Corso di Laurea in Economia Aziendale

Docente: Marta Nai Ruscone

Statistica

a.a. 2015/2016

Indici di posizione

- **GLI INDICI DI POSIZIONE**

sono indici sintetici che evidenziano le caratteristiche essenziali della distribuzione del carattere

Qual è il voto medio riportato nella prova intermedia di Statistica dagli studenti del II anno?

Attraverso gli indici di posizione è possibile confrontare statistiche che rappresentano

i livelli/valori tipici di due diverse distribuzioni

Hanno riportato voti più alti le femmine o i maschi?

- **Consentono di sintetizzare un insieme di misure tramite un unico valore "rappresentativo" indice che riassume o descrive i dati e dipende dalla scala di misura dei dati in oggetto**

Indici di posizione

ALCUNI INDICI TIPICI

- **Moda** *(per tutti i tipi di carattere)*
- **Mediana** *(solo per caratteri ordinati)*
- **Quartili / percentili** *(solo per caratteri ordinati)*
- **Media** *(solo per i caratteri quantitativi)*

Ogni carattere statistico ha l'indice di posizione adeguato, e non tutti gli indici si possono calcolare per ogni carattere.

Passeremo quindi in rassegna i vari caratteri, individuando l'indice di posizione **adeguato**.

Indici di posizione

INDICI ADEGUATI

Qualitativo sconnesso -> MODA

Qualitativo ordinato -> MEDIANA

Quantitativo -> MEDIA

Indici di posizione

ALCUNI INDICI TIPICI

- **Moda** *(per tutti i tipi di carattere)*

"modalità a cui è associata la max f_i (o *max* d_i)"

Funzione di excel: "*moda*"

- **Mediana/ Percentili** *(solo per caratteri ordinati)*

"modalità che occupa la posizione *centrale* nella sequenza *ordinata* dei dati"

Funzione di excel: "*mediana*" oppure "*percentili*"

- **Media** *(solo per i caratteri quantitativi)*

$$M(X) = \overline{X} = \frac{\sum_{i=1}^{k} X_i f_i}{n}$$

Funzione di excel: "*media*"

Carattere qualitativo sconnesso

- **MODA**: **modalità di massima frequenza**

(**N.B.**: SI PUO' CALCOLARE PER OGNI CARATTERE, anche se di fatto viene calcolata solo per i caratteri qualitativi sconnessi o nominali, in quanto per altri caratteri si possono calcolare altri indici più informativi)

Mo(X) = modalità con massimo valore di f_i

E' l'unico indice di tendenza centrale per i dati qualitativi misurati su scala nominale
Indice descrittivo poco informativo.

Carattere qualitativo sconnesso

- **Attenzione …….**

la moda è la modalità cui è associata la frequenza (o densità di frequenza nel caso di caratteri quantitativi in classi) massima e non il valore massimo!!!

- Data la seguente distribuzione della variabile X

{8,1,1,2,4}

la moda non è 8 (la modalità con valore massimo) ma è 1 (cioè la modalità cui è associata la frequenza massima)

in questo caso la modalità 1 ha frequenza 2 al contrario di 2,4,8 che hanno frequenza 1.

Carattere qualitativo sconnesso

v.s. Regione di residenza

Regione	fi	fi%
Lombardia	9	30
Piemonte	6	20
Liguria	3	10
Valle d'Aosta	3	10
Friuli Venezia Giulia	3	10
Veneto	6	20
	30	100

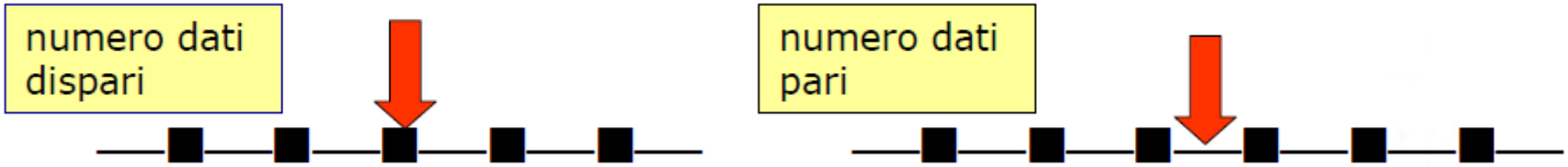
L'unico indice di posizione che si può calcolare è la moda

max fi=9 Mo=Lombardia

Carattere qualitativo ordinale

- **MEDIANA**: modalità/valore che occupa la posizione centrale o mediana (Pos_{Me}) nella distribuzione ordinata dei dati

  - preceduta da almeno 50% dei casi
  - superata da almeno 50% dei casi



Carattere qualitativo ordinale

- La posizione della mediana:

Posizione mediana = $\frac{n+1}{2}$ posizione della sequenza ordinata

- Se il numero di valori è dispari, la mediana è il valore centrale
- Se il numero di valori è pari, la mediana è la media dei due valori centrali

Nota che $\frac{n+1}{2}$ non è il *valore* della mediana, ma la *posizione d*ella mediana nella sequenza ordinata

Carattere qualitativo ordinale

- **MEDIANA**

E' l'indice di tendenza centrale,
insieme alla moda, per i dati
qualitativi misurati
su scala ordinale

Carattere qualitativo ordinale

- **PERCENTILI**: modalità/valori che dividono la distribuzione di frequenza ordinata in più parti

Permettono di rispondere ad es. alle seguenti domande:

- Qual è il reddito familiare che divide il 25% dei più poveri dal restante 75% ?
- Qual è la soglia di reddito oltre cui sta la fascia dei più abbienti ?
- Quanti bambini di 6 anni pesano più di 25 kg?

Carattere qualitativo ordinale

- **Alcuni esempi sono:**

quartili ⇒ dividono in **4** parti la distribuzione

decili ⇒ dividono in **10** parti la distribuzione

percentili ⇒ dividono in **100** parti la distribuzione

Carattere qualitativo ordinale

- **Per i quartili:**

$x_{0.25} = Q_1 = 1°$ quartile
(lascia alla sua sinistra il 25% e alla sua destra il 75%)

$x_{0.50} = Q_2 = 2°$ quartile
(lascia alla sua sinistra il 50% e alla sua destra il 50%)

$x_{0.75} = Q_3 = 3°$ quartile
(lascia alla sua sinistra il 75% e alla sua destra il 25%)

Carattere qualitativo ordinale

- **Per i quartili:**

In generale: il **percentile x_p di ordine p** è quella modalità che è:

- preceduta da almeno p% dei casi
- superata da almeno (1–p)% dei casi

Carattere qualitativo ordinale

Dunque…

- QUARTILI: percentili di ordine 25 – 50 – 75
- DECILI: percentili di ordine 10 – 20 - … - 90
- PERCENTILI: percentili di ordine 1 – 2 - … - 99

Carattere qualitativo ordinale

v.s. Interesse per la politica

Interesse per la politica	fi	fi%	Fi	Fi%
Non molto interessante	9	30	9	30
Abbastanza interessante	12	40	21	70
Molto interessante	9	30	30	100
	30	100		

- Mo= Abbastanza interessante

  max fi=12

- Me=Abbastanza interessante
- Q1=Non molto interessante
- Q3=Molto interessante

L'unico indice di posizione che non si può calcolare è la media

Carattere quantitativo

- **MEDIA**: è data dalla somma delle misure osservate diviso il numero delle osservazioni fatte (totale dei casi)
- Si indica con M(X) o con $\bar{x}$ per i campioni
- Quando ci si riferisce alla popolazione si indica con μ

Carattere quantitativo

- **MEDIA**: è data dalla somma delle misure osservate diviso il numero delle osservazioni fatte (totale dei casi)

$$M(X), \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Σ= sommatoria

X_i= osservazione i-ma

***n*= numero osservazioni**

Carattere quantitativo discreto

v.s. Ore dedicate a News in Internet

Ore News Internet	fi	fi%	Fi	Fi%	Xifi
1	4	13,33	4	13,33	4
2	7	23,33	11	36,67	14
3	4	13,33	15	50,00	12
4	4	13,33	19	63,33	16
5	4	13,33	23	76,67	20
6	4	13,33	27	90,00	24
7	1	3,33	28	93,33	7
8	2	6,67	30	100,00	16
	30	100,00		somma	113
				media	3,766667

- Mo=2 MODA(K2:K31)
- Me=3,5 MEDIANA(K2:K31)
- Media=3,77 MEDIA(K2:K31)
- Q1=2 PERCENTILE(K2:K31;0,25)
- Q3=5 PERCENTILE(K2:K31;0,75)

Carattere quantitativo: dati raggruppati in classi

- **MEDIA per dati raggruppati in classi**

si moltiplica la frequenza di ogni classe per il valore definito dal valore centrale di ogni classe, prima di fare la somma e dividere per il numero dei casi.

Carattere quantitativo: dati raggruppati in classi

- **MEDIA per dati raggruppati in classi**

$$M(X), \overline{X} = \frac{\sum_{i=1}^{k} {}_{c}x_i f_i}{n}$$

${}_{c}X_i$ = valore centrale della classe i-ma
f_i = frequenza assoluta classe i-ma
k = numero di classi
n = numero totale di osservazioni

Carattere quantitativo: dati raggruppati in classi

- **Valore centrale di classe**

Data la classe $[h_{i-1}, h_i]$ il valore centrale si ottiene:

$$c x_i = \frac{h_{i-1} + h_i}{2}$$

Esempio: data la classe 3–|12

$$c x_i = \frac{3+12}{2} = 7.5$$

Carattere quantitativo: dati raggruppati in classi

- **NB**

Se avessimo voluto calcolare la moda quale sarebbe stata?

Essendo un carattere in classi, avremmo dovuto calcolare la DENSITA' di frequenza.

La **moda** è la classe con **max d_i**

Carattere quantitativo in classi

v.s. Età

Età	linf	lsup	Xc	fi	fi%	Fi	Fi%	Xc*fi
20-|30	20	30	25	10	33,33	10	33,33	250
30-|40	30	40	35	9	30,00	19	63,33	315
40-|50	40	50	45	3	10,00	22	73,33	135
50-|60	50	60	55	8	26,67	30	100,00	440
				30	100,00		somma	1140
							media	38

- Mo=25
- Me=36 posizione (n+1)/2
- Q1=28 posizione (n+1)*25/100
- Q3=52 posizioni (n+1)*75/100
- Media=38

N.B. Si ottengono valori differenti se gli indici vengono calcolati sui dati non in classi

Proprietà media aritmetica

1. La media aritmetica di una variabile è sempre compresa tra il valore minimo e il valore massimo assunti dalla variabile stessa, cioè

$$x_{\min} \leq \overline{x} \leq x_{\max}$$

2. La media di una costante è uguale alla costante stessa, inoltre se una variabile X viene moltiplicata per una costante anche la sua media risulta moltiplicata per la stessa costante, cioè

$$M(a + bX) = a + bM(X),$$

dove M si dice operatore media aritmetica

e a e b sono due costanti, vale quindi

$$M(a) = a$$

$$M(bX) = bM(X)$$

Carattere quantitativo: dati raggruppati in classi

RIASSUMENDO

- per caratteri qualitativi sconnessi si può calcolare solo la MODA
- per caratteri qualitativi ordinabili si possono calcolare la MODA e la MEDIANA
- per caratteri quantitativi discreti/continui si possono calcolare TUTTI gli indici (MODA, MEDIANA, MEDIA).

N.B: Nel caso di caratteri in classi la MODA e' la CLASSE cui e' associata la densita' di frequenza massima, e non la classe con frequenza massima!

Variabilità

LA STATISTICA (Trilussa)

Sai ched'è la statistica? E' 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa
Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che lì la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
E se nun entra ne le spese tue
t'entra ne la statistica lo stesso
perché c'e' un antro che ne magnia due!!!

Variabilità

Gli indici di posizioni sono utili per alcune informazioni sui caratteri

- appare tuttavia insufficiente
- sintesi troppo spinta $\longrightarrow$ perdita di informazioni

$$\Longrightarrow \textbf{POSIZIONE + VARIABILITÀ}$$

- interessano anche indicatori della diversità (molteplicità) dei valori di un carattere

Variabilità

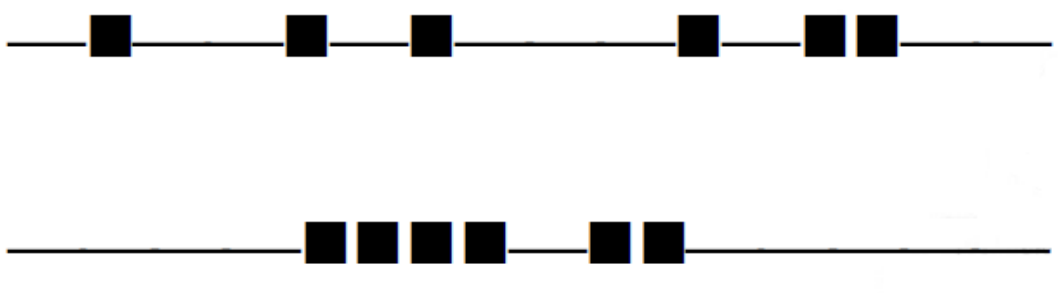
Attitudine del carattere ad assumere modalità differenti

Variabilità

Per capire

$X = \{x_1, \ldots, x_6\}$

$Y = \{y_1, \ldots, y_6\}$



- è più variabile (disperso) X oppure Y??

Variabilità

Indici di dispersione:

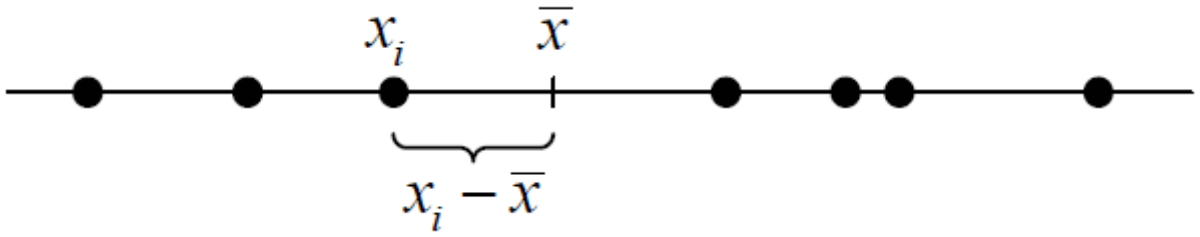
- VARIANZA
- SCARTO QUADRATICO MEDIO o DEVIAZIONE STANDARD
- COEFFICIENTE DI VARIAZIONE
- CAMPO DI VARIAZIONE
- DIFFERENZA INTERQUARTILE

Variabilità

Indici di dispersione:

- Si può ottenere un indice di dispersione che tenga conto del contributo dei singoli casi:

  a) si calcolano gli scarti dei valori osservati dalla media



  b) si fa una media di questi scarti

Varianza della popolazione

- La **VARIANZA** è la media degli scarti da M al quadrato

  (Si considerano gli scostamenti al quadrato per evitare compensazioni tra distanze positive e negative.)

$$\sigma^2 = Var(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

FORMULA DI DEFINIZIONE

Varianza della popolazione –
formula operativa

- Negli esercizi si utilizza solitamente una formula più semplice per il calcolo della varianza.

  E' possibile dimostrare che:

$$\sigma^2 = Var(X) = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2 =$$

$$= M\left(X^2\right) - \bar{X}^2$$

Varianza della popolazione –
formula operativa

$$\sigma^2 = \frac{\sum_{i=1}^{k} X_i^2 f_i}{n} - \overline{X}^2$$

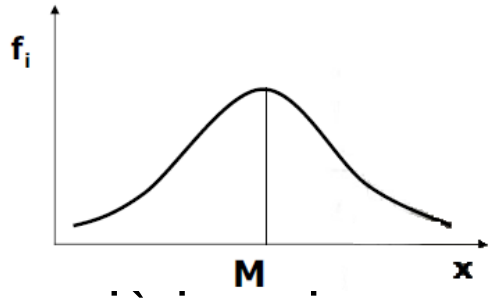
→ **Distribuzioni di frequenza**

$$\sigma^2 = \frac{\sum_{i=1}^{k} {}_c X_i^2 f_i}{n} - \overline{X}^2$$

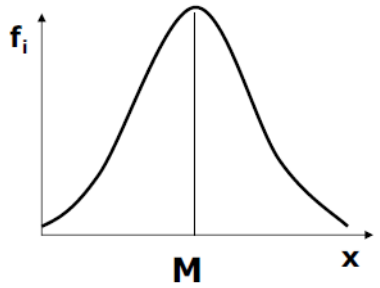
→ **Dati in classi**

Proprietà Varianza

- La varianza non è mai negativa
- Maggiore è la varianza più i casi sono dispersi attorno alla media



- Minore è la varianza più i casi sono concentrati attorno alla media



Proprietà Varianza

1. La varianza di una costante è uguale a 0, cioè

$$V(a) = 0$$

2. E' invariante per translazione, cioè se ad ogni x_i viene aggiunta una quantità a costante, la varianza non cambia, cioè

$$V(X + a) = \sigma_x^2$$

3. Se ogni x_i viene moltiplicata per una quantità b costante, la varianza risulta moltiplicata per la costante b al quadrato, cioè

$$V(bX) = b^2 V(X)$$

IN SINTESI (varianza di una trasformazione lineare)

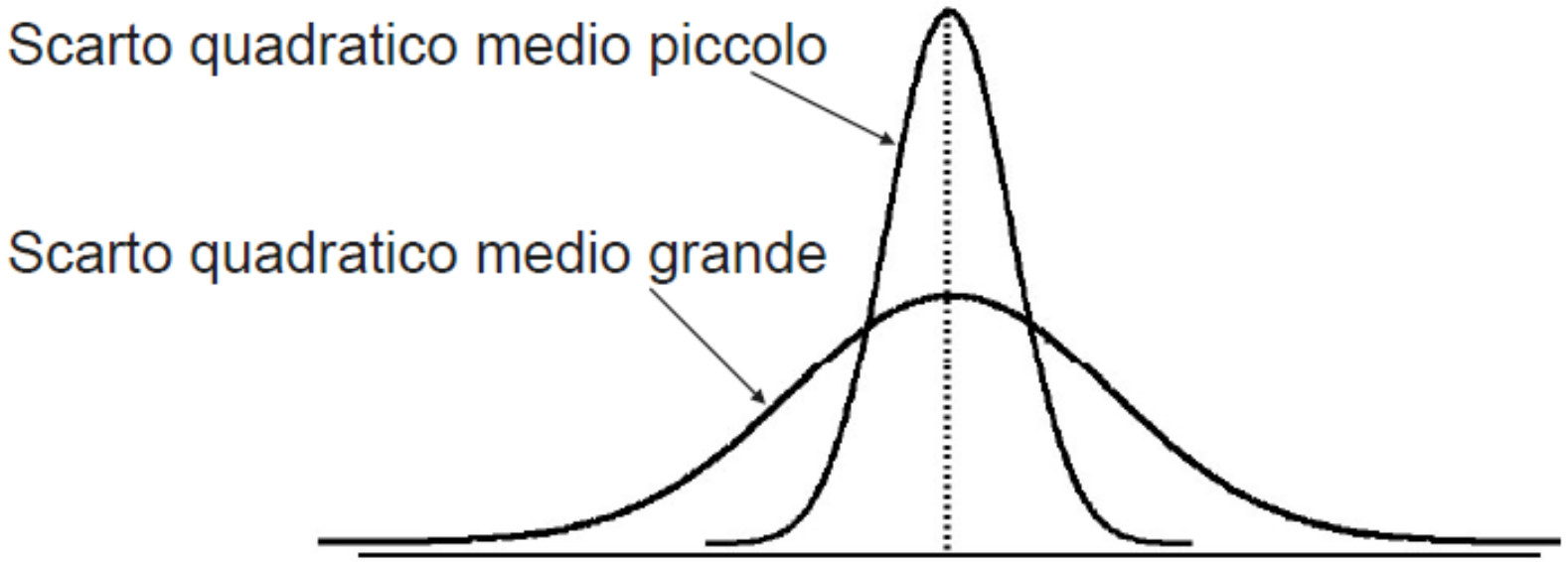
$$V(a + bX) = b^2 \sigma_x^2$$

Scarto quadratico medio della popolazione (o deviazione standard)

- Radice quadrata della Varianza
- Indice di dispersione con unità di misura uguale alla media.
- Indica di quanto, mediamente, i dati osservati si discostano dalla loro media.

$$\sigma_X = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2 f_i}{n}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 f_i}{n} - M^2}$$

Scarto quadratico medio della popolazione (o deviazione standard)



Coefficiente di variazione

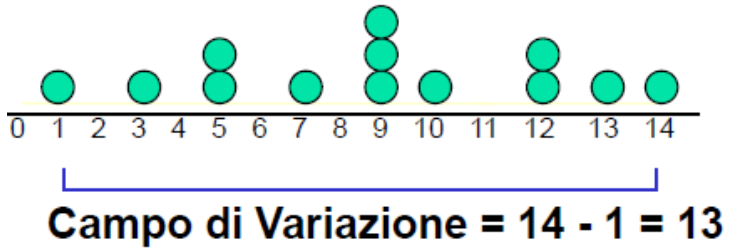
- Il coefficiente di variazione sintetizza il rapporto tra Media e Deviazione Standard
- Determina la dispersione dei dati osservati mediante l'uso della Media come unità di misura
- E' un indicatore di variabilità relativa
- E' particolarmente utile per confrontare due differenti distribuzioni

$$CV = \frac{\sigma}{\overline{X}}$$

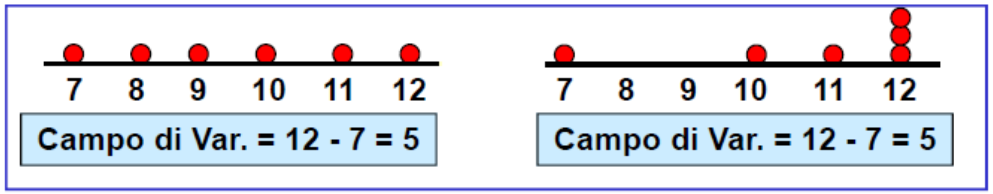
Campo di variazione

- La più semplice misura di variabilità
- Differenza tra il massimo e il minimo dei valori osservati

$$\text{Campo di variazione} = X_{\text{massimo}} - X_{\text{minimo}}$$



- Svantaggi:
- ignora il modo in cui i dati sono distribuiti
- sensibile agli outlier



Differenza interquartile

- Possiamo eliminare il problema degli outlier usando la **differenza interquartile**
- Elimina i valori osservati più alti e più bassi e calcola il campo di variazione del 50% centrale dei dati
- Differenza Interquartile = $Q_3 - Q_1$
- Ricordando che il primo quartile è l'osservazione di posizione 0.25(n+1) nella serie ordinata, mentre il terzo quartile occupa la posizione 0.75(n+1)

$$IQR = Q_3 - Q_1$$

Indici di variabilità

- **Variazione**: X_{max}-X_{min}
  Funzione di excel: "*max-min*"
- **Differenza interquantile**: Q_3-Q_1
  Funzione di excel: "*percentile(;0,75)-percentile(;0,25)*"
- ***Varianza:***

$$VAR(X) = \sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

Dati grezzi

$$VAR(X) = \sigma^2 = \frac{\sum_{i=1}^{k}(X_i - \overline{X})^2 f_i}{n}$$

Distribuzioni di frequenza

Funzione di excel: "*var.pop*"

Carattere quantitativo discreto

v.s. Ore dedicate a News in Internet

Ore News Internet	fi	fi%	Fi	Fi%	Xifi	Xi^2 fi
1	4	13,33	4	13,33	4	4
2	7	23,33	11	36,67	14	28
3	4	13,33	15	50,00	12	36
4	4	13,33	19	63,33	16	64
5	4	13,33	23	76,67	20	100
6	4	13,33	27	90,00	24	144
7	1	3,33	28	93,33	7	49
8	2	6,67	30	100,00	16	128
	30	100,00		somma	113	553
				media	3,766667	18,43333
				varianza		4,245556

- Mo=2 MODA(K2:K31)
- Me=3,5 MEDIANA(K2:K31)
- Media=3,77 MEDIA(K2:K31)
- Q1=2 PERCENTILE(K2:K31;0,25)
- Q3=5 PERCENTILE(K2:K31;0,75)

- Xmax-Xmin=8-1=7
- Q3-Q1=5-2=3
- Var=4,25 VAR.POP(K2:K31)

Carattere quantitativo in classi

v.s. Età

Età	linf	lsup	Xc	fi	fi%	Fi	Fi%	Xc*fi	Xc^2*fi
20-|30	20	30	25	10	33,33	10	33,33	250	6250
30-|40	30	40	35	9	30,00	19	63,33	315	11025
40-|50	40	50	45	3	10,00	22	73,33	135	6075
50-|60	50	60	55	8	26,67	30	100,00	440	24200
				30	100,00		somma	1140	47550
							media	38	1585

- Mo=25
- Me=36 posizione (n+1)/2
- Q1=28 posizione (n+1)*75/100
- Q3=52 posizioni (n+1)*25/100

- Media=38
- Var=141

N.B. Si ottengono valori differenti se gli indici vengono calcolati sui dati non in classi

Carattere quantitativo discreto

v.s. Ore dedicate alla Televisione

Ore televisione	linf	lsup	fi	fi%	Fi%	xc	xi *fi	xi^2 * fi
8 -| 17	8	17	12	40	40	12,5	150	1875
17 -| 26	17	26	7	23,33333	63,33333	21,5	150,5	3235,75
26 -| 35	26	35	8	26,66667	90	30,5	244	7442
35 -| 44	35	44	3	10	100	39,5	118,5	4680,75
			30	100		somma	663	17233,5
						media	22,1	574,45
						varianza	86,04	

- Mo=14 MODA(K2:K31)
- Me=20,5 MEDIANA(K2:K31)
- Media=22 MEDIA(K2:K31)
- Q1=14 PERCENTILE(K2:K31;0,25)
- Q3=29 PERCENTILE(K2:K31;0,75)

- Xmax-Xmin=42-8=34
- Q3-Q1=29-14=15
- Var=84,93 VAR.POP(K2:K31)

Confronto

v.s. Ore dedicate alla Televisione
v.s. Ore dedicate a News in Internet

	MEDIA	VAR.POP	DEV.ST	CV
Ore settimanali dedicate alla televisione	22,00	84,93	9,22	0,42
Ore settimanali dedicate a News in Internet	3,77	4,25	2,06	0,55

- $\sigma^2_{Televisione}$=84,93 > $\sigma^2_{News\ Internet}$ =4,25
- CV Televisione=$\frac{\sigma}{\mu} = \frac{9{,}22}{22}$ =0,42
- CV News Internet=$\frac{\sigma}{\mu} = \frac{2{,}06}{3{,}77}$ =0,55

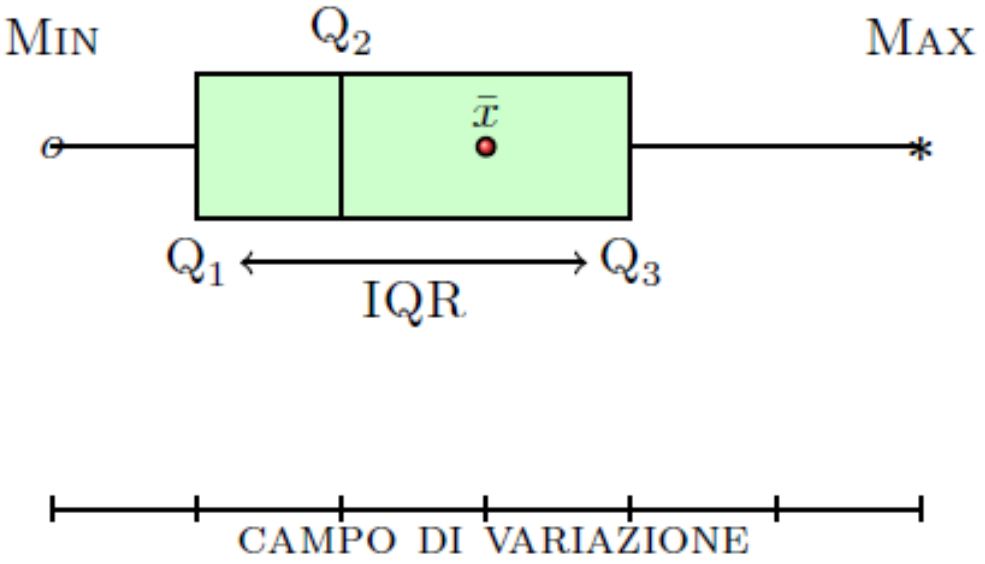
CV Televisione < CV News Internet

Grafici BOX-PLOT (o Box&Whiskers)

- GRAFICO RIASSUNTIVO DEI MAGGIORI INDICI DESCRITTIVI UNIVARIATI CHE CONSENTE CONFRONTI "VISIVI" TRA DIVERSE VARIABILI

- Per ogni variabile vengono rappresentate:

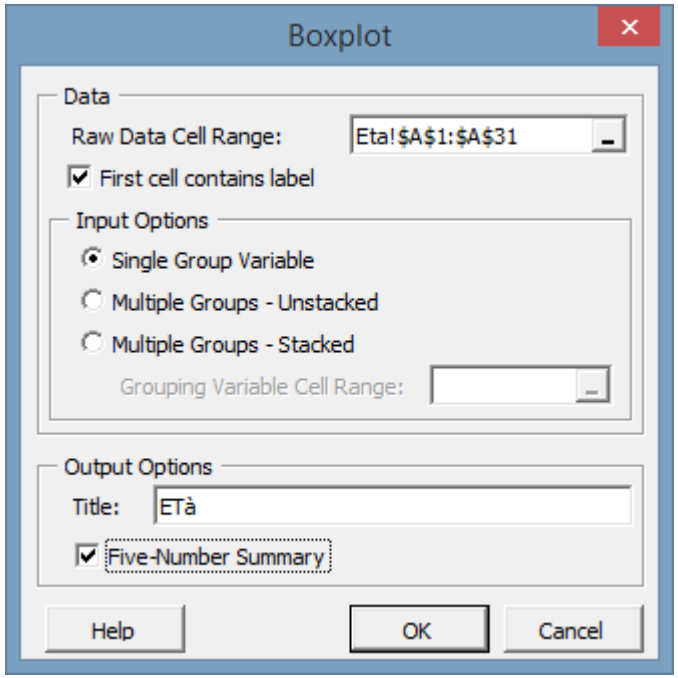
- mediana (Q2)
- I e III quartile (Q1 e Q3)
- Differenza interquartile IQR = Q3 – Q1
- minimo e massimo



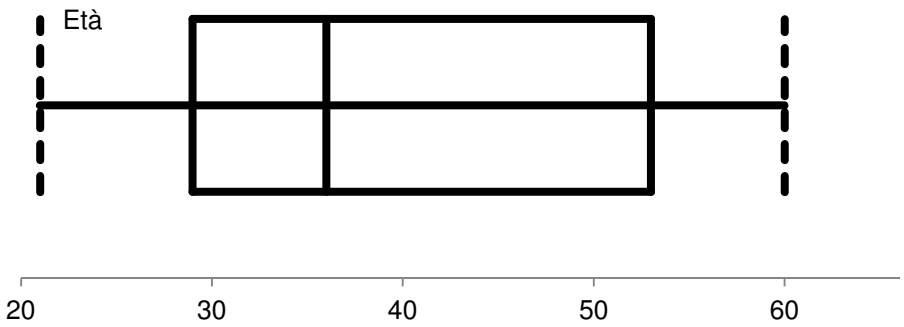
Carattere quantitativo

v.s. Età

PHStat2 -> Descriptive Statistics -> Boxplot



	A	B
1	ETà	
2		
3	Five-number Summary	
4	Minimum	21
5	First Quartile	29
6	Median	36
7	Third Quartile	53
8	Maximum	60

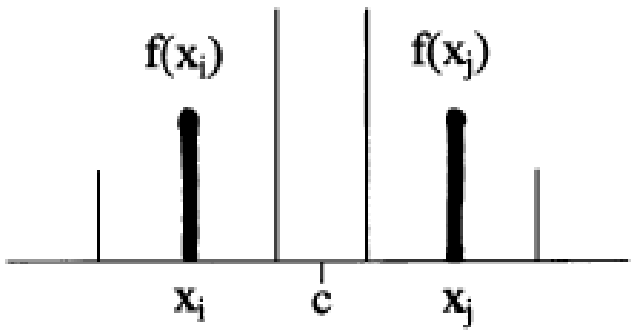


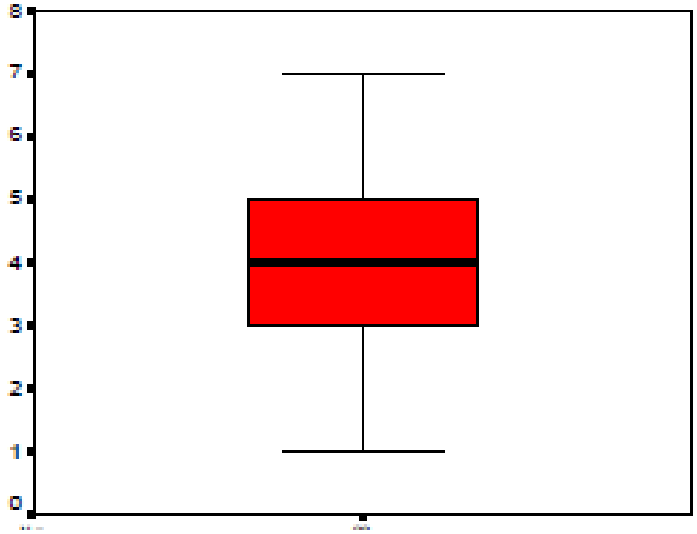
Forma di una distribuzione di frequenza

Una v.s. è **simmetrica** rispetto ad un centro **c** se:

- per ogni $x_i = c - k$
- esiste un $x_j = c + k$ (simmetrico)

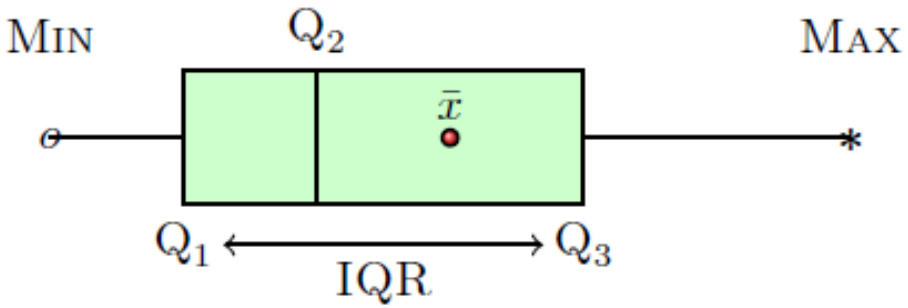
con stessa frequenza: $f(x_i) = f(x_j)$



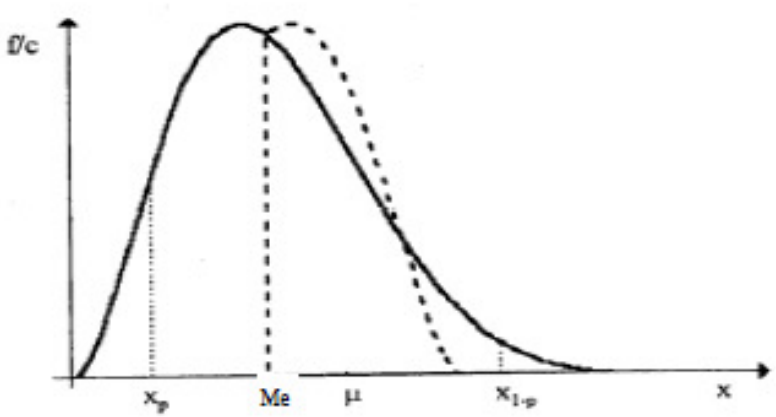


Grafici BOX-PLOT (o Box&Whiskers)

- **Distribuzione obliqua a destra (asimmetria positiva)**

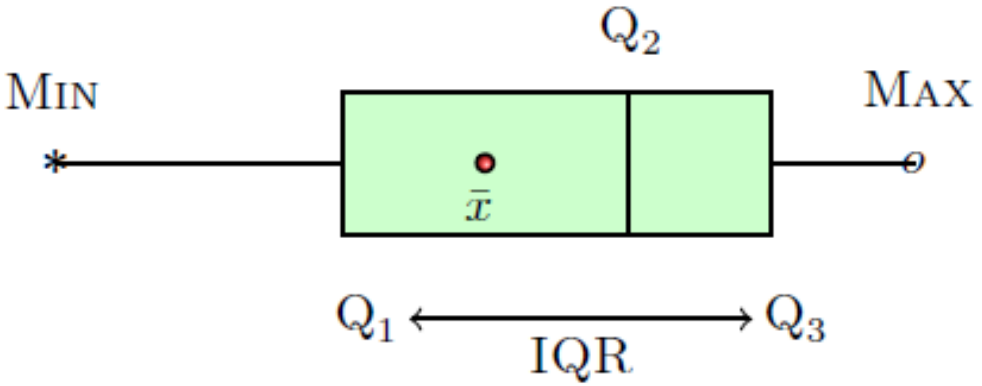


1. $\bar{x} > Q_2$
2. $Q_3 - Q_2 > Q_2 - Q_1$
3. $\text{Max} - Q_3 > Q_1 - \text{Min}$

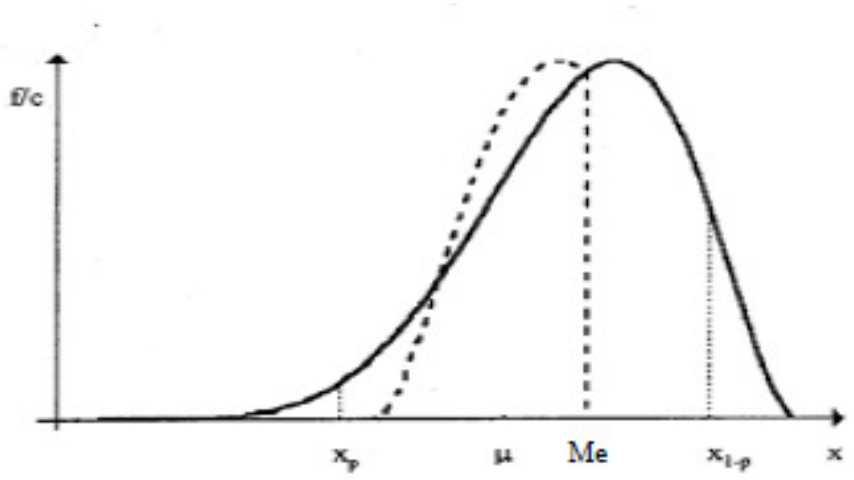


Grafici BOX-PLOT (o Box&Whiskers)

- **Distribuzione obliqua a sinistra (asimmetria negativa)**

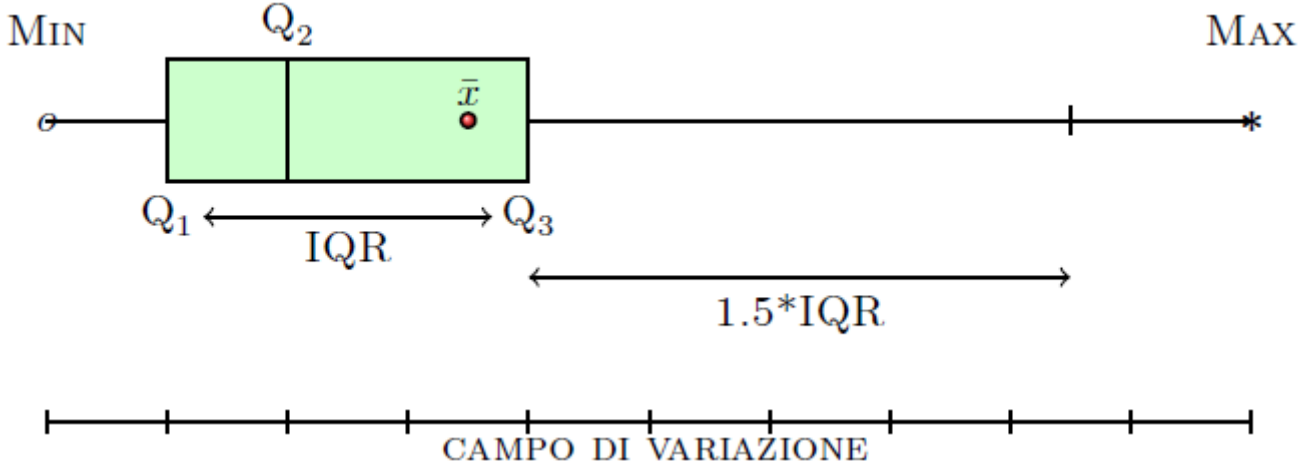


1. $\bar{x} < Q_2$
2. $Q_3 - Q_2 < Q_2 - Q_1$
3. $Max - Q_3 < Q_1 - Min$



Grafici BOX-PLOT (o Box&Whiskers)

- **Box – plot con outlier**



Outliers

un singolo dato x_i è definito *outlier* se

- $x_i < Q_1 - 1.5 \cdot IQR$
- $x_i > Q_3 + 1.5 \cdot IQR$

Indici di forma

- **Indice di FISHER o di SKEWNESS**

  (più comunemente usato)

$$\gamma_1 = \frac{M[(X-\mu)^3]}{\sigma^3} = \frac{\mu_3}{\sigma^3}$$

se **(asimmetria positiva)** $\Rightarrow \quad \gamma_1 > 0$

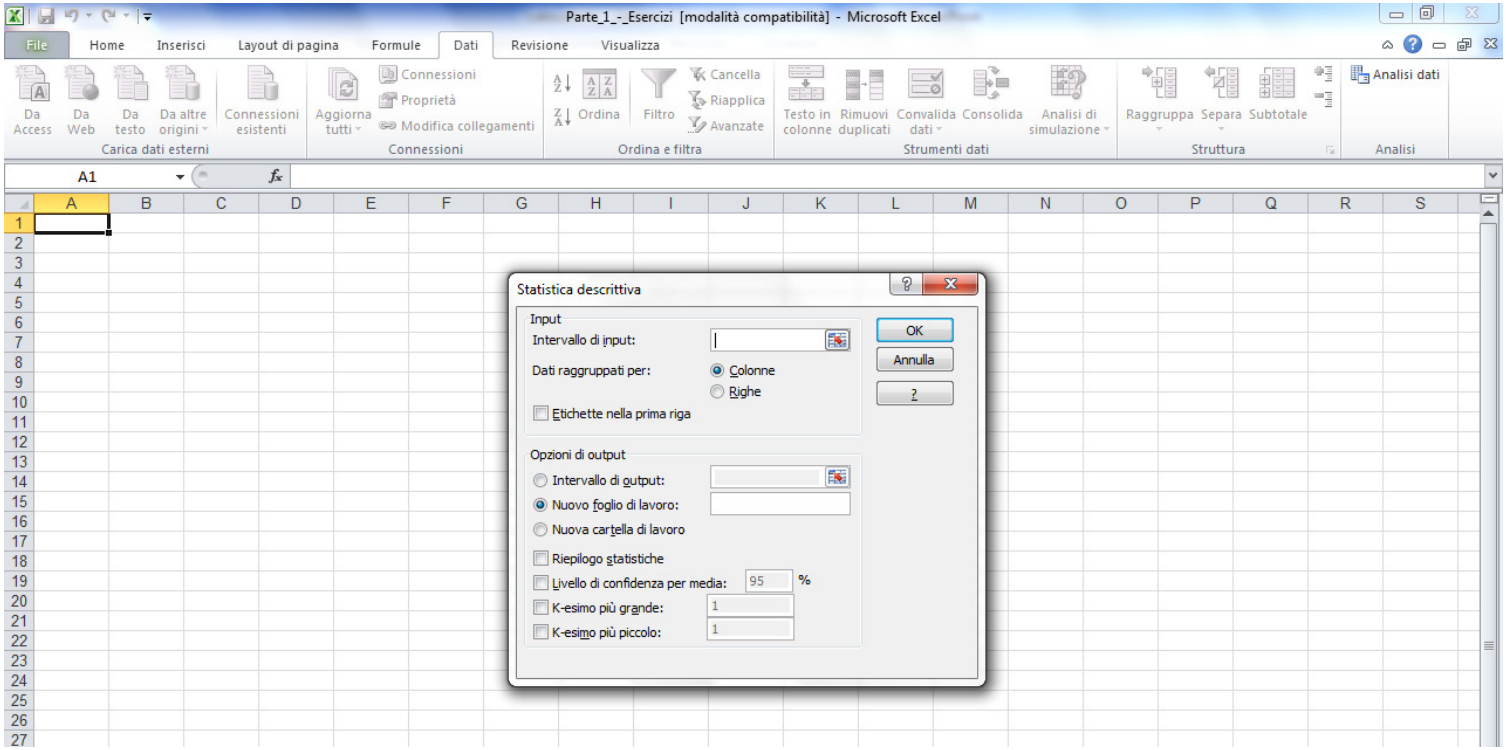
se **(asimmetria negativa)** $\Rightarrow \quad \gamma_1 < 0$

se simmetria $\Rightarrow \quad \gamma_1 = 0$

IMPORTANTE

Si può utilizzare anche lo strumento di excel:

"Dati → Analisi dati → statistica descrittiva"



Esempio carattere quantitativo discreto

v.s. Ore settimanali dedicate ai News in Internet

Ore dedicate a News in Internet		
Media	3,767	**=MEDIA(K2:K31)**
Errore standard	0,383	
Mediana	3,500	**=MEDIANA(K2:K31)**
Moda	2,000	**=MODA(K2:K31)**
Deviazione standard	2,096	
Varianza campionaria	4,392	**=VAR(K2:K31)**
Curtosi	-0,777	
Asimmetria	0,452	
Intervallo	7,000	**=MAX(K2:K31)-MIN(K2:K31)**
Minimo	1,000	**=MIN(K2:K31)**
Massimo	8,000	**=MAX(K2:K31)**
Somma	113,000	**=SOMMA(K2:K31)**
Conteggio	30,000	**=CONTA.NUMERI(K2:K31)**