
3- INDICI SULLA FORMA, PER DARE UN'IDEA SULLA SIMMETRIA O ASIMMETRIA DI COME I DATI SI DISTRIBUISCONO.

Statistica descrittiva (ESEMPI)

LA STATISTICA DESCRITTIVA È QUELLA DISCIPLINA CHE PERMETTE DI ELABORARE E SINTETIZZARE GRANDI MOLI DI DATI RAPPRESENTANDOLI MEDIANTE TABELLE E GRAFICI, E RACCOGLIENDO LE INFORMAZIONI MEDIANTE SONDAGGI CON LA SOMMINISTRAZIONE DI QUESTIONARI.
SUPPONIAMO DI CONDURRE UN'INDAGINE PER CONOSCERE GLI ISCRITTI AL PRIMO ANNO DI UN CERTO CORSO DI LAUREA ALL'UNIVERSITÀ DI MILANO E DI ESSERE INTERESSATI ALLA LORO REGIONE DI ORIGINE.
UNA VOLTA RACCOLTI I DATI SI COSTRUISCE UNA TABELLA IN CUI SI DESCRIVONO LE REGIONI E PER OGNUNA SI INDICA IL NUMERO DI SOGGETTI, CHE PRENDE IL NOME DI FREQUENZE ASSOLUTE, CIOÈ IL NUMERO DI VOLTE CHE SI PRESENTA IL CORRISPONDENTE VALORE DELLA PRIMA COLONNA:

REGIONI	FREQUENZE ASSOLUTE
Piemonte	11
Campania	9
Sicilia	7
Calabria	6
Veneto	5
Abruzzo	3
Sardegna	3
Lombardia	2
Liguria	1
Lazio	1

TALE TABELLA SARÀ FORMATA DA 2 COLONNE, LA PRIMA RAPPRESENTA L'ELENCO DEI VALORI DELLA VARIABILE REGIONE MENTRE LA SECONDA INDICA I CORRISPONDENTI NUMERI DI VOLTE CHE OGNI

Statistica descrittiva (ESEMPI)

REGIONE SI È PRESENTATA NELL'INDAGINE.
OGNI NUMERO DELLA SECONDA COLONNA È DETTO PESO O FREQUENZA.
L'INSIEME DEI DATI COSÌ RAPPRESENTATO PRENDE IL NOME DI DISTRIBUZIONE DI FREQUENZE E SPESSO È UTILE RIPORTARE NELLA TABELLA ANCHE LE FREQUENZE RELATIVE (CIOÈ IL RAPPORTO TRA LA SINGOLA FREQUENZA ASSOLUTA ED IL TOTALE DELLE FREQUENZE DELLA DISTRIBUZIONE) E LE FREQUENZE PERCENTUALI (CIOÈ LE FREQUENZE RELATIVE MOLTIPLICATE PER 100), CIOÈ:

REGIONI	FREQUENZE ASSOLUTE	FREQUENZE RELATIVE	FREQUENZE %
Piemonte	11	0,23	23
Campania	9	0,19	19
Sicilia	7	0,15	15
Calabria	6	0,13	13
Veneto	5	0,10	10
Abruzzo	3	0,06	6
Sardegna	3	0,06	6
Lombardia	2	0,04	4
Liguria	1	0,02	2
Lazio	1	0,02	2
TOTALI	**48**	**1**	**100**

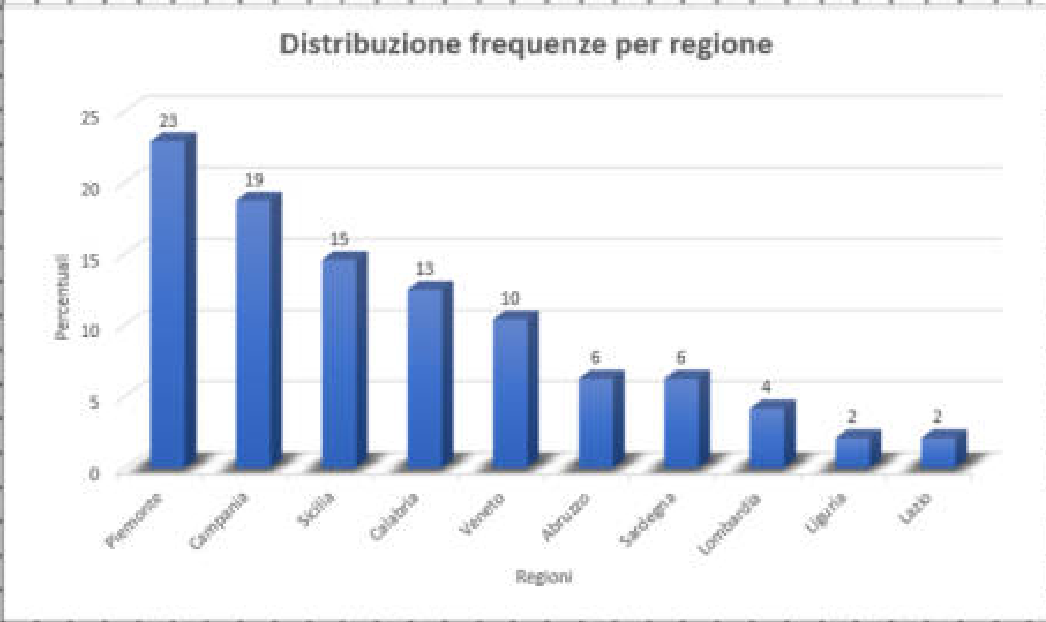
NATURALMENTE LE FREQUENZE RELATIVE VENGONO APPROSSIMATE, IN QUANTO NEL NOSTRO CASO AD ESEMPIO PER QUANTO RIGUARDA LA REGIONE SICILIA, ESSENDO IL TOTALE DELLE FREQUENZE ASSOLUTE PARI A 48, IL RAPPORTO

$$\frac{7}{48} = 0{,}145833333$$

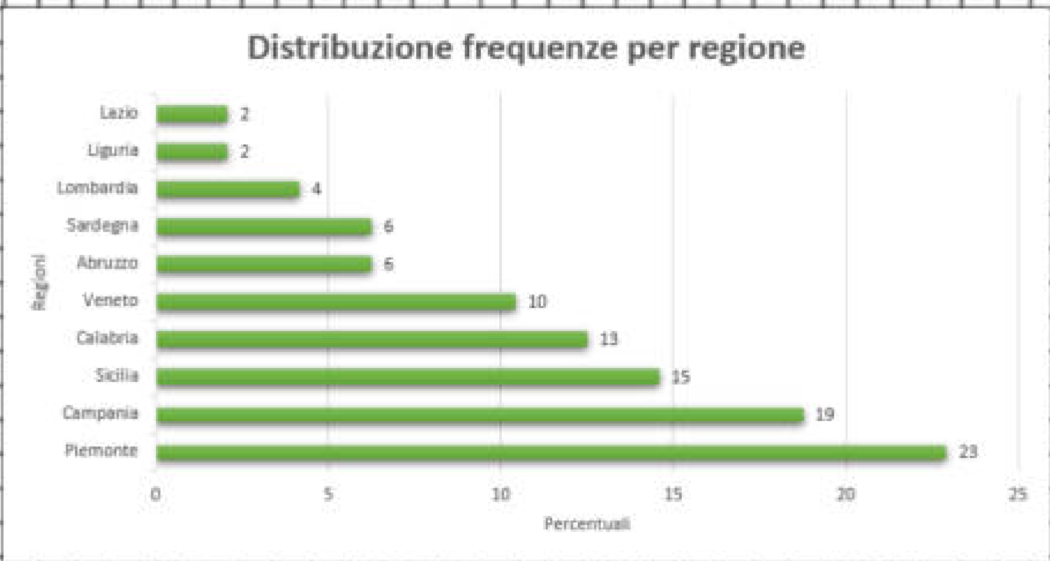
VIENE APPROSSIMATO A 0,15_

I DATI RIPORTATI NELLA TABELLA POSSONO ESSERE POI RAPPRESENTATI GRAFICAMENTE MEDIANTE ISTOGRAMMI O DIAGRAMMI A TORTA_

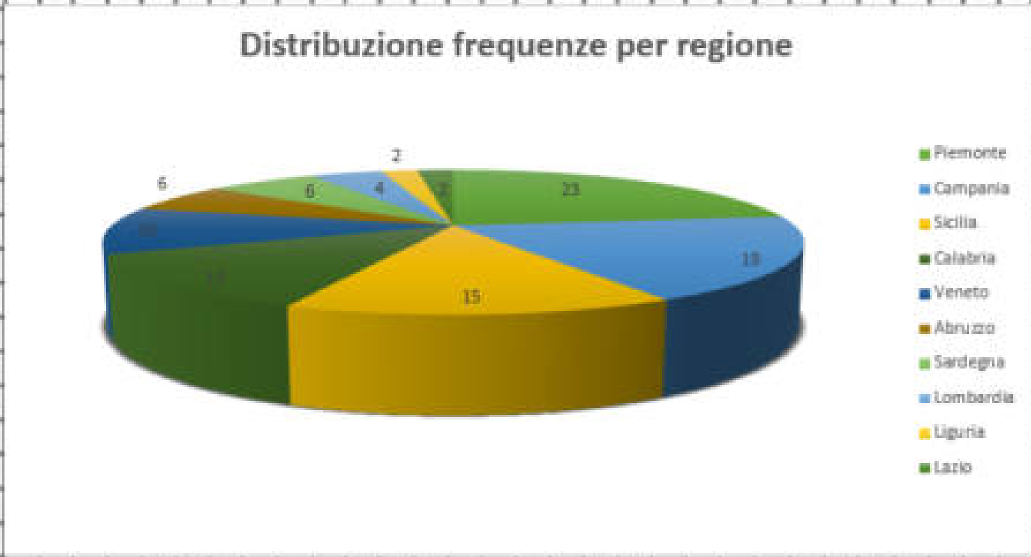
IL GRAFICO AD ISTOGRAMMA O DI PARETO PUÒ ESSERE IN VERTICALE:



O IN ORIZZONTALE:



MENTRE IL DIAGRAMMA A TORTA SARÀ:



LE VARIABILI STATISTICHE CON CUI SI RAPPRESENTANO I DATI POSSONO ESSERE DI DUE TIPI:

- QUALITATIVE, COME AD ESEMPIO NEL NOSTRO CASO LA REGIONE DI PROVENIENZA, OPPURE IL COLORE DEGLI OCCHI DI UN INDIVIDUO, IL TIPO DI VEICOLO CHE SI POSSIEDE, ECCETERA.
- QUANTITATIVE, CIOÈ MISURABILI ED ESPRIMIBILI NUMERICAMENTE, COME AD ESEMPIO NEL NOSTRO CASO IL NUMERO DI SOGGETTI DI QUELLA REGIONE, OPPURE L'ALTEZZA DI UN INDIVIDUO, IL REDDITO DI UNA FAMIGLIA, ECCETERA.

A LORO VOLTA LE VARIABILI QUANTITATIVE POSSONO ESSERE DISCRETE O CONTINUE.

UNA VARIABILE SI DICE DISCRETA SE PUÒ ASSUMERE UN NUMERO FINITO DI VALORI O UNA INFINITÀ NUMERABILE DI VALORI, AD ESEMPIO IL NUMERO DI STANZE DELLE ABITAZIONI, LE VOTAZIONI RIPORTATE AD UN ESAME O IL NUMERO DI VOLTE CHE ESCE TESTA NEI LANCI CONSECUTIVI DI UNA MONETA.

Statistica descrittiva (ESEMPI)

UNA VARIABILE SI DICE CONTINUA SE PUÓ ASSUMERE DEI VALORI IN UN QUALSIASI INTERVALLO DI NUMERI REALI, COME AD ESEMPIO L'ALTEZZA DELLE PERSONE, IL PESO DI UN OGGETTO, LA LUNGHEZZA ECCETERA.

PER LE VARIABILI QUANTITATIVE IL **CAMPO DI VARIAZIONE** È LA DIFFERENZA TRA IL VALORE MAGGIORE E QUELLO MINORE.

SPESSO LE VARIABILI QUANTITATIVE SI RAPPRESENTANO CON RAGGRUPPAMENTI IN CLASSI, COME AD ESEMPIO I DATI RELATIVI ALLE ALTEZZE IN CENTIMETRI DI UN CAMPIONE DI 50 STUDENTI DI UNA SCUOLA RIPORTATI NELLA TABELLA SEGUENTE, CON LE MISURAZIONI ARROTONDATE PER ECCESSO O PER DIFETTO AL VALORE INTERO:

Classi di altezza	Frequenze assolute	Valori centrali
151–155	4	153
156–160	9	158
161–165	15	163
166–170	7	168
171–175	8	173
176–180	3	178
181–185	3	183
186–190	1	188

IN CUI L'ALTEZZA 175,3 È ARROTONDATA A 175 E È CONTEGGIATA NELLA QUINTA CLASSE.

GLI ESTREMI DI OGNI CLASSE SONO DETTI **LIMITE INFERIORE E LIMITE SUPERIORE** DELLA CLASSE.

INDICI DI POSIZIONE

MEDIA ARITMETICA

DATI $x_1, x_2, \ldots, x_n$ n VALORI ASSUNTI DA UNA VARIABILE STATISTICA, SI DEFINISCE MEDIA ARITMETICA E SI INDICA CON $\bar{x}$, IL VALORE

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

PROPRIETÁ FONDAMENTALE DELLA MEDIA ARITMETICA È CHE SE AD OGNI x_i SI SOSTITUISCE IL SUO VALORE $\bar{x}$ LA SOMMA NON CAMBIA, INFATTI:

$$\frac{\bar{x} + \bar{x} + \cdots + \bar{x}}{n} = \frac{\cancel{n} \cdot \bar{x}}{\cancel{n}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

ESEMPIO

UNA FAMIGLIA HA PAGATO IN UN ANNO, I SEGUENTI IMPORTI IN EURO, RELATIVI RELATIVI AL CONSUMO DI ENERGIA ELETTRICA:

145,10 / 137,50 / 134,20 / 156,80 / 148,30 / 139,70

LA CIFRA BIMESTRALE PAGATA IN MEDIA È:

$$\frac{145{,}10 + 137{,}50 + 134{,}20 + 156{,}80 + 148{,}30 + 139{,}70}{6} = \frac{861{,}60}{6} = 143{,}60$$

PER LA PROPRIETÁ FONDAMENTALE SE LA FAMIGLIA AVESSE PAGATO OGNI BIMESTRE €143,60, ALLA FINE DELL'ANNO AVREBBE PAGATO LA STESSA SOMMA, INFATTI

$$143{,}60 \cdot 6 = €861{,}60$$

MEDIA ARITMETICA PONDERATA

NEL CASO IN CUI I DATI SONO RAGGRUPPATI SECONDO UNA DISTRIBUZIONE DI FREQUENZE, CIOÈ SIANO $x_1, x_2, \dots, x_n$ n VALORI ASSUNTI DA UNA VARIABILE STATISTICA, CON FREQUENZE ASSOLUTE $f_1, f_2, \dots, f_n$ RISPETTIVAMENTE, SI DEFINISCE MEDIA ARITMETICA PONDERATA IL VALORE:

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \dots + x_n \cdot f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

ESEMPIO

UN COMMERCIANTE HA VENDUTO 5 DIVERSI MODELLI DI UN PRODOTTO RIPORTANDO I GUADAGNI UNITARI INDICATI NELLA SEGUENTE TABELLA:

Modello	Guadagno netto per un pezzo	Numero di pezzi venduti
A	0.35	60
B	0.53	100
C	0.60	85
D	0.75	35
E	0.80	10

IL GUADAGNO MEDIO È

$$\bar{x} = \frac{0{,}35 \cdot 60 + 0{,}53 \cdot 100 + 0{,}60 \cdot 85 + 0{,}75 \cdot 35 + 0{,}80 \cdot 10}{290} = 0{,}5456 \cong 0{,}55$$

NEL CASO IN CUI I DATI SONO RAGGRUPPATI PER CLASSI SI ASSUME COME VALORE RAPPRESENTATIVO DI OGNI CLASSE IL SUO VALORE CENTRALE.

MODA

IN UNA DISTRIBUZIONE DI FREQUENZE È QUEL VALORE O MODALITÀ A CUI CORRISPONDE LA MASSIMA FREQUENZA.
UNA DISTRIBUZIONE PUÒ AMMETTERE PIÙ DI UN VALORE MODALE.
NEL CASO IN CUI I DATI SONO RAGGRUPPATI PER CLASSI SI INDIVIDUA LA CLASSE MODALE, E SE LE CLASSI HANNO LA STESSA AMPIEZZA, LA MODA È DATA DALLA FORMULA

$$moda = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot a$$

DOVE

l_1: LIMITE INFERIORE CLASSE MODALE

Δ_1: FREQUENZA CLASSE MODALE MENO FREQ. CLASSE PREC.

Δ_2: FREQUENZA CLASSE MODALE MENO FREQ. CLASSE SUCC.

a: AMPIEZZA CLASSE

ESEMPI

1 NELL'ESEMPIO PRECEDENTE ABBIAMO UNA DISTRIBUZIONE UNIMODALE PERCHÉ È PRESENTE UN SOLO VALORE MODALE RAPPRESENTATO DAL "MODELLO B" CON FREQUENZA PARI A 100.

2 LA DISTRIBUZIONE IN CLASSI DELLE ALTEZZE IN CENTIMETRI DI 20 STUDENTI È:

Classi di altezza	Frequenze assolute
151–155	4
156–160	9
161–165	15
166–170	7
171–175	8
176–180	3
181–185	3
186–190	1

Statistica descrittiva (ESEMPI)

LA CLASSE MODALE È 161-165, l_1 È 160,5 PER VIA DEI VALORI CHE APPROSSIMATI PER ECCESSO ENTRANO NELLA CLASSE, $\Delta_1 = 15 - 9 = 6$, $\Delta_2 = 15 - 7 = 8$ E $a = 5$, COSÌ LA MODA È:

$$moda = 160{,}5 + \frac{6}{6+8} \cdot 5 = 162{,}64$$

MEDIANA

DATI $x_1, x_2, \ldots, x_n$ n VALORI QUANTITATIVI ORDINATI IN SENSO CRESCENTE, SE n È DISPARI LA MEDIANA È IL DATO CHE OCCUPA LA POSIZIONE CENTRALE, MENTRE SE n È PARI LA MEDIANA È LA MEDIA ARITMETICA DEI DUE DATI CENTRALI_

ESEMPIO

SI CHIEDONO A 7 OPERATORI DI UN CALL-CENTER IL NUMERO DI TELEFONATE CHE HANNO EFFETTUATO IN UN CERTO GIORNO E SI OTTENGONO I SEGUENTI DATI

8 / 5 / 5 / 8 / 7 / 6 / 8

PER DETERMINARE LA MEDIANA BISOGNA DISPORLI IN MODO CRESCENTE:

5 / 5 / 6 / 7 / 8 / 8 / 8

ESSENDO $n = 7$ DISPARI, LA MEDIANA CORRISPONDE AL DATO CENTRALE, CIOÈ 7_

SE INVECE IL NUMERO DEI DATI È 6 CIOÈ PARI:

6 / 7 / 7 / 9 / 11 / 12

LA MEDIANA È LA MEDIA ARITMETICA DEI DUE DATI CENTRALI, CIOÈ

$$\frac{7+9}{2} = \frac{16}{2} = 8$$

Statistica descrittiva (ESEMPI)

SE INVECE I DATI SONO RAGGRUPPATI CON LE FREQUENZE BISOGNA INNANZI TUTTO DETERMINARE LA O LE POSIZIONI CENTRALI E STABILIRE POI QUALE DATO OCCUPA QUELLA POSIZIONE MEDIANTE LE FREQUENZE CUMULATE, CIOÈ A PARTIRE DALLA SECONDA, SI SOMMA OGNI FREQUENZA CON LE PRECEDENTI.

LA FREQUENZA CUMULATA MAGGIORE O UGUALE ALLA POSIZIONE CENTRALE FORNISCE IL VALORE DELLA MEDIANA.

ESEMPIO

RIPORTIAMO LA DISTRIBUZIONE DI FREQUENZE DELLE TELEFONATE EFFETTUATE DA UN CALL-CENTER NELLA TABELLA SEGUENTE, NELLA QUALE L'ULTIMA COLONNA RAPPRESENTA LE FREQUENZE CUMULATE:

numero di telefonate	Frequenze assolute	Frequenze cumulate
3	11	11
4	17	28
5	38	66
6	32	98
7	43	141
8	50	191

DOVE

$11=11$

$28=17+11$

$66=38+17+11$

$98=32+38+17+11$

$141=43+32+38+17+11$

$191=50+43+32+38+17+11$

ESSENDO IL TOTALE DEI DATI 191 CIOÈ DISPARI, LA POSIZIONE CENTRALE È LA 96-ESIMA.

IN CORRISPONDENZA DELLA FREQUENZA CUMULATA $98>96$ SI HA IL VALORE DELLA MEDIANA, CIOÈ 6.

SE INVECE SI HA UNA DISTRIBUZIONE IN CLASSI, SI UTILIZZA LA FORMULA SEGUENTE:

$$\text{mediana} = l_1 + (P - F_c)\frac{a}{f}$$

DOVE:

l_1: LIMITE INFERIORE CLASSE MEDIANA (CHE CONTIENE LA MEDIANA)

$P-F_c$: POSIZIONE CENTRALE MENO FREQUENZA CUMULATA CLASSE PRECEDENTE QUELLA MEDIANA

a: AMPIEZZA CLASSE MEDIANA

f: FREQUENZA ASSOLUTA CLASSE MEDIANA

ESEMPIO

CONSIDERIAMO LA DISTRIBUZIONE IN CLASSI DELLE ALTEZZE IN CENTIMETRI DI 20 STUDENTI CON LE FREQUENZE CUMULATE:

Classi di altezza	Freq. Assolute	Freq. Cumulate
151–155	4	4
156–160	9	13
161–165	15	28
166–170	7	35
171–175	8	43
176–180	3	46
181–185	3	49
186–190	1	50

IL NUMERO TOTALE DEI DATI È 50 CIOÈ DISPARI QUINDI LE POSIZIONI CENTRALI SONO LA 25-ESIMA E LA 26-ESIMA ED ESSENDO $28>25$ E $28>26$ CADONO ENTRAMBE NELLA CLASSE 161-165 CHE È QUINDI LA CLASSE MEDIANA. LA POSIZIONE CENTRALE P È LA MEDIA ARITMETICA DELLE 2 POSIZIONI CENTRALI:

$$P = \frac{25+26}{2} = \frac{51}{2} = 25{,}5$$

Statistica descrittiva (ESEMPI)

MENTRE

$$\ell_1 = 160{,}5 \quad \text{(PER LE APPROSSIMAZIONI PER ECCESSO)}$$

E

$$a = 5 \quad f = 15 \quad F_c = 13$$

COSÌ

$$\text{mediana} = 160{,}5 + (25{,}5 - 13) \cdot \frac{5}{15} = 164{,}666 \simeq 164{,}67$$

INDICI DI VARIABILITÀ

LA VARIANZA

SI INTENDE SCARTO DALLA MEDIA LA DIFFERENZA TRA UN QUALSIASI VALORE DELLA DISTRIBUZIONE E LA MEDIA ARITMETICA, CIOÈ

$$x_i - \bar{x}$$

SI DEFINISCE COSÌ LA VARIANZA CHE È LA MEDIA ARITMETICA DEI QUADRATI DEGLI SCARTI, CIOÈ

$$V = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \dots + (x_n-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n}$$

E NEL CASO IN CUI SONO NOTE LE FREQUENZE

$$V = \frac{(x_1-\bar{x})^2 f_1 + (x_2-\bar{x})^2 f_2 + \dots + (x_n-\bar{x})^2 f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2 f_i}{\sum_{i=1}^{n} f_i}$$

ESEMPIO

RIPRENDENDO LA DISTRIBUZIONE DI FREQUENZE DELLE TELEFONATE EFFETTUATE DA UN CALL-CENTER CON LA TABELLA SEGUENTE

numero di telefonate	Frequenze assolute
3	11
4	17
5	38
6	32
7	43
8	50
Totale	191

SI HA

$$\bar{X}=\frac{3\cdot 11+4\cdot 17+5\cdot 38+6\cdot 32+7\cdot 43+8\cdot 50}{11+17+38+32+43+50}=\frac{1184}{191}=6{,}1989\cong 6{,}20$$

E

$$V=\frac{(3-6{,}20)^2\cdot 11+(4-6{,}20)^2\cdot 17+(5-6{,}20)^2\cdot 38+(6-6{,}20)^2\cdot 32+(7-6{,}20)^2\cdot 43+(8-6{,}20)^2\cdot 50}{191}$$

$$V=\frac{440{,}4398}{191}\cong\frac{440{,}44}{191}=2{,}3059\cong 2{,}31$$

DEVIAZIONE STANDARD (SCARTO QUADRATICO MEDIO)

SI DEFINISCE DEVIAZIONE STANDARD O SCARTO QUADRATICO MEDIO, IN SIMBOLI $\hat{\sigma}$ (SIGMA) LA RADICE QUADRATA DELLA VARIANZA, CIOÈ

$$\hat{\sigma}=\sqrt{VAR}$$

NELL'ESEMPIO PRECEDENTE SI HA

$$\hat{\sigma}=\sqrt{2{,}31}=1{,}5198\cong 1{,}52$$

INDICI SULLA FORMA

CONFRONTO TRA MEDIA E MEDIANA

UN PRIMO INDICE SULLA FORMA DI UNA DISTRIBUZIONE DI FREQUENZE È IL CONFRONTO TRA MEDIA E MEDIANA

PERCHÉ SE

$$MEDIA = MEDIANA$$

LA DISTRIBUZIONE PUÒ ESSERE SIMMETRICA, QUINDI SE

$$MEDIA - MEDIANA > 0$$

LA DISTRIBUZIONE È ASIMMETRICA POSITIVA, MENTRE SE

$$MEDIA - MEDIANA < 0$$

LA DISTRIBUZIONE È ASIMMETRICA NEGATIVA.

COEFFICIENTE DI ASIMMETRIA

PER STABILIRE SE UNA DISTRIBUZIONE DI FREQUENZE È SIMMETRICA, CIOÈ IL SUO GRAFICO È TAGLIATO DA UNA RETTA (DETTO ASSE DI SIMMETRIA) CHE LO DIVIDE IN DUE PARTI UGUALI, SI PUÒ RICORRERE AL CALCOLO DEL COEFFICIENTE DI SIMMETRIA, IN SIMBOLI (GAMMA):

$$\gamma = \frac{1}{n \cdot \sigma^3} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3 \cdot f_i$$

NEL CUI CALCOLO RIENTRANO IL CUBO DELLO SCARTO QUADRATICO MEDIO E IL CUBO DEGLI SCARTI DALLA MEDIA E NON AVENDO UNITÀ DI MISURA SI PRESTA COSÌ AI CONFRONTI.

SE

$\gamma > 0$ ASIMMETRICA POSITIVA

$\gamma = 0$ SIMMETRICA

$\gamma < 0$ ASIMMETRICA NEGATIVA