

CORSO DI STATISTICA 1

Questi appunti sono una raccolta riveduta e corretta di dispense e appunti di lezione tratti da numerosi corsi di statistica delle più frequentate facoltà universitarie italiane. Sono dedicati agli studenti di svariate discipline scientifiche ma non solo e rappresentano la base per affrontare lo studio della statistica. Per iniziare con questo livello è richiesta una conoscenza elementare dell'algebra di base.

Se desideri approfondire gli argomenti trattati in queste pagine o chiarire dubbi o semplicemente fare pratica ed esercitazioni non esitare a contattarci alla mail corsionline@geaservizi.org per fissare un' appuntamento con il tutor a te dedicato.

Indice

Introduzione.....	3
Definizioni.....	4
Rilevazione dati e frequenza	5
Variabili statistiche e distribuzioni di frequenze.....	7
La funzione di ripartizione	10
I quantili	13
Box Plot.....	15
Misure in sintesi delle variabili statistiche	17
Indici di posizione	17
Indici di variabilità	20
Esercizi svolti	22

Introduzione

La statistica è la disciplina che fornisce la metodologia per lo studio di fenomeni collettivi. Per programmare l'analisi statistica di un fenomeno vanno seguiti diversi punti generici:

- Definizione degli obiettivi
- Raccolta delle informazioni utili
- Definizione dei mezzi da utilizzare
- Scale di misura
- Limitazioni spazio temporali
- Tecniche di elaborazione dati
- Presentazione dei dati.

I primi 5 punti costituiscono il piano di rilevazione, ovvero la fase di raccolta dati. Questa fase è fondamentale per una buona elaborazione, infatti seguendo il gergo “best in best out” migliori dati raccolti corrispondono a migliori output.

Parlando di tecniche di rilevazione, possiamo definire due tipologie:

- Totale (ovvero un censimento) o parziale (ovvero un campionamento)
- Occasionale o periodica
- Con raccolta dati da fonti o misurati.

Definizioni

Collettivo statistico (popolazione) U : insieme delle unità reali o virtuali “ u ” (individui, oggetti, etc..) mediante le quali si manifesta il fenomeno oggetto di studio.

- Se $u = \infty$ unità reale
- Se $u \neq \infty$ unità virtuale

Numerosità N : $N > 0$, numero di unità statisticamente oggetto di osservazione.

$$U = \{u_1; u_2; \dots u_n\}$$

Carattere: grandezza attraverso la quale viene percepito il fenomeno in esame. Risponde alla domanda cosa?

Modalità M : il modo in cui concretamente si manifesta il carattere relativamente all' unità statistica osservata. Risponde alla domanda come?

Scale di misura: (classificazione di Stevens, 1946)

- 1) **Scala nominale**: le modalità si identificano in attributi non legate alle relazioni di ordine naturale (sesso, corsi di laurea etc..)
- 2) **Scala ordinale**: le modalità si identificano in attributi con una relazione di ordine naturale (voto, classifiche..)
- 3) **Scala per intervalli**: le modalità sono numeri di un sistema. Per un confronto faccio la differenza (temperature, voto diploma..)
- 4) **Scala per rapporti**: modalità sono numeriche con origine assoluta coincidente con lo zero del carattere. Gli elementi possono essere relazionati con la divisione (reddito, spesa, statura...)

Le scale permettono di definire il carattere: qualitativo o quantitativo.

Qualitativo: definito tramite attributi non necessariamente numerabili, può essere a sua volta diviso in:

- Sconnesso: nominale, senza ordine di lettura
- Ordinale: tramite un ordine gerarchico non necessariamente numerico.

Qualitativo: definito tramite valori numerici, può essere a sua volta diviso in:

- Discreto: numeri relativi
- Continuo: numeri naturali
- In classi: raggruppamenti di ampiezza continua.

Rilevazione dati e frequenza

La fase di rilevazione dei dati può essere suddivisa in base al numero di elementi che costituiscono la popolazione, in particolare, la rilevazione può essere:

- Totale, ovvero svolta tramite un censimento (U finito)
- Parziale, ovvero un campionamento (indago solo una parte di U)

La modalità di rilevazione a sua volta si divide in:

- Misurazione diretta
- Per intervista o questionario

Frequenza assoluta n

Numero di unità statistiche che presentano una data modalità (x_i). Sono numeri interi, positivi e la somma da

$$\sum_{i=1}^n n_i = N$$

con n positivo.

Frequenza relativa f

Numero di unità statistiche sul totale che presentano una data modalità (x_i). Sono positive e la somma da 1:

$$\sum_{i=1}^n f_i = 1$$

Con $0 \leq f_i \leq 1$

Esempio

$U = \{u_1, u_2, \dots, u_n\}$ definire la frequenza assoluta e quella relativa degli elementi con posizione pari.

Svolgimento: considero l'insieme $A = \{u_2, u_4, u_6\}$

$$n(A) = 3$$

$$f(A) = 3/7$$

Proprietà della frequenza

- 1) $n(A), f(A) \geq 0 \quad \forall A \in U$
- 2) $n(U) = N; f(U) = 1$
- 3) Dati $A, B \subseteq U, A \cap B = \emptyset$
- 4) Dati A e A^C allora $F(A^C) = 1 - F(A)$

Variabili statistiche e distribuzioni di frequenze

Condizioni:

- U con $N > 0$
- M : $\forall u \exists!$ elemento di M (non si possono avere due risposte diverse).

Variabile statistica:

È la funzione X : U ha immagini in M : $\forall u \in U, \exists! M$

Se M è costituita da modalità con carattere qualitativo, allora X è nominale od ordinale.

Se M è costituita da modalità con carattere quantitativo, allora X è discreta o continua in classi.

Distribuzione di frequenze

Considero l'insieme IMMAGINE della funzione X

$Im(X) = \{X=x_i\} = \{x_1, x_2, x_3, \dots, x_n\}$ con $k > 0$ numero delle modalità e $k \leq n$

Indico con $\{X = x_i\}$ l'insieme di tutte le unità statistiche su cui è stata osservata la modalità x_i

Es: $X = \{25\} = \{\text{matematica, inglese}\} = \{u_2, u_3\}$

Indico la $N(A)$ di X con n_i (numerosità) e la $F(A)$ di X con f_i , $\rightarrow f_i = F(X = x_i) = n_i/N$

Con la frequenza assoluta n_i o relativa f_i , effettuata per tutte le modalità osservate posso costruire la DISTRIBUZIONE DI FREQUENZE

Def: distribuzione di frequenze assolute N_x è l'insieme delle coppie $\{(x_i, n_i), i=1 \dots k\}$. In forma tabellare

X	$N(x=x_i)$
x_1	n_1
x_2	n_2
x_k	n_k

$$\sum_{i=1}^k n_i = N$$

Def: distribuzione di frequenze relative: F_x è l'insieme delle coppie $\{(x_i, f_i), i=1 \dots k\}$. In forma tabellare

X	$F(x=x_i)$
x_1	f_1
x_2	f_2
x_k	f_k

$$\sum_{i=1}^k f_i = 1$$

In forma algebrica:

$$\sum_{i=1}^k n_i = N$$

Da cui:

$$\sum_{i=1}^k f_i = \sum_{i=1}^k n_i / N = 1/N \sum_{i=1}^k n_i = 1$$

Esempio

Viene svolta un'indagine sul gradimento di un determinato prodotto. Viene chiesto un giudizio (scarso, sufficiente, buono, ottimo). Identificare: Collettivo statistico, numerosità, carattere, modalità, variabile statistica ed organizzare un data set calcolando le frequenze assolute e relative.

Svolgimento:

- Collettivo statistico: risponde alla domanda chi? O cosa? – i clienti che hanno acquistato il prodotto.
- Numerosità: $N = 6$
- Carattere: giudizio, qualitativo ordinato.
- Modalità: voti $M = \{\text{scarso, suff, buono, ottimo}\}$
- Variabile statistica: $x = \text{"giudizio espresso"}$
- Data set

Unità U	$F(x=x_i)$
u₁	Suff
u₂	Buono
u₃	Ottimo
u₄	Ottimo
u₅	Buono
u₆	Buono

X (giudizio)	N(x)	$F(x=x_i)$
SUFF	1	1/6
BUONO	3	3/6
OTTIMO	2	2/6

Tabella con frequenze assolute (N) e relative (F)

La funzione di ripartizione

Def: funzione di ripartizione: Fdr di una variabile statistica X nel punto $x \in \mathbb{R}$ è la porzione di collettivo su cui sono osservati valori non superiori a x

1) Caso di v.s. discreta o continua

$$F(x) = F(X \leq x)$$

$$\text{Se } x_1 \leq x < x_2 \text{ allora } F(x) = F_1$$

In modo compatto:

$$F(x) = \begin{cases} 0 & \text{se } x < x_1 \\ F_i & \text{se } x_i \leq x < x_{i+1} \\ 1 & \text{se } x \geq x_i \end{cases}$$

Proprietà

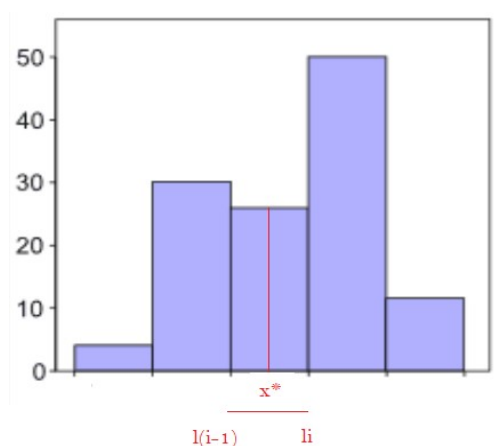
- a) Dominio: $[0; 1]$
- b) Funzione monotona non decrescente
- c) Funzione continua a destra
- d) $F(a < X \leq b) = F(b) - F(a)$
- e) Caso di v.s. discreta o continua

2) Caso di dati in classi

In questo caso non si hanno informazioni circa il valore assunto esattamente da ciascuna unità statistiche all' interno della classe.

Per uno studio semplificato consideriamo una **distribuzione uniforme per ogni classe**.

Con questa semplificazione possiamo costruire un diagramma per istogrammi



Per calcolare la Fdr in x^* devo sommare le aree di tutti i rettangoli prima di x^* + la frazione del rettangolo in cui ho x^*

$$F(x^*) = F(x \leq l_{i-1}) + F(l_{i-1} < x \leq x^*)$$

$$= F_{i-1} + \frac{f_i}{w_i} (x^* - l_{i-1})$$

In modo compatto:

$$F(x) = \begin{cases} 0 & \text{se } x < l_0 \\ F(i-1) + \frac{f_i}{w_i}(x - l_{i-1}) & \text{se } l_{i-1} \leq x < l_i \\ 1 & \text{se } x \geq l_k \end{cases}$$

Rappresenta una funzione lineare a tratti (spezzata). Graficamente, il coefficiente angolare dei segmenti di ciascuna classe è dato dalla densità di frequenza della classe.

$$d(x) = \frac{dF(x)}{dx} = \frac{f_i}{w_i} = h_i$$

ESEMPIO

Data la seguente distribuzione di frequenze di X con dati in classi, trovare $F(15)$.

X	$F(x)$
-10 → 0	0.1
0 → 10	0.5
10 → 30	0.3
30 → 70	0.1

Svolgimento:

Partendo dalla tabella fornita costruisco una tabella con w (ampiezza della classe), F_{cum} (frequenza cumulata) e d (densità della classe).

Applico la formula $F_{i-1} + \frac{f_i}{w_i}(x^* - l_{i-1})$ con $x^* = 15$

X	F(x)	Fcum	w	d
-10 → 0	0.1	0.1	10	0.01
0 → 10	0.5	0.6	10	0.05
10 → 30	0.3	0.9	20	0.015
30 → 70	0.1	1.0	40	0.025

$F(15) \in X_3 (10 \rightarrow 30)$, quindi:

$$F(15) = F_2 + (15 - l_2) \frac{f_i}{w_i} = 0.6 + (15 - 10) \frac{0.3}{20} = 0.675$$

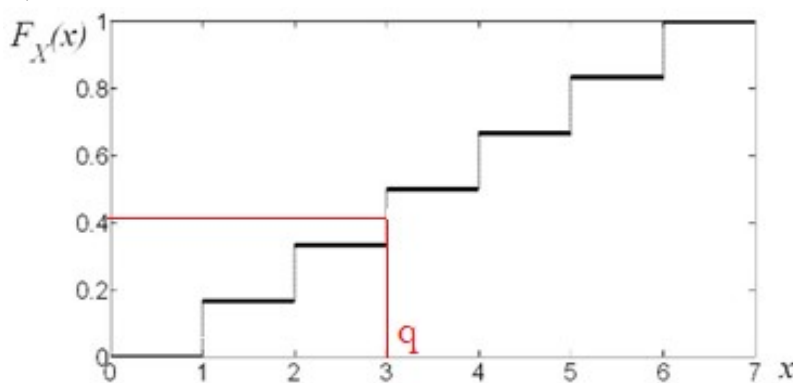
I quantili

Def: quantile di ordine k: sia k un numero reale compreso tra 0 e 1 (100%). Definiamo quantile (o frattile) d'ordine k della variabile statistica x quel valore q_x tale che suddivide le $u \in U$, ordinate crescenti, in 2 gruppi disgiunti aventi frequenze rispettive k e $1-k$.

Es: se $k = 0.25$ di una U di lavoratori dipendenti è $q_{0.25} = 1500\text{€}$ significa che almeno il 25% di U ha reddito ≤ 1500 . Si può sempre costruire, partendo dai quantili, una distribuzione di frequenza con dati in classi:

Redditi	F. relativa
Da 0 a 1500	0.25
>1500	0.75

1) Caso di variabili statistiche con Fdr costante a tratti



In questo caso $q(\alpha) = x_3$

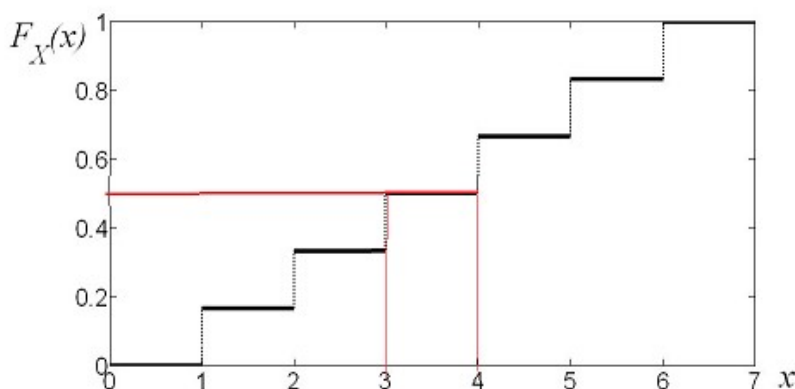
$$F(x) \geq \alpha \rightarrow [x_3; +\infty]$$

$$F(X < x) \leq \alpha \rightarrow (-\infty; x_3)$$

$$\text{Quindi: } q(\alpha) \in (-\infty; x_3] \cup [x_3; +\infty]$$

$$q(\alpha) = x_3$$

Caso 2



In questo caso $x_3 < q(\alpha) < x_4$

$$F(x) \geq \alpha \rightarrow [x_3; +\infty]$$

$$F(X < x) \leq \alpha \rightarrow (-\infty; x_4)$$

$$\text{Quindi: } q(\alpha) \in [x_3; +\infty) \cap (-\infty; x_4]$$

$$q(\alpha) = (x_3 + x_4)/2$$

2) Caso di variabili statistiche con dati in classi

Si deve risolvere l'equazione $F(q\alpha) = \alpha$

$$q(\alpha) = l_{i-1} + w_i \frac{\alpha - F(i-1)}{f_i}$$

Def: quartile: valore q_k tale che $q = 1/4$. Valore che suddivide l'insieme delle osservazioni in due parti contenenti rispettivamente $\frac{1}{4}$ e $\frac{3}{4}$ degli individui, ordinati secondo i valori crescenti di x .

Def: mediana (Me): secondo quartile, $k = \frac{1}{2}$. Valore che suddivide l'insieme delle osservazioni in due parti contenenti ognuna $\frac{1}{2}$ degli individui, ordinati secondo i valori crescenti di x .

Def: decile: valore di k espresso in decimi.

Def: percentile: valore di k espresso in percentuale o in numero reale ≤ 1 . Es: $k = 0.8 \rightarrow q(k) = 80$ esimo percentile.

Quindi: Me = secondo quartile = quinto decile = cinquantesimo percentile.

Box Plot

Si tratta di una rappresentazione grafica utile a descrivere la distribuzione di un campione tramite indici di dispersione e di posizione.



Lower extreme (Adiacente inferiore): è il più piccolo valore di x superiore a

$$q_1 - 1.5(q_3 - q_1)$$

Upper extreme (Adiacente superiore): è il più grande valore di x inferiore a

$$q_1 + 1.5(q_3 - q_1)$$

Esempio

Il numero di esami sostenuti da studenti nella facoltà A in un semestre è:

2 1 2 2 0 3 1 1 5 2

Costruire il Box Plot e stabilire se ci sono dati anomali

X	N	F	F _{cum}
0	1	0.1	0.1
1	3	0.3	0.4
2	4	0.4	0.8
3	1	0.1	0.9
5	1	0.1	1

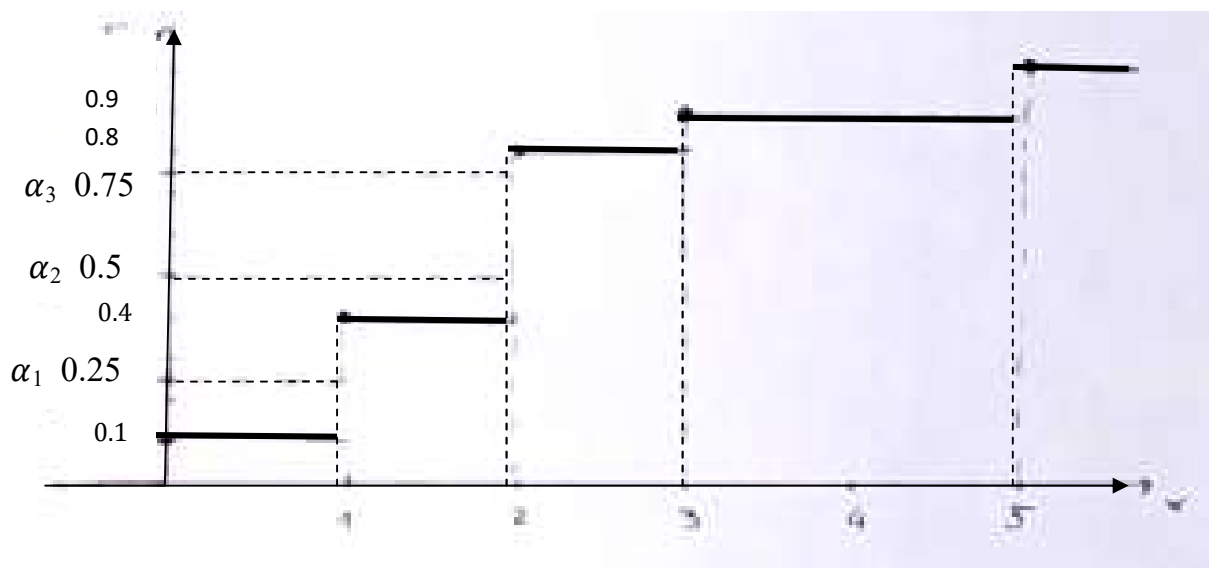
Svolgimento:

Per prima cosa si costruisce la tabella delle frequenze

Costruisco il grafico della Fdr

$$F(x) = \begin{cases} 0 & \text{se } x < 0 \\ 0.1 & \text{se } 0 \leq x < 1 \\ 0.4 & \text{se } 1 \leq x < 2 \\ 0.8 & \text{se } 2 \leq x < 3 \\ 0.9 & \text{se } 3 \leq x < 5 \\ 1 & \text{se } x \geq 5 \end{cases}$$

Grafico



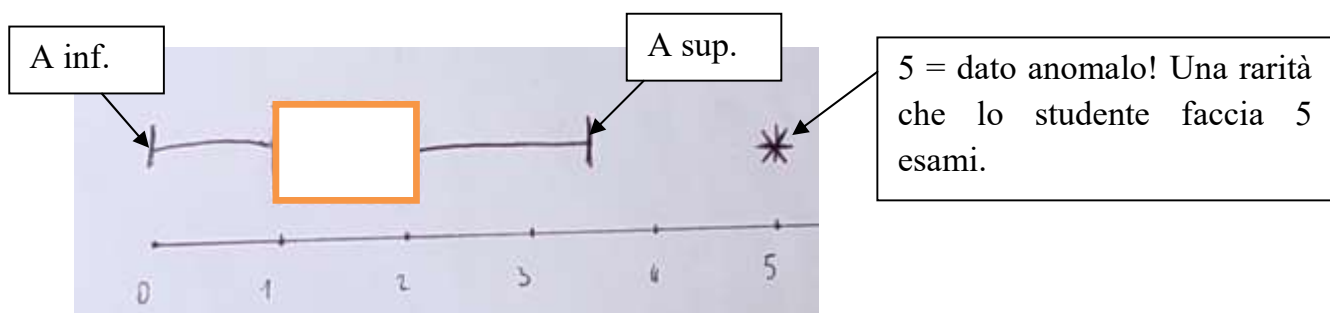
Quantili:

$\alpha_1 = 0.25 \rightarrow Q_1 = 1$ *il 25% ha superato non più di un esame*
 $\alpha_2 = 0.5 \rightarrow Q_2 = 2$ $\left. \begin{array}{l} \alpha_3 = 0.75 \rightarrow Q_3 = 2 \end{array} \right\}$ *il 50% e il 75% non ha superato più di 2 esami*

Box plot:

$$A_{inf} = Q_1 - 1.5(Q_3 - Q_1) = 1.5(2 - 1) = 1.5$$

$$A_{sup} = Q_3 + 1.5(Q_3 - Q_1) = 2 + 1.5 = 3.5$$



Misure in sintesi delle variabili statistiche

Sono degli indici che danno informazioni in merito a taluni aspetti specifici. In particolare si dividono in indici di posizione e indici di variabilità.

Indici di posizione

Valore medio di Cauchy: sia x una variabile statistica quantitativa discreta o continua, con $Im(x) = \{x_1, x_2, ..x_k\}$. Si dice valore medio di x un qualsiasi valore m t.c:

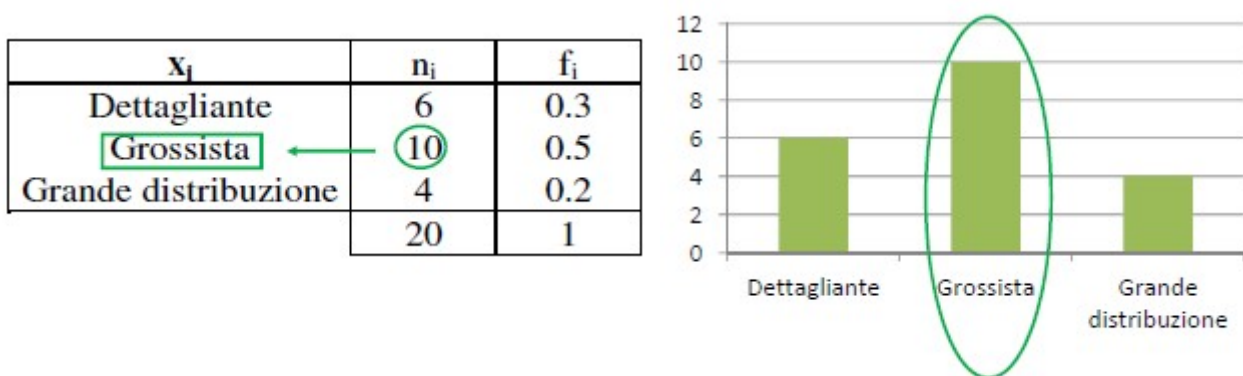
$$x_1 \leq m \leq x_k$$

Mid Range (MR): è il valore centrale che sta a metà tra il minimo (x_1) e il massimo (x_k)

$$MR(x) = (x_1 + x_k)/2$$

N.B: ha poco senso se la distribuzione non è omogenea

Moda (Mo(x)): è quella modalità che si presenta con la massima frequenza. Di una variabile statistica con dati in classi si parla di “classe modale” cioè la classe con massima densità di frequenza (rettangolo più alto nell’ istogramma).



Nell’ esempio riportato la moda $Mo(x) = \text{grossista}$. La moda è una modalità e non una frequenza!

In un carattere espresso in dati in classi la $Mo(x)$ non è il carattere con la massima frequenza ma quello con la massima densità (d), ovvero il rapporto tra la frequenza (n) e l’ ampiezza della classe (a).

$$Mo(x) = \max(d) = \max\left(\frac{n}{a}\right)$$

Classi	ni	ai	di = ni/ai
7- 9	40	2	20
9- 11	25	2	12.5
11- 15	120	4	30
15- 21	145	6	24.17
	330		

Cerco la densità massima

Mediana (Me(x)): è il quantile di ordine 0.5. ovvero quel valore che non viene superato dal 50% delle unità statistiche ordinate crescenti. Suddivide U in due gruppi di pari numerosità, una sotto e uno sopra la mediana.

Proprietà :

- Internalità: Me(x) è media secondo Cauchy
- Robustezza: Me(x) non è sensibile ai valori anomali

Per calcolare la Mediana devo procedere nel seguente modo:

- 1) Le modalita devono presentarsi ordinate
- 2) Calcolo la posizione mediana p: per n dispari/pari

Dispari: $(n+1)/2$

Pari: $(n/2)$ e $(n/2)+1$

- 3) Cerco la posizione mediana sulla colonna delle FREQUENZE CUMULATE N_i

si individua la prima N_i maggiore o uguale di p ($N_i \geq p$)

- 4) Trovo la modalita corrispondente: mediana

Istruzione	n_i	N_i
Elementare	12	12
Media	17	29
Superiore	123	152
Laurea	51	203
TOTALE	203	

n dispari

$$p = (n+1)/2 = (203+1)/2 = 102$$

152 è il primo valore maggiore di 102

$$Me(x) = X_{102} = \text{"Superiore"}$$

Media aritmetica (μ): valore riferito ad un carattere trasferibile, ossia misurato in scala per rapporti che può essere suddiviso e ceduto da una unità ad un' altra. Es: reddito, PIL, n. turisti...

$$\mu = \frac{T}{N} = \frac{1}{N} \sum_{v=1}^N x(u)_v = E(x)$$

Con T = ammontare cumulato $\forall t$

μ rappresenta quel valore che verrebbe attribuito ad ogni unità se fosse possibile ridistribuire T tra tutte le modalità in modo invariante.

Considerando le frequenze, possiamo scrivere come segue:

$$\mu = \frac{1}{N} \sum_{i=1}^k x_i u_i$$

$$\mu = \sum_{i=1}^k x_i f_i$$

Proprietà:

- 1) Internalità: $x_1 \leq \mu \leq x_k$
- 2) Linearità: $E[a+bx] = a + bE(x)$, con $a, b \in \mathbb{R}$ e $b \neq 0$
- 3) Nullità degli scarti: $E(x - \mu) = 0$
- 4) Minimi quadrati: $E[(x - \mu)^2] \leq E[(x - m)^2]$ con M = valore di sintesi
- 5) Sensibilità: μ è poco robusta perché è sensibile ai valori anomali.
- 6) $\mu = Me(x)$ se la distribuzione delle frequenze è simmetrica rispetto a μ

Indici di variabilità

Variabilità: sia $X: U \rightarrow M$ una v.s. quantitativa con distribuzione di frequenze n_x (o f_x). Diciamo variabilità di X la sua attitudine a disperdersi su $Im(x)$. Sono indici che manifestano la dispersione delle misure rispetto ad un valore centrale.

Range = $x_{\max} - x_{\min}$, definisce l' intervallo dove rientra il 100% della popolazione.

Differenza interquantile (IQR) = $Q_3 - Q_1$ è l' intervallo dove si distribuisce il 50% dei dati e corrisponde alla larghezza del BOX PLOT.

Scarto quadratico medio (σ ; sqm): rappresenta la media delle differenze tra il valore x_i e la media μ . È la radice quadrata della varianza.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$$

Varianza (σ^2 ; $Var(x)$): è il quadrato dello scarto quadratico medio (σ^2) e rappresenta l' indice di variabilità più usato.

Proprietà della varianza

- 1) $Var(k) = 0$ la varianza di una costante è pari a zero.
- 2) $Var(bx) = b^2 Var(x) = Var(bx + k)$ La varianza non è un operatore lineare
- 3) $Var(x + y) = Var(x) + Var(y) + a$ (termine non nullo) Proprietà associativa
- 4) **Teorema di scomposizione della varianza:** la varianza totale di una popolazione divisa in gruppi è data dalla somma della varianza intragruppo e quella intergruppi, o più comunemente **varianza whitin e varianza between**

$$\sigma^2 = \sigma_W^2 + \sigma_B^2$$

In particolare:

varianza whitin: media delle varianze dei gruppi $\sigma_W^2 = M(\sigma_i^2)$

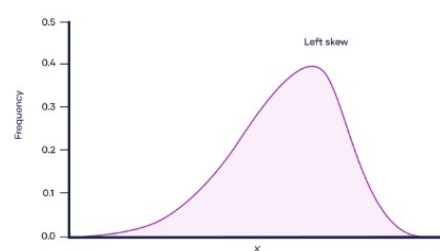
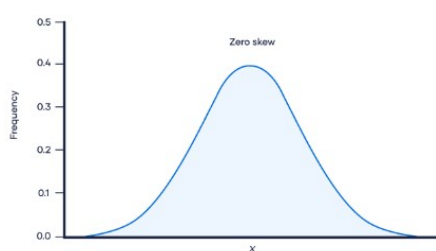
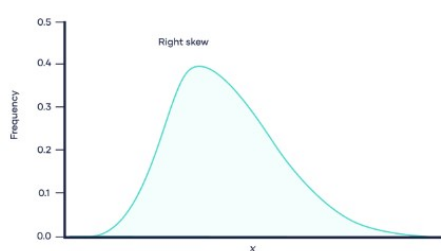
varianza between: varianza delle medie dei gruppi $\sigma_B^2 = V(\mu_i)$

corollario: applicando la proprietà 3) e 4) se una popolazione è divisa in sottogruppi e si conoscono solo numerosità, media e varianza di ogni gruppo, allora è possibile calcolare la media generale e la varianza totale

Misure relative alla variabilità

Si ottengono dividendo gli indici di variabilità per un qualsiasi indice di posizione. Si ottengono numeri adimensionali.

1. Range / μ
2. IQR / μ
3. Coefficiente di variazione $CV = \sigma / \mu$
4. Skweness $SK = \frac{\mu - Me}{SQM}$



Lo Skweness indica la asimmetria della distribuzione. Nelle immagini vediamo una asimmetria positiva ($SK > 0$) nella curva a sinistra, una asimmetria nulla ($SK = 0$) nella curva in centro e una asimmetria negativa ($SK < 0$) nella curva a destra.

Esercizi svolti

- 1) In una catena di montaggio vengono prodotti bulloni di acciaio. Il controllo post produzione individua i pezzi difettosi come riportato nella tabella sottostante. Calcolare la deviazione standard e la varianza.

Bulloni difettosi	0	1	2	3	4	5	6
Frequenza	0.17	0.18	0.25	0.2	0.1	0.08	0.02

Svolgimento:

Costruisco una tabella identificando frequenze, scarti, scarti quadratici e frequenze degli scarti quadratici,

x_i	f_i	$x_i f_i$	$(x_i - \mu)$	$(x_i - \mu)^2$	$f_i(x_i - \mu)^2$
0	0.17	0	-2.2	4.84	0.8228
1	0.18	0.18	-1.2	1.44	0.2592
2	0.25	0.5	-0.2	0.04	0.01
3	0.2	0.6	0.8	0.64	0.128
4	0.1	0.4	1.8	3.24	0.324
5	0.08	0.4	2.8	7.84	0.6272
6	0.02	0.12	3.8	14.44	0.2888
	$\mu = 2.2$				Tot: 2.46

Con i dati indicati calcolo la deviazione standard con la formula (n = numerosità).

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$$

La varianza è il valore sotto la radice cioè totale dell' ultima colonna: $\sigma^2 = 2.46$.
Pertanto la deviazione standard $\sigma = 1.56$

- 2) In una classe quinta di un Istituto superiore sono state rilevate le altezze degli alunni e riportate nella tabella sottostante. Calcolare: media, moda, mediana, deviazione standard e varianza.

Svolgimento:

Costruisco le colonne di frequenze cumulate, scarti e quadrati degli scarti

X_i	n	N_{cum}	$(x_i - \mu)$	$(x_i - \mu)^2$
166	1	1	-4.9	24.01
168	3	4	-2.9	8.41
169	6	10	-1.9	3.61
170	11	21	-0.9	0.81
171	8	29	0.1	0.01
172	6	35	1.1	1.21
173	4	39	2.1	4.41
174	3	42	3.1	9.61
175	1	43	4.1	16.81
178	1	44	5.1	26.01

Media: $\mu = \frac{1}{N} \sum_{i=1}^k x_i u_i = 170.9 \text{ cm}$

Moda: $Mo(x) = 171$ (valore con massima n)

Mediana: $Me(x) = 171$ ($44/2 = 22$, si prende la classe appena superiore)

Varianza: $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} = 135.49/44 = 3.079$

Sqm: $\sigma = 1.754$

- 3) Nella seguente tabella sono riportate le durate di un gruppo di mutui. Disegna il Box-plot.

Durata in anni	n
0- 4	10
4- 8	53
8- 12	51
12- 16	30
16- 20	15

Svolgimento:

Per prima cosa si calcolano Mediana, primo e terzo quartile;

$$Med = Q_2 = X_{i-1} + \frac{(X_i - X_{i-1})(0,5 - F_{i-1})}{(F_i - F_{i-1})}$$

$$Q_1 = X_{i-1} + \frac{(X_i - X_{i-1})(0,25 - F_{i-1})}{(F_i - F_{i-1})}$$

$$Q_3 = X_{i-1} + \frac{(X_i - X_{i-1})(0,75 - F_{i-1})}{(F_i - F_{i-1})}$$

X classi	n	f	F cumulate
0- 4	10	$10/160 = 0.0625$	0.0625
4- 8	53	$10/160 = 0.1563$	0.2188
8- 12	51	$10/160 = 0.3125$	0.5313
12- 16	30	$10/160 = 0.2813$	0.8126
16- 20	15	$10/160 = 0.1875$	1
Totale	160		

Per calcolare la mediana trovare la classe con frequenza relativa cumulata F_i immediatamente superiore al 0,5. In questo caso la classe è 8 -| 12, quindi $X_i=12$ e $X_{i-1}=8$; $F_i=0,5313$; $F_{i-1}=0,2188$.

$$Med = Q_2 = X_{i-1} + \frac{(X_i - X_{i-1})(0,5 - F_{i-1})}{(F_i - F_{i-1})} = 8 + \frac{(12 - 8) \cdot (0,5 - 0,2188)}{0,5313 - 0,2188} = 11,5994$$

Per calcolare il primo quartile Q_1 occorre considerare la classe con frequenza relativa cumulata immediatamente superiore a 0,25. In questo caso la classe è ancora la 8 -| 12, quindi $X_i=12$ e $X_{i-1}=8$; $F_i=0,5313$; $F_{i-1}=0,2188$

$$Q_1 = X_{i-1} + \frac{(X_i - X_{i-1})(0,25 - F_{i-1})}{(F_i - F_{i-1})} = 8 + \frac{(12 - 8) \cdot (0,25 - 0,2188)}{0,5313 - 0,2188} = 8,3994$$

Per calcolare il terzo quartile Q_3 si prende la classe con frequenza relativa cumulata immediatamente superiore a 0,75. In questo caso la classe è la 12 -| 16, quindi

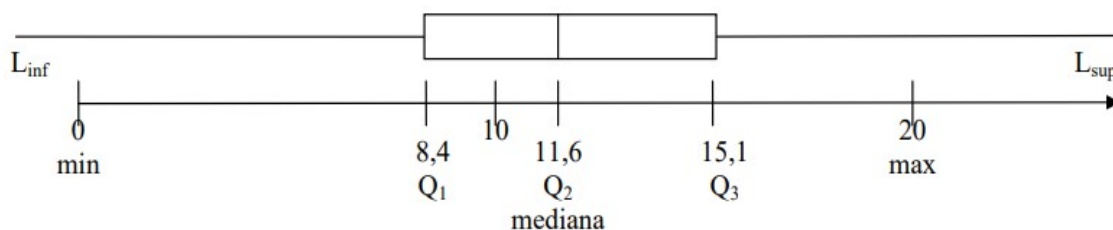
$X_i=16$ e $X_{i-1}=12$; $F_i=0,8126$, $F_{i-1}=0,5313$.

Si calcolano il limite inferiore L_{inf} e il limite superiore L_{sup} :

$$L_{inf} = Q_1 - 1,5(Q_3 - Q_1) = 8,3994 - 1,5 \cdot (15,1098 - 8,3994) = -1,6662$$

$$L_{sup} = Q_3 + 1,5(Q_3 - Q_1) = 15,1098 + 1,5 \cdot (15,1098 - 8,3994) = 25,1754$$

Infine rappresentiamo su un segmento i valori X_{min} , X_{max} , Q_1 , Q_2 , Q_3 , L_{inf} , L_{sup} , tenendo conto che: se $L_{inf} < X_{min}$ rappresentiamo L_{inf} , altrimenti X_{min} ; se $L_{sup} > X_{max}$ rappresentiamo L_{sup} , altrimenti X_{max} .⁴



(fonte: matematicamente.it)