01 – Confrontare più distribuzioni

Consideriamo la distribuzione per età di tre diversi collettivi di studenti

Gruppo A	Gruppo B	Gruppo C
22	22	25
22	22	25
23	22	25
23	22	25
24	22	25
25	25	25
26	28	25
27	28	25
27	28	25
28	28	25
28	28	25

In che modo possiamo confrontare i tre gruppi?

Se calcoliamo la media aritmetica e la mediana osserviamo per i tre collettivi lo stesso valore:

$$\bar{x}_a = 25 \qquad Me = 25$$

ma da una prima analisi dei valori delle diverse distribuzioni si osserva subito che ci sono delle differenze

Risulta quindi difficile effettuare un confronto utilizzando i soli indici di posizione

Abbiamo bisogno di un'altra classe di indici che tenga conto della variabilità del fenomeno

02 – Variabilità di un fenomeno

Unità n° 04

Con il termine **variabilità** si suole indicare l'**attitudine di un carattere quantitativo ad assumere modalità diverse**. Lo studio della variabilità è di fondamentalmente importanza:

1. **Valore Intrinseco**
La conoscenza della variabilità è alla base della Statistica: se tutte le manifestazioni di un fenomeno fossero uguali tra loro la rilevazione di una singola modalità consentirebbe la conoscenza della totalità del fenomeno, quindi non avrebbe più senso uno studio statistico

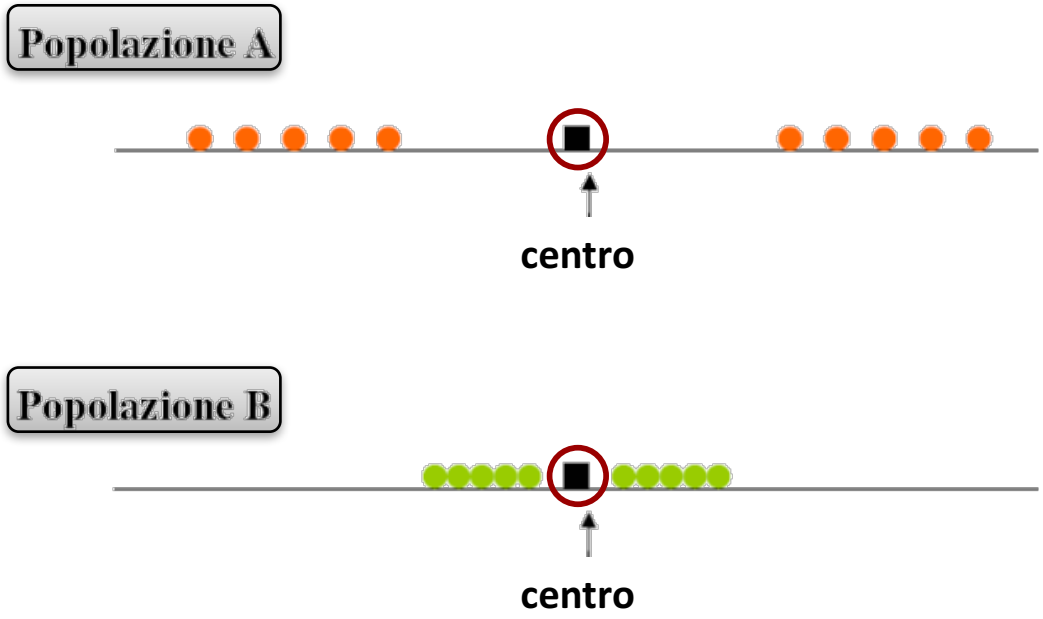
2. **Accuratezza della Sintesi dei Dati**
L'impiego delle medie (sia di posizione che analitiche) non è sufficiente a sintetizzare le informazioni rilevate sulla popolazione oggetto di studio, specialmente quando occorre confrontare tra loro popolazioni

03 – Dispersione dei dati nella distribuzione

Unità n° 04

Le misure di variabilità consentono di valutare il grado di **dispersione** delle modalità e la bontà della sintesi della distribuzione operata mediante le misure di centralità

Può accadere, come visto, che due o più popolazioni presentino lo stesso centro, ma che il livello di sintesi sia completamente differente. Consideriamo la seguente esemplificazione relativa a due popolazioni:



E' immediato che la sintesi effettuata tramite l'indice di centralità è più significativa nella popolazione B, perché le osservazione sono maggiormente addensate intorno al centro

04 – Indici di Variabilità

Unità n° 04

A seconda degli aspetti della variabilità che si vuole mettere in evidenza è necessario calcolare indici di variabilità diversi:

A) Indici che si basano sulla differenza tra i valori che occupano determinate posizioni in un dato ordinamento delle unità del collettivo

B) Indici che si basano sugli scostamenti delle osservazioni da un valore medio

C) Indici che si basano sulle differenze tra tutte le modalità osservate

Un'altra classificazione che viene spesso adottata è quella fra **indici assoluti** e **indici relativi**

Gli **indici assoluti di variabilità** sono espressi nella stessa unità di misura con la quale si rilevano le modalità del carattere

Gli **indici relativi di variabilità** sono invece adimensionali, sono cioè "numeri puri" (non sono espressi in nessuna unità di misura)

05 – Proprietà generali degli indici di variabilità

Unità n° 04

Affinché un indice $V(x_1, ..., x_N)$ calcolato a partire dalle osservazioni $(x_1, ..., x_N)$ di un carattere X su un collettivo di numerosità N sia idoneo a misurare la variabilità occorre che:

1) $V(x_1, ..., x_N) = 0$ se e solo se il carattere assume lo stesso valore $x_1 = ... = x_N$ in tutte le unità del collettivo (il fenomeno si manifesta sempre nello stesso modo)

2) se almeno due osservazioni x_i e x_j sono diverse tra loro allora $V(x_1, ..., x_N) > 0$ (la variabilità aumenta all'aumentare della diversità tra modalità)

3) l'indice è invariante rispetto a traslazioni, cioè $V(x_1, ..., x_N) = V(x_1 + c, ..., x_N + c)$

4) un carattere X è più variabile di un carattere Y, in uno stesso collettivo, se risulta $V(x_1, ..., x_N) > V(y_1, ..., y_N)$

N.B.: per i caratteri qualitativi sarebbe più corretto parlare di <u>MUTABILITA'</u>

06 – Eterogeneità e omogeneità

Unità n° 04

A prima vista una distribuzione con una elevata variabilità potrebbe sembrare più complessa da analizzare rispetto ad una distribuzione con una bassa o nulla variabilità

In Statistica in realtà la variabilità può essere vista da diversi punti di vista, a seconda che si focalizzi l'attenzione sul fenomeno o sulle unità del collettivo

variabilità come ricchezza di informazione

Possiamo considerare il fatto che una elevata variabilità, ossia una maggior dispersione dei valori della distribuzione intorno al suo centro, implica una maggior ricchezza di informazione:
da questo punto di vista più la distribuzione è variabile più il fenomeno tende a manifestarsi in modo diverso nel collettivo e quindi abbiamo maggiori elementi per poterlo studiare

variabilità come elemento di discriminazione

se il nostro obiettivo è quello di classificare le unità statistiche in gruppi omogenei rispetto ad una o più caratteristiche allora una bassa variabilità in ciascun gruppo, rispetto ad una elevata variabilità tra i gruppi, consente di separare le unità statistiche e quindi ottenere una migliore informazione

07 – Campo di Variazione

Uno degli indici più semplici tra quelli basati sulla differenza tra valori che occupano determinate posizioni è il cosiddetto **campo di variazione** (o range) della distribuzione:

$$\Delta = x_{(N)} - x_{(1)}$$

E' un indice di variabilità di facile interpretazione poiché rappresenta l'ampiezza della distribuzione del carattere nel collettivo. Il suo impiego è comunque limitato solo a poche applicazioni per una serie di inconvenienti:

1) dipende solo da due osservazioni e non tiene conto delle altre

2) essendo espressione dell'osservazione più grande e di quella più piccola è poco stabile, in quanto estremamente sensibile ai valori anomali

3) presenta difficoltà di calcolo in presenza di classi aperte

08 – Differenza Interquartile

Un altro interessante indice basato sulla differenza tra valori che occupano determinate posizioni è la **differenza interquartile**:

$$IR = Q_3 - Q_1$$

È calcolato come differenza tra terzo e primo quartile della distribuzione e rappresenta l'ampiezza dell'intervallo centrale (quello intorno alla mediana), nel quale si collocano il 50% delle osservazioni

Tanto più piccola è la differenza interquartile tanto più la metà delle osservazioni risulta addensata intorno alla mediana. In tal senso, la distanza interquartile risulta un indice di variabilità interno, nel senso che si riferisce solo al 50% delle unità che presentano valori intorno alla mediana

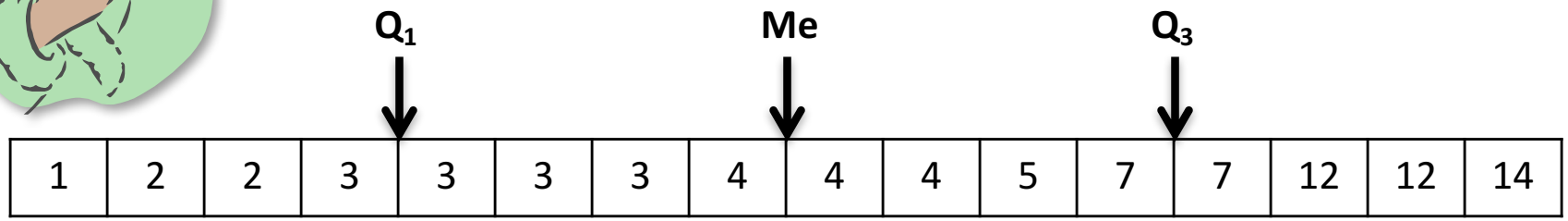
La distanza interquartile presenta alcune peculiarità:

1) è un indice più stabile del campo di variazione perché non si basa sulle osservazioni estreme

2) potrebbe essere nulla senza che il carattere risulti degenere

09 – Esempio

Consideriamo la distribuzione dei giorni di assenza per malattia negli ultimi due mesi osservati sui 16 dipendenti di un'azienda



Il numero medio di giorni di assenza è pari a $\bar{x}_a$ **= 5,375 giorni** mentre il numero mediano di giorni di assenza è **Me = 4 giorni**

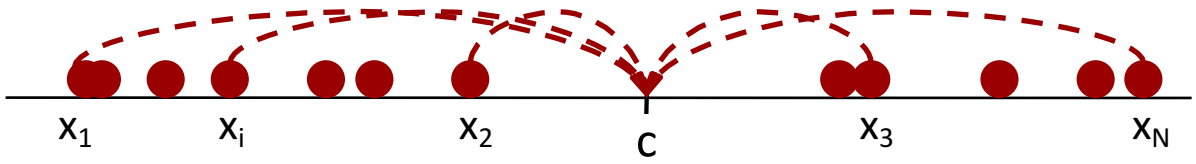
- Se vogliamo calcolare il campo di variazione avremo **Δ = 14 – 1 = 13 giorni**
- Per calcolare la differenza interquartile abbiamo bisogno del primo e del terzo quartile: in questo caso si ottiene $Q_1 = 3$ e $Q_3 = 7$, quindi **IR = $Q_3 - Q_1$ = 4 giorni**

Si nota subito come la presenza di possibili valori anomali, già evidenziata nel calcolo della media aritmetica e delle mediana, influenza il valore del campo di variazione rispetto a quello della differenza interquartile (vedremo meglio in seguito come sfruttare questa informazione)

10 – Indici basati sugli scostamenti dalle medie

Uno dei modi più utilizzati per studiare la variabilità del fenomeno è osservare come si comportano le osservazioni (in termini di carattere osservato) rispetto ad un valore che sia rappresentativo della distribuzione (il suo "centro")

Consideriamo la distribuzione di un carattere quantitativo **X** e supponiamo che **c** sia il centro



Ogni differenza $(x_i - c)$ è chiamata **scarto** e può essere utilizzata per costruire un indice che valuti il livello di dispersione del fenomeno nel collettivo oggetto di studio

In generale possiamo considerare **scostamenti semplici** o **scostamenti quadratici**

11 – Scostamenti semplici

Possiamo costruire due indici di variabilità considerando le differenze in valore assoluto delle modalità della distribuzione dalla media o dalla mediana

Una volta sommati tutti gli scarti è necessario dividere per il numero di unità statistiche del collettivo studiato: il principio è sempre quello della sintesi del fenomeno, ma da un diverso punto di vista -> di quanto in media (±: cioè per valori sopra o sotto la media) le osservazioni si discostano dal valore medio scelto

$$S_{\bar{x}_a} = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \bar{x}_a\right|$$

scostamento semplice medio

$$S_{Me} = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - Me\right|$$

scostamento semplice mediano

Come già visto, l'uso della media nel calcolo dell'indice di variabilità può essere rischiosa nel caso in cui vi siano nella distribuzione dei valori anomali

12 – Esempio

Sono stati registrati gli arrivi di passeggeri all'aeroporto di Lamezia S. Eufemia (in migliaia)

GEN	FEB	MAR	APR	MAG	GIU	LUG	AGO	SET	OTT	NOV	DIC
33.0	30.9	31.0	42.4	37.5	40.0	41.2	48.5	41.7	38.0	32.8	44.2

Il numero medio di passeggeri per mese è pari a 38.4, il numero mediano è invece pari a 39.0 (migliaia)

$\lvert x_i - \bar{x}_a \rvert$	5.43	7.53	7.43	3.97	0.93	1.57	2.77	10.07	3.27	0.43	5.63	5.77
$\lvert x_i - Me \rvert$	6.00	8.10	8.00	3.40	1.50	1.00	2.20	9.50	2.70	1.00	6.20	5.20

$$S_{\bar{x}_a} = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \bar{x}_a\right| \cong 4.57 \qquad S_{Me} = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - Me\right| \cong 4.57$$

Il numero di passeggeri in arrivo mensilmente a Lamezia si è discostato in media di ±4570 unità rispetto al numero medio (o mediano) di passeggeri riscontrato sui 12 mesi

13 – Distribuzioni di frequenza e in classi

Se si calcolano gli scostamenti semplici su distribuzioni di frequenza è necessario tener conto non solo delle modalità osservate ma anche della loro *importanza* in termini di frequenza

$$S_{\bar{x}_a} = \frac{1}{N}\sum_{i=1}^{N}\left[\left|x_i - \bar{x}_a\right| \cdot n_i\right] \qquad S_{Me} = \frac{1}{N}\sum_{i=1}^{N}\left[\left|x_i - Me\right| \cdot n_i\right]$$

Se la distribuzione è in classi la formulazione è la stessa ma al posto delle modalità bisogna considerare i valori centrali di ciascuna classe

$$S_{\bar{x}_a} = \frac{1}{N}\sum_{i=1}^{N}\left[\left|c_i - \bar{x}_a\right| \cdot n_i\right] \qquad S_{Me} = \frac{1}{N}\sum_{i=1}^{N}\left[\left|c_i - Me\right| \cdot n_i\right]$$

Ovviamente è possibile in entrambe le tipologie di distribuzione utilizzare per il calcolo anche le frequenze relative (ricordando che $f_i = n_i/N$)

14 – Scostamenti quadratici

Nella costruzione di un indice di variabilità possiamo considerare al posto del valore assoluto il quadrato degli scarti: il risultato è che a valori più distanti dal centro si assegna un maggior peso mentre ai valori più vicini si assegna minor peso

A ciò si aggiunge inoltre il vantaggio di poter utilizzare alcune importanti proprietà delle misure di centralità come la media aritmetica che, si ricorda, minimizza la somma degli scarti al quadrato

L'indice più utilizzato nella famiglia degli scostamenti quadratici è la cosiddetta **varianza**, indicata per convenzione con la lettera sigma dell'alfabeto greco -> σ^2

Quando tutti i valori della distribuzione sono uguali allora la varianza è nulla: infatti se tutte le unità del collettivo presentano lo stesso valore ciò indica che non c'è variabilità

La varianza non ha un massimo: più si allontana dallo 0 più il fenomeno è variabile

ATTENZIONE: poiché i valori nel calcolo della varianza sono elevati al quadrato, anche l'unità di misura sarà elevata al quadrato -> esempio: l'altezza media sarà espressa in *cm* come le singole altezze osservate sulle unità ma la varianza del carattere altezza sarà espressa in cm^2

15 – Varianza per distribuzioni unitarie

Consideriamo la distribuzione unitaria del carattere quantitativo X

u	X
u_1	x_1
...	...
u_i	x_i
...	...
u_N	x_N

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}_a\right)^2$$

La **varianza** è ottenuta come media degli scarti dalla media della distribuzione al quadrato

Il numeratore è chiamato **devianza** ed è utilizzato in alcuni casi come indice di variabilità al posto della varianza

Tale indice, con gli opportuni accorgimenti già visti per le medie analitiche, può essere utilizzato per tutti i tipi di distribuzione dei caratteri quantitativi

16 – Esempio

Si consideri la distribuzione unitaria della spesa per pubblicità nel 2008 delle aziende del comparto manifatturiero aventi sede operativa nella provincia di Cosenza (in migliaia di €)

Spesa per Pubblicità (migliaia di €)	76	112	78	63	115	82	61	73	65	54

Calcolare la spesa media e studiare la variabilità del fenomeno

$$\bar{x}_a = \frac{76+112+78+63+115+82+61+73+65+54}{10} = 77,9$$

$$\sigma^2 = \frac{\begin{array}{c}(76-77,9)^2+(112-77,9)^2+(78-77,9)^2+(63-77,9)^2+(115-77,9)^2+\\ +(82-77,9)^2+(61-77,9)^2+(73-77,9)^2+(65-77,9)^2+(54-77,9)^2\end{array}}{10} = 382,89$$

Mediamente nel 2008 le aziende del comparto manifatturiero hanno speso 77900 € per pubblicità, con una varianza pari a 382890 (possiamo omettere l'unità di misura poiché è priva di significato)

17 – Varianza per distribuzioni di frequenza

Per le distribuzioni di frequenza è necessario tenere in considerazione sia le modalità del carattere che il numero di unità statistica sulle quali ciascuna modalità è stata osservata

frequenze assolute

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}[(x_i - \overline{x}_a)^2 \cdot n_i]$$

frequenze relative

$$\sigma^2 = \sum_{i=1}^{k}[(x_i - \overline{x}_a)^2 \cdot f_i]$$

Per le distribuzioni in classi è sufficiente sostituire alle modalità il valore centrale di ciascuna classe: come già detto, il valore ottenuto non è che una approssimazione di quello realmente osservabile a partire dalla distribuzione unitaria

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}[(c_i - \overline{x}_a)^2 \cdot n_i]$$

18 – Esempio

Consideriamo la distribuzione del numero di impiegati per anni di servizio in una industria

Anni di servizio	n. impiegati	vc
0 -| 1	7	0.5
1 -| 5	18	3.0
5 -| 10	45	7.5
10 -| 20	25	15.0
20 -| 30	20	25.0
Totale	115	

Mediamente ogni impiegato lavora in questa azienda 11.04 anni

Studiamo ora la variabilità della distribuzione degli anni di servizio calcolando la varianza:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(c_i - \bar{x}_a)^2 \cdot n_i = \frac{1}{115}[(0.5-11.04)^2 \cdot 7 + (3.0-11.04)^2 \cdot 18 + (7.5-11.04)^2 \cdot 45 + (15.0-11.04)^2 \cdot 25 + (25.0-11.04)^2 \cdot 20]$$

$$\sigma^2 = 59.08$$

19 – Proprietà della varianza

La varianza presenta alcune peculiarità

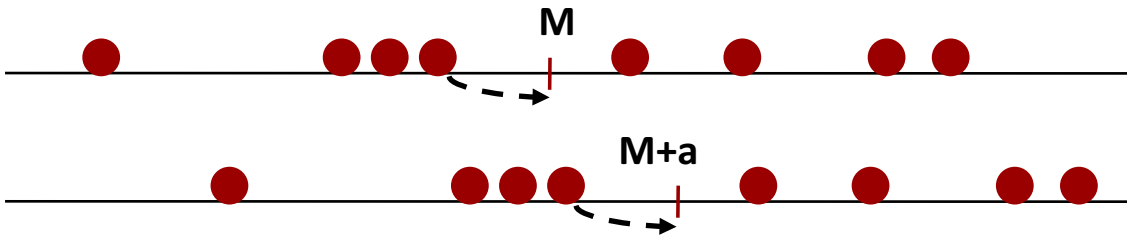
1) dipende da tutte le modalità del carattere

2) assume solo valori non negativi: $\sigma^2 > 0$

3) è nulla se e solo se il carattere è "degenere": **se $x_1 = x_2 = \ldots = x_N$ allora $\sigma^2 = 0$**

4) è sensibile ai valori anomali (poiché lo è la media aritmetica in essa contenuta)

5) è invariante per traslazioni del tipo Y = X + a: infatti si dimostra che $\sigma_Y^2 = \sigma_X^2$



Se spostiamo tutti i valori a destra gli scostamenti rispetto alla media non si modificano

non è più invariante per trasformazioni del tipo Y = bX: si ha che $\sigma_Y^2 = b^2 \sigma_X^2$

20 – Formula alternativa per il calcolo della varianza

È possibile calcolare la varianza anche attraverso l'utilizzo di una formula alternativa

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}_a\right)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \overline{x}_a^{\,2}$$

La varianza è data dalla differenza tra la media quadratica (media dei valori della distribuzione al quadrato) e il quadrato della media aritmetica

N.B.: se i dati sono rappresentati con una distribuzione di frequenza si deve tener conto di questa informazione!

ES. Distribuzione del voto all'esame

MODO A

x_i	$x_i - x$	$(x_i - x)^2$
30	2.5	6.25
19	-8.5	72.25
24	-3.5	12.25
30	2.5	6.25
30	2.5	6.25
30	2.5	6.25
24	-3.5	12.25
30	2.5	6.25
28	0.5	0.25
30	2.5	6.25

MODO B

x_i	x_i^2
30	900
19	361
24	576
30	900
30	900
30	900
24	576
30	900
28	784
30	900

MODO A

$$\sigma^2 = \frac{1}{10}\sum_{i=1}^{10}\left(x_i - 27.5\right)^2 = 13.45$$

MODO B

$$\sigma^2 = \frac{1}{10}\sum_{i=1}^{10}\left(x_i^{\,2}\right) - 756.25 = 13.45$$

21 – Variabilità di variabili doppie miste e quantitative — Unità n° 04

	y_1	...	y_j	...	y_c	TOT
x_1	n_{11}	...	n_{1j}	...	n_{1c}	$n_{1.}$
...	...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	...	n_{ic}	$n_{i.}$
...	...	...	...	...	...	...
x_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r.}$
TOT	$n_{.1}$	...	$n_{.j}$	...	$n_{.c}$	$n_{..}$

Consideriamo una variabile doppia **(X,Y)** e supponiamo che sia stata organizzata in una tabella che contiene sulle righe le r modalità di X e sulle colonne le c modalità di Y

Supponiamo che la variabile **(X,Y)** sia mista o quantitativa (cioè che almeno una delle due variabili in gioco sia quantitativa)

Se entrambe le variabili sono quantitative è possibile calcolarne la varianza

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{r}\left(x_i - \bar{x}_a\right)^2 \cdot n_{i.} = \frac{1}{n}\sum_{i=1}^{r} x_i^2 \cdot n_{i.} - \bar{x}_a^2$$

varianza generale di X

$$\sigma_Y^2 = \frac{1}{n}\sum_{j=1}^{c}\left(y_j - \bar{y}_a\right)^2 \cdot n_{.j} = \frac{1}{n}\sum_{i=1}^{c} y_j^2 \cdot n_{.j} - \bar{y}_a^2$$

varianza generale di Y

22 – Esempio

Consideriamo la distribuzione doppia di frequenza del ***tipo di birra*** preferito e dell'***età*** di un collettivo di consumatori: ovviamente possiamo calcolare la varianza solo della variabile in colonna

		Età			
		18 - 22	***23 - 26***	***27 - 30***	
Tipo di Birra	*Bionda*	12	22	11	**45**
	Rossa	5	9	14	**28**
	Scura	3	15	18	**36**
		20	**46**	**43**	**109**

v.c.	20	24,5	28,5	
$n_{.j}$	20	46	43	109

Individuiamo i valori centrali delle diverse classi di età e consideriamo l'età media di 25,25 anni

$$\overline{y}_a^2 = \frac{20.0^2 \cdot 20 + 24.5^2 \cdot 46 + 28.5^2 \cdot 43}{109} = 647.14$$

$$\sigma_Y^2 = 647.14 - 25.25^2 = 9.57$$

I consumatori del collettivo che stiamo esaminando hanno una età media di 25.25 anni e una varianza di 9.57 (anni^2)

23 – Varianze condizionate

Unità n° 04

Con lo stesso criterio già visto è possibile calcolare anche la varianza della distribuzione condizionata **Y|x** : siamo interessati a studiare la variabilità nel collettivo del carattere misurato dalla variabile Y fissato un certo valore della variabile X

$$\sigma^2_{Y|x_i} = \frac{1}{n_{i.}} \sum_{j=1}^{c} \left[y_j - M(Y|x_i) \right]^2 \cdot n_{ij}$$

$$\sigma^2_{X|y_j} = \frac{1}{n_{.j}} \sum_{i=1}^{r} \left[x_i - M(X|y_j) \right]^2 \cdot n_{ij}$$

VARIANZA CONDIZIONATA Y|x

$$\sigma^2_{Y|x_i} = \frac{1}{n_{i.}} \sum_{j=1}^{c} \left[y_j - M(Y|x_i) \right]^2 \cdot n_{ij}$$

numerosità del gruppo definito dalla modalità x_i

media di Y nel gruppo definito dalla modalità x_i

n. di unità del gruppo che presentano le diverse modalità di Y

Anche in questo caso è possibile considerare la deviazione standard come indice di variabilità espresso nell'unità di misura dei dati

24 – Il problema dell'unità di misura

Per ovviare al problema del cambio di unità di misura è possibile utilizzare un altro indice di variabilità, detto **scarto quadratico medio** (o deviazione standard)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k}\left(x_i - \bar{x}_a\right)^2}$$

È praticamente la radice quadrata della varianza

Lo s.q.m. ci dice di quanto in media i valori della distribuzione si discostano dal valore rappresentativo dell'intera distribuzione, cioè la media

Qual è il senso di elevare al quadrato tutte gli scostamenti se poi dobbiamo considerare la radice quadrata della varianza?

L'elevazione al quadrato delle differenze tra i valori della distribuzione e la media trova di fatto due ragioni: la prima deriva dalle proprietà della media aritmetica, poiché la somma degli scarti dalla media è nullo, quindi è necessario utilizzare i quadrati; la seconda è invece data dal fatto che elevando al quadrato possiamo attenuare l'importanza delle modalità più vicine al valore medio e contemporaneamente dare il giusto peso a quelle più lontane, come se utilizzassimo una lente di ingrandimento

25 – Esempio

Analizziamo il n° di imprese dolciarie operanti nel 1991 in cinque regioni italiane

Regioni	N. Imprese
Piemonte	268
Marche	106
Abruzzo	76
Campania	238
Calabria	88

$$\overline{x}_a = 155.2 \text{ (Imprese)}$$

$$\sigma^2 = \frac{1}{5}\sum_{i=1}^{5}\left(x_i - 155.2\right)^2 = 6557.76$$

$$\sigma = \sqrt{6557.76} = 80.98 \text{ (Imprese)}$$

La deviazione standard ci dice di quanto in media i valori della distribuzione si discostano dal valore centrale: in media nelle regioni analizzate sono state create 155 ± 81 imprese

26 – Potere informativo della media

Abbiamo visto come la varianza presenti talvolta un problema di lettura: poiché per costruzione tutti i valori sono elevati al quadrato avremo un indice non espresso nella stessa unità di misura dei valori della distribuzione

Per ovviare a questo inconveniente lo scarto quadratico medio, ossia la radice quadrata della varianza, è molto utilizzato: ha di fatto uguale contenuto informativo della varianza ma allo stesso tempo è espressa nell'unità di misura dei dati. Possiamo inoltre interpretarla come lo scostamento (quadratico) medio dal centro della distribuzione, in questo caso rappresentato dalla media

Proprio per tale motivo possiamo valutare attraverso lo s.q.m il potere informativo della media, cioè quanto è effettivamente un buon indicatore sintetico della distribuzione del carattere nel collettivo: a valori bassi dell'indice corrisponde una media che meglio sintetizza la distribuzione, mentre a valori alti dell'indice corrisponde una media con un minor potere informativo, perché in tal caso i valori della distribuzione si allontanano mediamente di più dal centro della distribuzione

27 – Deviazione Standard, ampiezza e numero delle classi

E' possibile utilizzare l'informazione data dalla deviazione standard anche per determinare l'ampiezza (e quindi il numero) degli intervalli di modalità in una distribuzione in classi:

Formula di Scott

$$\omega \cong \frac{3{,}5 \cdot \sigma}{\sqrt{N}}$$

ES.: Se abbiamo un collettivo di 200 famiglie e vogliamo rappresentare in classi il carattere spesa mensile per trasporti, noto che in media la spesa di ciascuna famiglia si discosta dalla spesa media di ±23,5€, avremo una ampiezza di circa 6€ per classe (ovviamente se le classi sono equiampie)

Ovviamente come per la formula di Sturges abbiamo una indicazione che però spesso non può prescindere o sostituire il buon senso e l'esperienza del ricercatore!

28 – Indici relativi di variabilità

Sia la varianza che lo scarto quadratico medio sono indici assoluti di variabilità

Questo aspetto fa sì che tali indici non possano essere utilizzati per effettuare confronti tra:

1) più collettivi sui quali si manifesta uno stesso fenomeno, con un diverso ordine di grandezza
2) più fenomeni, espressi anche con diversa unità di misura

Per poter confrontare la variabilità di fenomeni con differente unità di misura o con un diverso ordine di grandezza si può far ricorso ad **indici relativi**. Possono essere costruiti in due modi:

- rapportando l'indice di variabilità assoluto ad una media
- rapportando l'indice di variabilità assoluto al massimo valore che può assumere

Per costruzione gli indici relativi si presentano come il rapporto tra due quantità espresse nella stessa unità di misura, quindi il valore numerico che ne scaturisce è un numero puro, dal quale è stata cioè eliminata l'influenza esercitata dall'unita di misura e dall'ordine di grandezza

29 – Il coefficiente di variazione

Per confrontare la variabilità di due distribuzioni può essere utilizzato il cosiddetto **coefficiente di variazione**, costruito come rapporto tra lo scarto quadratico medio e la media in valore assoluto

$$CV = \frac{\sigma}{|\bar{x}_a|} \left(x\ 100\right)$$

→ Il CV è generalmente è espresso in termini percentuali

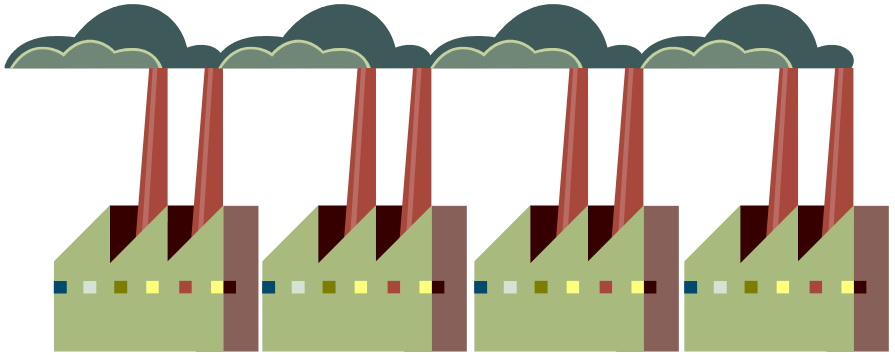
Per quanto già evidenziato, il rapporto tra **σ** e $\bar{x}_a$ da come risultato un valore che non è espresso in nessuna unità di misura (adimensionale)

Il rapportare la deviazione standard alla media consente di eliminare l'influenza esercitata sulla variabilità dalla magnitudine del fenomeno, fornendo la media un'utile indicazione dell'ordine di grandezza del fenomeno

Può essere letto come *"mediamente gli scostamenti dal valore medio del carattere sono pari al …% di quest'ultimo"*

30 – Esempio

Consideriamo 9 industrie nelle quali è installato un dispositivo anti-inquinante di tipo A, e altre 9 con un dispositivo di tipo B: vogliamo confrontare le due distribuzioni della quantità di pulviscolo (g. per minuto) prodotto dalle diverse industrie



Quantità di pulviscolo	
Tipo A	Tipo B
69	35
80	62
44	43
52	23
54	30
54	28
86	22
77	40
66	25

Tipo A

$\bar{x}_A = 64.67$

$\sigma_A = 13.65$

$CV_A = 21\%$

Tipo B

$\bar{x}_B = 34.22$

$\sigma_B = 12.02$

$CV_B = 35\%$

Dal confronto del CV calcolato sulle due distribuzioni si può concludere che la distribuzione della quantità di pulviscolo prodotta da chi installa B è più variabile della distribuzione della quantità di pulviscolo prodotta da chi installa A

31 – Centralità, variabilità e forma

Una volta studiata la distribuzione attraverso il calcolo delle misure di centralità e variabilità abbiamo delle informazioni sintetiche per poter comprendere il comportamento di un certo fenomeno rispetto al collettivo oggetto di studio

La centralità e la variabilità di una distribuzione non esauriscono le informazioni contenute nei dati, in alcuni casi non sono esaustive per poter interpretare come il carattere si manifesta

Abbiamo bisogno quindi anche di un altro elemento per meglio definire le caratteristiche della distribuzione: due variabili possono avere infatti, ad esempio, la stessa media/mediana e la stessa variabilità ma differire per il peso dei valori più grandi o più piccoli rispetto al valore centrale, a causa del comportamento differenziato delle “code” della distribuzione, cioè delle parti più esterne dell’insieme ordinato dei dati

Tale studio può essere effettuato considerando la cosiddetta **forma della distribuzione**

Tale argomento meriterebbe una trattazione separata, ma per semplicità lo consideriamo nell’ambito della variabilità, essendo ad essa strettamente collegato

32 – Gli intervalli di variabilità

Data la distribuzione unitaria di un carattere X, ordinata in senso crescente

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$$

è possibile considerare 5 valori rappresentativi:

$x_{(1)} = x_{min}$ è il valore più piccolo della distribuzione

Q_1 = primo quartile (25° percentile)

Me = mediana (50° percentile)

Q_3 = terzo quartile (75° percentile)

$x_{(N)} = x_{max}$ è il valore più grande della distribuzione

Attraverso tali valori possiamo costruire i cosiddetti **intervalli di variabilità** della distribuzione

33 – Centralità e Variabilità

Da $|x_{min}|\ Q_1\ |\ M_e\ |\ Q_3\ |x_{max}|$ è possibile ottenere due misure di posizione e due di variabilità

posizione

$$\text{MidRange} = \frac{x_{min} + x_{max}}{2}$$

$$\text{Media Interquartile} = \frac{Q_1 + Q_3}{2}$$

variabilità

$$\text{Campo di var.ne} = x_{max} - x_{min}$$

$$\text{Differenza Interquartile} = Q_3 - Q_1$$

Le misure forniscono delle indicazioni di massima sulla distribuzione dei dati ma sono influenzate dai valori anomali o considerano solo il 50% dei dati a disposizione: possiamo comunque utilizzare tali quantità per analizzare la forma della distribuzione di X

34 – La sintesi a cinque

Utilizzando i cinque valori rappresentativi $|x_{min}|\ Q_1\ |\ M_e\ |\ Q_3\ |x_{max}|$ è possibile studiare il comportamento di un carattere in un collettivo, osservando:

- la distanza tra il primo quartile e la mediana e tra la mediana e il terzo quartile
- la distanza tra x_{min} e il primo quartile e tra il terzo quartile e x_{max}
- la relazione tra la mediana, la media interquartile e il midrange

La distribuzione si dice **simmetrica** se:
- la distanza tra primo quartile e mediana e tra mediana e terzo quartile è uguale
- la distanza tra x_{min} e primo quartile e tra terzo quartile e x_{max} è uguale
- la mediana, la media interquartile e il midrange coincidono

In questo caso anche la moda e la media aritmetica coincidono con la mediana

La distribuzione si dice **asimmetrica** se:
- la distanza tra primo quartile e mediana e tra mediana e terzo quartile è diversa
- la distanza tra x_{min} e primo quartile e tra terzo quartile e x_{max} è diversa
- la mediana, la media interquartile e il midrange non coincidono

35 – Asimmetria positiva e negativa

In generale si distingue tra una asimmetria **positiva** e una asimmetria **negativa**

La distribuzione si dice **asimmetrica negativa** (o "obliqua a sinistra") se:
- la distanza tra x_{min} e primo quartile è maggiore di quella tra terzo quartile e x_{max}
- la mediana è maggiore della media interquartile, la media interquartile è maggiore del midrange

→ In questo caso si osservano più frequentemente modalità con *alta* intensità e più raramente modalità con *bassa* intensità, quindi in generale (ma non sempre): moda>mediana>media

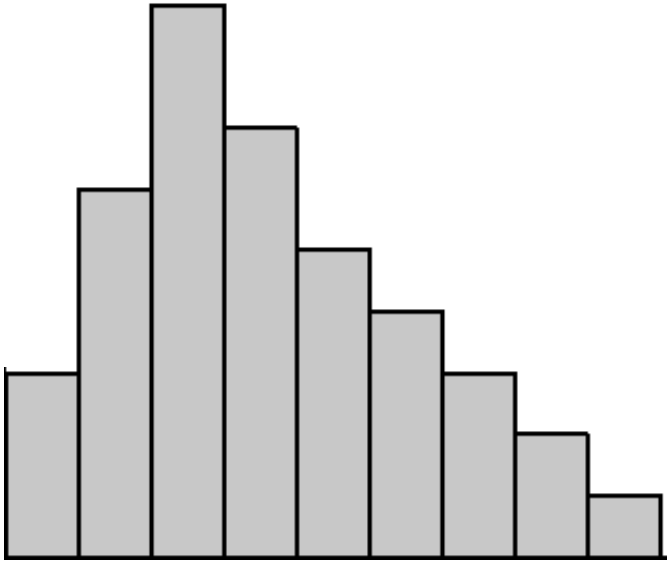
La distribuzione si dice **asimmetrica positiva** (o "obliqua a destra") se:
- la distanza tra x_{min} e primo quartile è minore di quella tra terzo quartile e x_{max}
- la mediana è minore della media interquartile, la media interquartile è minore del midrange

→ In questo caso si osservano più frequentemente modalità con *bassa* intensità e più raramente modalità con *bassa* intensità, quindi in generale (ma non sempre): moda<mediana<media

36 – Rappresentazione grafica

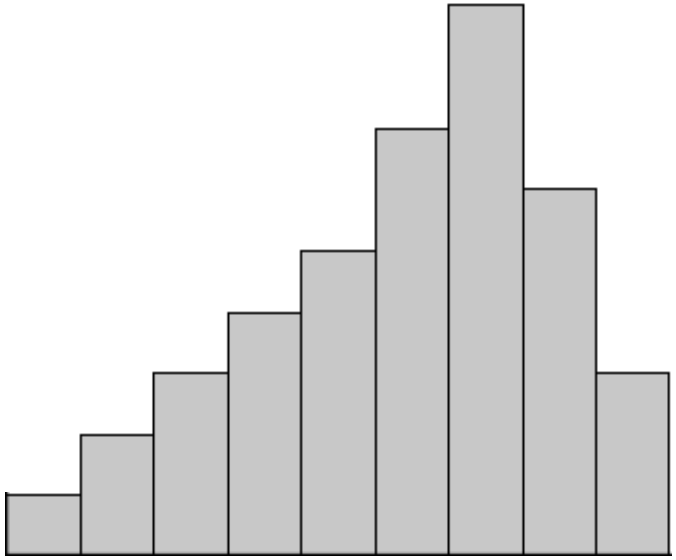
Unità n° 04

Possiamo studiare la forma di una distribuzione di frequenza o in classi osservando il corrispondente diagramma a barre o istogramma



Distribuzione asimmetrica positiva

i valori più piccoli sono più frequenti e la moda è minore del centro della distribuzione: abbiamo valori alti che "disturbano" la distribuzione



Distribuzione asimmetrica negativa

i valori più grandi sono più frequenti e la moda è maggiore del centro della distribuzione:abbiamo valori bassi che "disturbano" la distribuzione

37 – Un diverso modo di studiare la forma della distribuzione — Unità n° 04

Possiamo ricorrere ai soli intervalli di variabilità per descrivere graficamente la distribuzione

La rappresentazione ottenuta è detta **box plot** (diagramma a scatola e baffi)

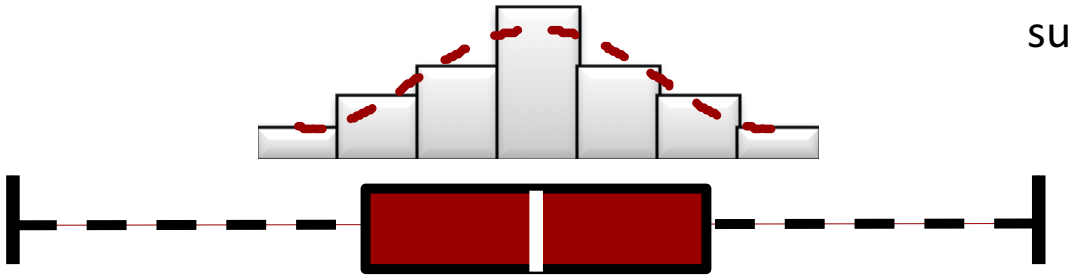
Il box-plot è un grafico caratterizzato da tre elementi:

1) un rettangolo (box) la cui dimensione indica la variabilità dei valori "prossimi" al centro della distribuzione
2) una linea o punto, che indica la posizione del centro della distribuzione
3) due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione

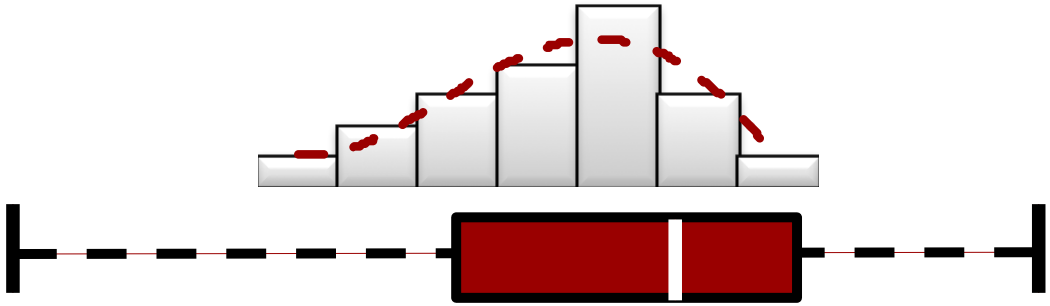
→ Generalmente come valore centrale si considera la mediana, come altezza/larghezza della scatola la distanza interquartile e come estremi dei segmenti il valore minimo e massimo della distribuzione

38 – Box plot e forma della distribuzione

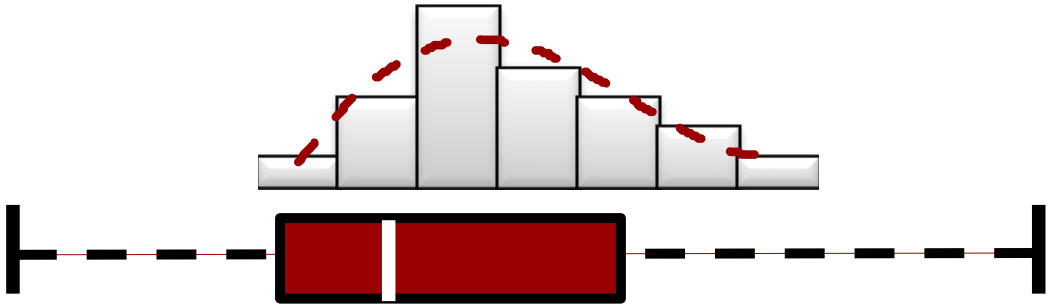
Dal Box plot possiamo dedurre informazioni su variabilità e forma della distribuzione di X



Distribuzione simmetrica



Distribuzione Asimmetrica negativa

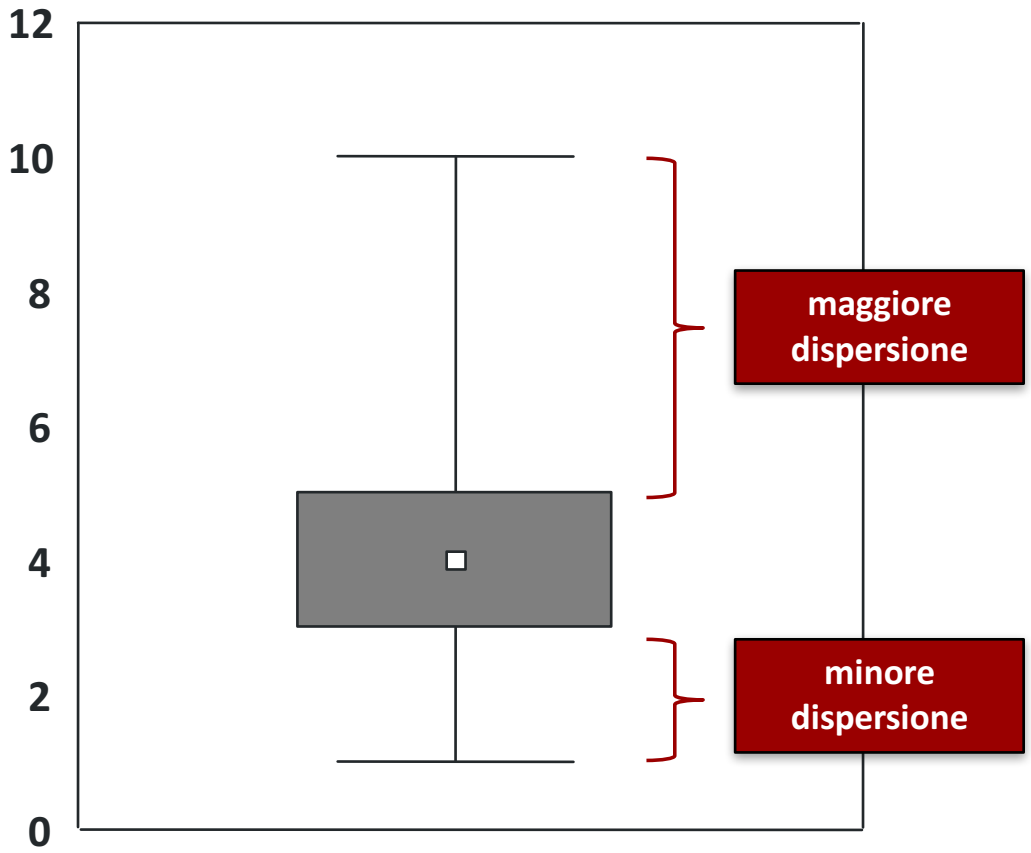


Distribuzione Asimmetrica positiva

39 – Esempio

N°atti aggressivi	1	2	3	4	5	6	7	8	9	10
Bambini	3	8	30	45	22	12	10	5	2	1

Studio sull'aggressività infantile (138 bambini)



Max = 10
Min = 1

Q_3=5
Q_1=3

Valore mediano:
Me=4

Dall'analisi del box plot si evince come ci sia una maggior frequenza di valori medio-bassi, il che spiega lo spostamento verso il basso della scatola (o verso sinistra se consideriamo un box plot orizzontale)

40 – Box plot e valori anomali

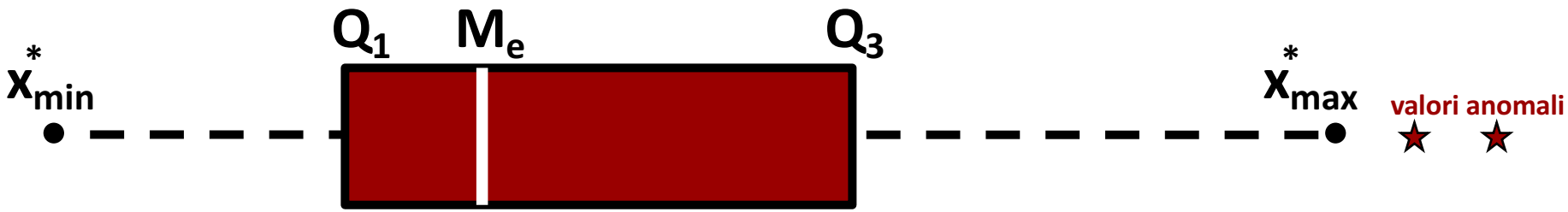
Attraverso il box plot è possibile evidenziare la presenza di eventuali valori anomali. Abbiamo già detto che un valore anomalo è un valore molto più piccolo o molto più grande rispetto ai valori della distribuzione: per poter evidenziare tali modalità particolari è necessario calcolare i cosiddetti valori minimo e massimo "teorici" e confrontarli con quelli effettivamente osservati

E' possibile considerare come minimo e massimo della distribuzione i valori così ottenuti:

x^*_{min} => valore più grande tra x_{min} e $[Q_1-1.5(Q_3-Q_1)]$

x^*_{max} => valore più piccolo tra x_{max} e $[Q_3+1.5(Q_3-Q_1)]$

Gli eventuali valori esterni a tali valori sono considerati anomali



41 – Esempio

Consideriamo la distribuzione dell'età degli studenti iscritti a un Master post-laurea

Età	Studenti

Età studenti del Corso	Frequenze assolute (n_i)
25	3
26	6
27	8
28	5
29	5
30	3
61	1
	31

$$M(X) = \frac{(25\times 3)+(26\times 6)+(27\times 8)+(28\times 5)+(29\times 5)+(30\times 3)+(61\times 1)}{31}$$

$$= \frac{75+156+216+140+145+90+61}{31} = \frac{883}{31} = 28,48$$

Mediana e Quartili

3 6 8 5 5 3 1

25 25 25 | 26 26 26 26 26 26 | 27 27 27 27 27 27 27 27 | 28 28 28 28 28 | 29 29 29 29 29 | 30 30 30 | 61

Q_1 Q_2=Me Q_3

$Q_1 = 26$

$Me = 27$

$Q_3 = 29$

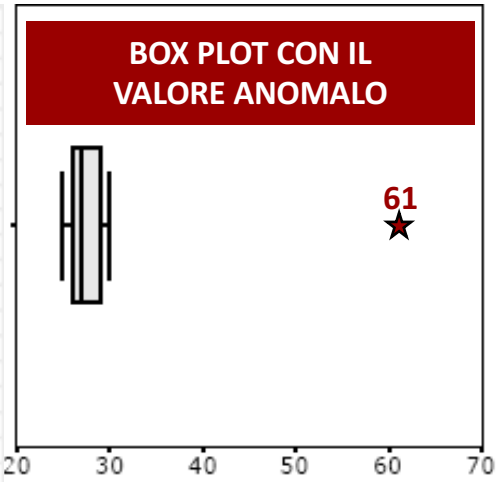
$Q_3 - Q_1 = 29-26 = 3$

Dati anomali:

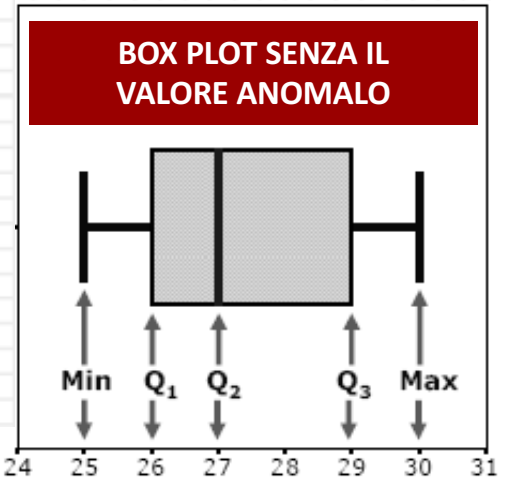
a) $Q_3 + 1,5\,(Q_3-Q_1) = 29 + 1,5\times 3 = 33,5$

b) $Q_1 - 1,5\,(Q_3-Q_1) = 26 - 1,5\times 3 = 21,5$

BOX PLOT CON IL VALORE ANOMALO

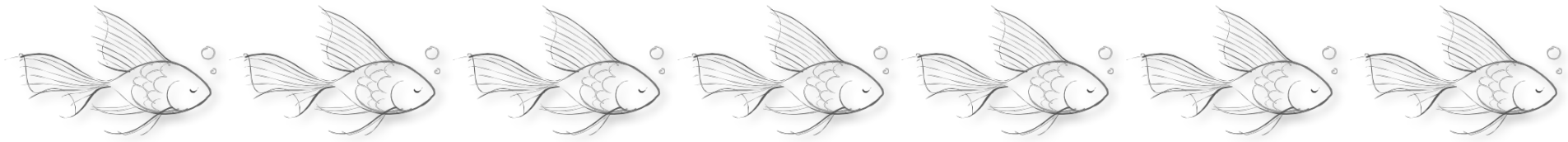


BOX PLOT SENZA IL VALORE ANOMALO



Dal box plot si rileva che 61 è un valore anomalo!

42 – Esercizio



Si consideri la distribuzione del peso di 10 maschi e 10 femmine (in Kg) di una particolare specie di pesce

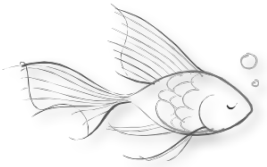
M	1.2	3.0	5.2	4.0	3.5	4.3	3.3	4.8	3.8	3.2
F	1.3	2.2	1.5	2.3	1.8	1.7	2.1	2.0	1.9	2.1

1) Calcolare per ciascuna sottopopolazione il peso medio e la deviazione standard

2) Confrontare la variabilità del peso di maschi e femmine con il coefficiente di variazione

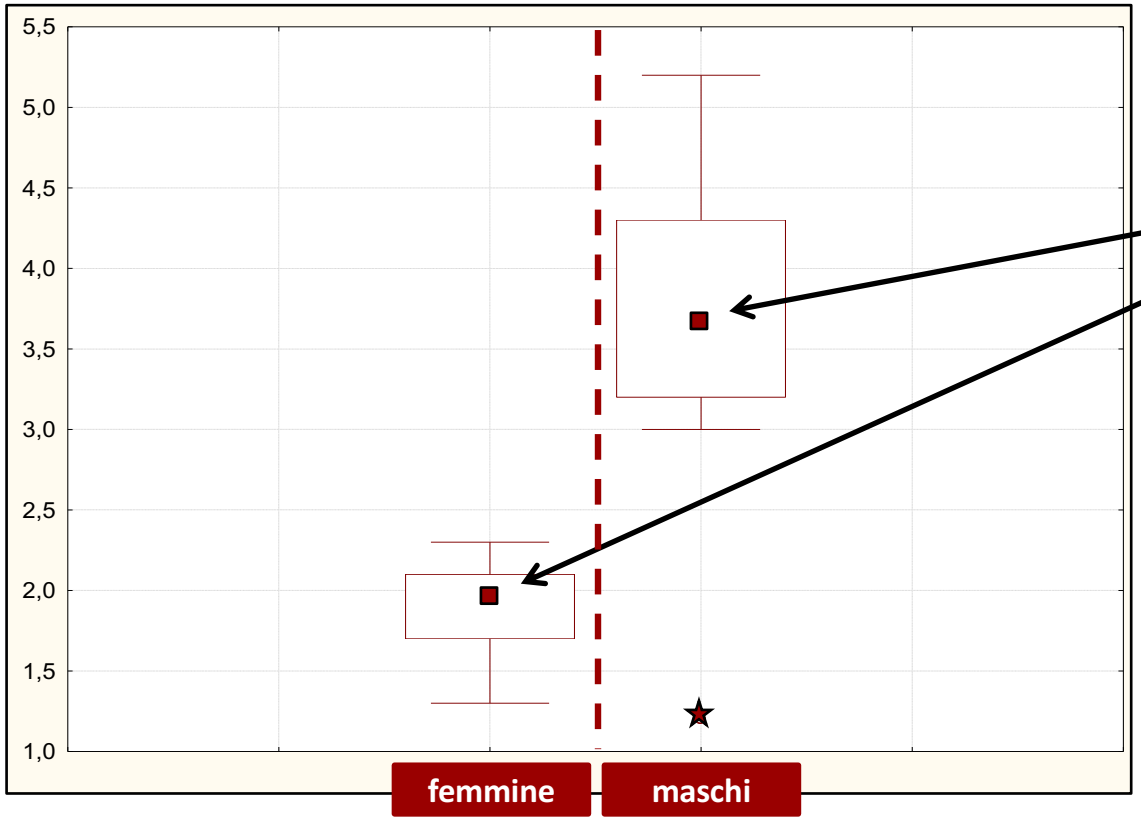
3) Costruire e commentare le rappresentazioni box plot

43 – Soluzione (2) e (3)



La variabilità del peso è maggiore nei maschi rispetto alle femmine

$CV_f = 17\%$
$CV_m = 31\%$



In media gli esemplari maschi pesano più degli esemplari femmine

Osserviamo come ci sia una maggiore dispersione nel peso degli esemplari maschi rispetto agli esemplari femmine. Rispetto alla forma delle diverse distribuzioni si vede come nel caso delle femmine ci sia una lieve asimmetria negativa, mentre nel caso dei maschi l'asimmetria è positiva