Statistica

Capitolo 3

Descrizione Numerica dei Dati

Obiettivi del Capitolo

Dopo aver completato il capitolo, sarete in grado di:

- Calcolare ed interpretare la media, la mediana e la moda di un set di dati
- Trovare il campo di variazione, varianza, scarto quadratico medio, e coefficiente di variazione e conoscere il loro significato
- Applicare la regola empirica per descrivere la variazione dei valori della popolazione attorno alla media
- Spiegare la media pesata e quando usarla
- Spiegare come una retta di regressione ottenuta con il metodo dei minimi quadrati stima la relazione lineare fra due variabili

Argomenti Trattati nel Capitolo

- Misure di tendenza centrale, variabilità, e forma
  - Media, mediana, moda, media geometrica
  - Quartili
  - Campo di variazione, differenza interquartile, varianza e scarto quadratico medio, coefficiente di variazione
  - Distribuzioni simmetriche e asimmetriche
- Misure di sintesi per la popolazione
  - Media, varianza, e scarto quadratico medio
  - La regola empirica e la disuguaglianza di Chebyshev

Argomenti Trattati nel Capitolo

(continuazione)

- Cinque numeri di sintesi e Box Plot
- Covarianza e coefficiente di correlazione
- Problemi con le misure usate per descrivere i dati numericamente e considerazioni etiche

Descrizione Numerica dei Dati

- **Descrizione numerica dei dati**
  - **Tendenza Centrale**
    - Media Aritmetica
    - Mediana
    - Moda
  - **Variabilità**
    - Campo di Variazione
    - Differenza Interquartile
    - Varianza
    - Scarto Quadratico Medio
    - Coefficiente di Variazione

Misure di Tendenza Centrale

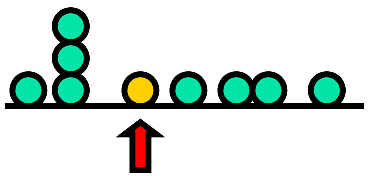
Panoramica

Tendenza Centrale

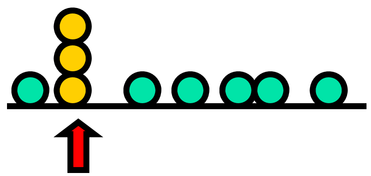
Media | **Mediana** | **Moda**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Media Aritmetica



Valore centrale delle osservazioni ordinate



Valore più frequente

Media Aritmetica

- La media aritmetica (media) è la misura di tendenza centrale più comune
  - Per una popolazione di N valori:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Valori della popolazione

Dimensione della popolazione

  - Per un campione di dimensione n:

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

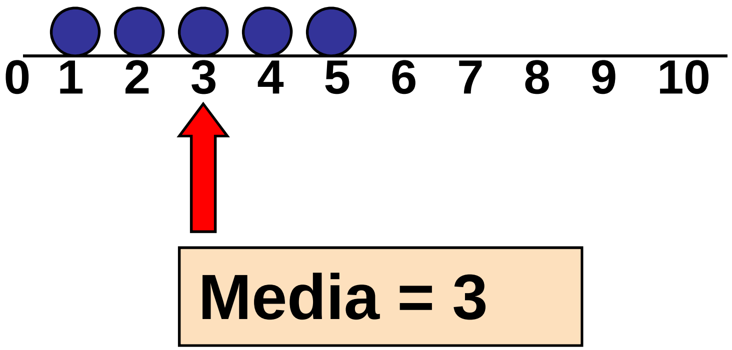
Valori osservati

Dimensione del campione

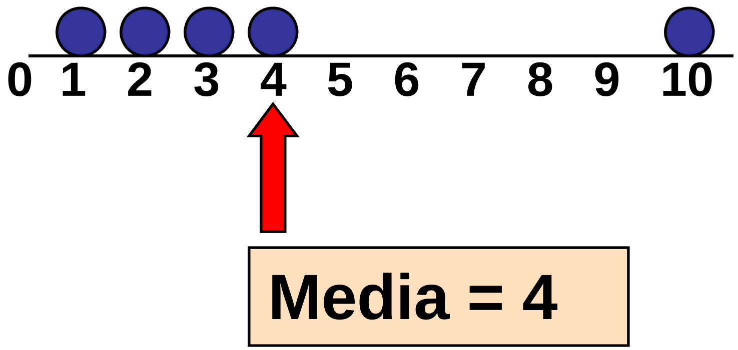
Media Aritmetica

(continuazione)

- La misura di tendenza centrale più comune
- Media = somma dei valori diviso il numero di valori
- Influenzata da valori estremi (outlier)



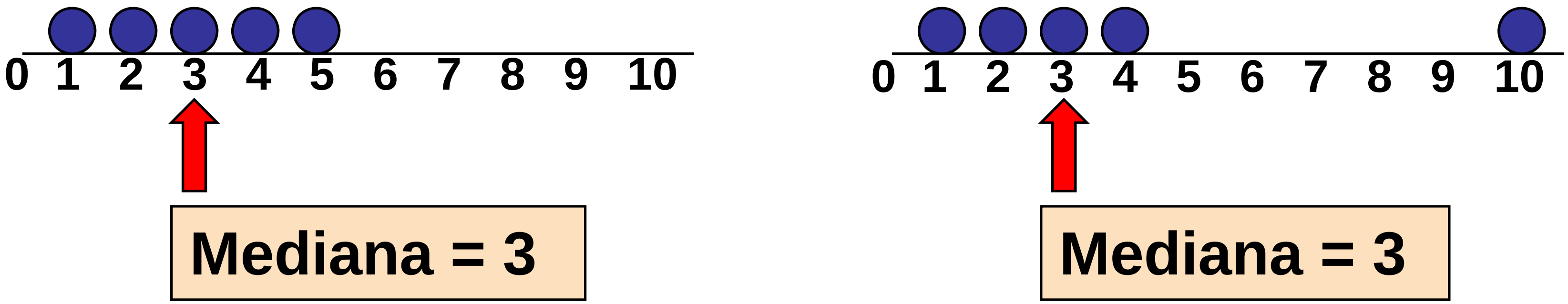
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Mediana

- In una lista ordinata, la mediana è il valore "centrale" (50% prima, 50% dopo)



- Non influenzata da valori estremi

Trovare la Mediana

- La posizione della mediana:

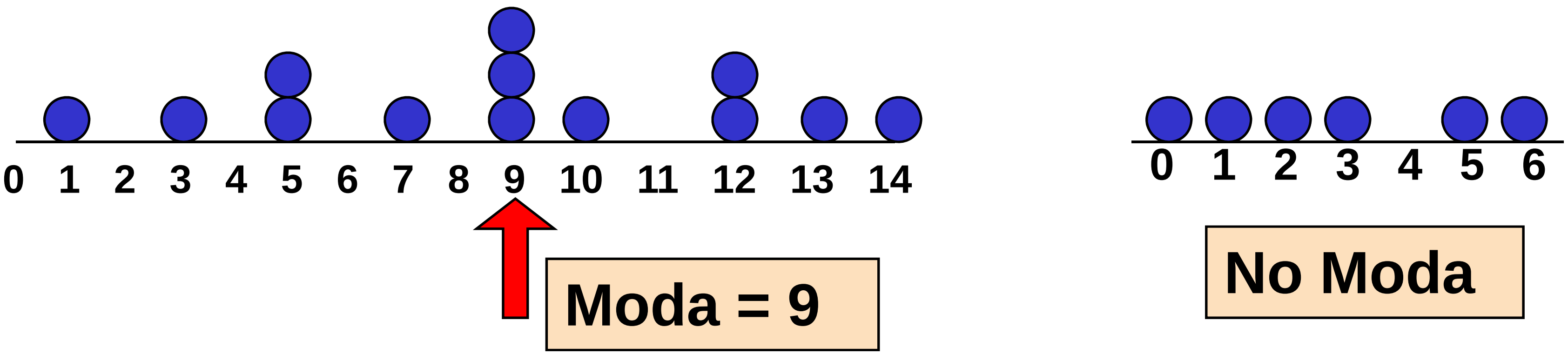
$$\text{Posizione Mediana} = \frac{n+1}{2} \text{ posizione nella sequenza ordinata}$$

  - 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* della mediana nella sequenza ordinata

Moda

- Una misura di tendenza centrale
- Valore che ricorre più frequentemente
- Non influenzata da valori estremi
- Usata sia per dati numerici che categorici
- Può non esserci una moda
- Ci può essere più di una moda

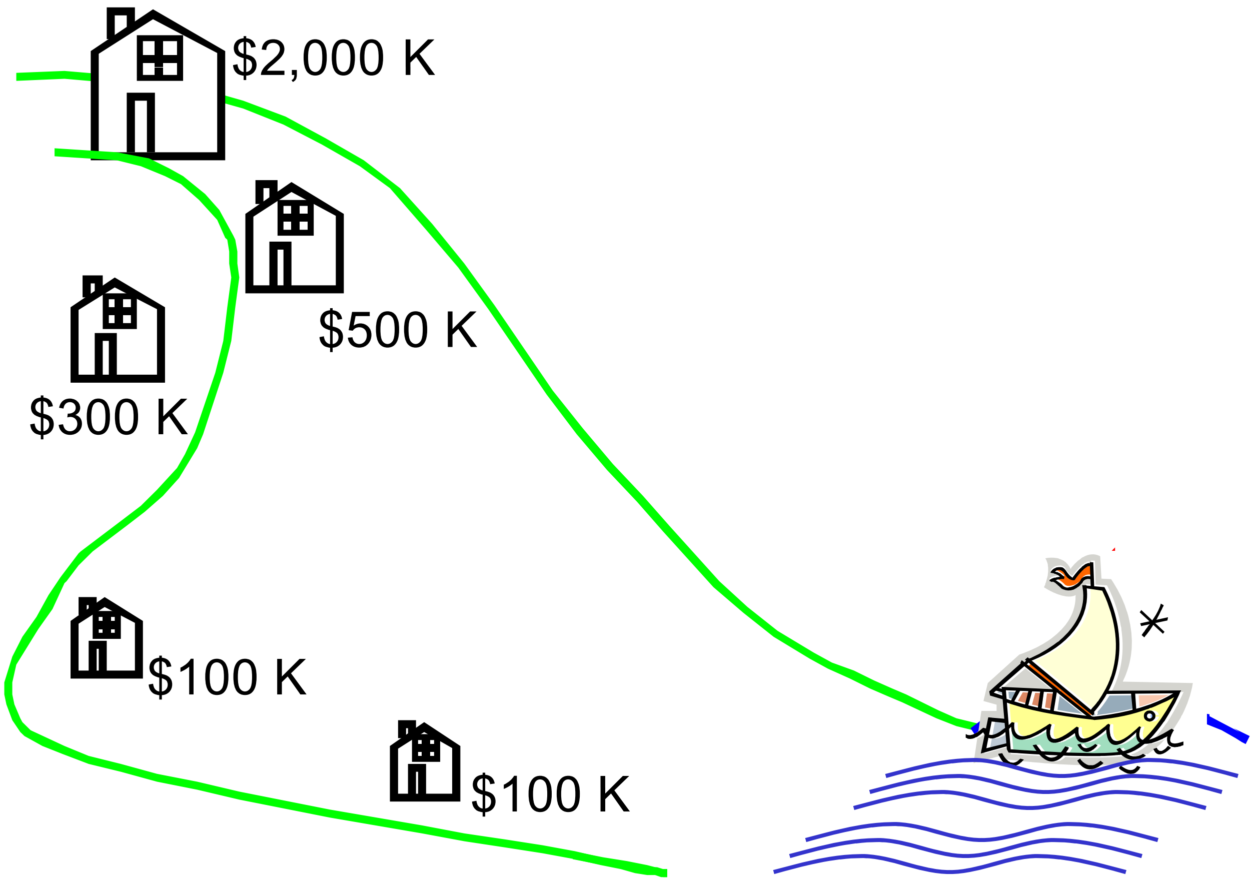


Esempio Riepilogativo

- Cinque case su una collina presso una spiaggia

Prezzi delle case:

$2,000,000
500,000
300,000
100,000
100,000



Esempio Riepilogativo: Misure di Sintesi

Prezzi delle case:

$2,000,000
500,000
300,000
100,000
100,000

Somma **3,000,000**

- **Media:** ($3,000,000/5)
  = **$600,000**

- **Mediana:** valore centrale dei dati ordinati
  = **$300,000**

- **Moda:** valore più frequente
  = **$100,000**

Quale misura di tendenza centrale è la "migliore"?

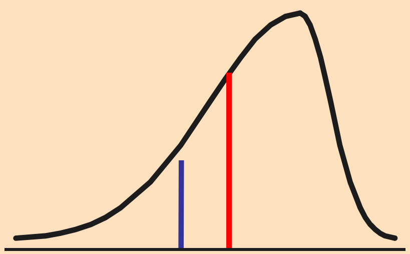
- La **media** è usata in generale, a meno che ci siano valori estremi (outlier)
- La **mediana** è usata spesso siccome non è influenzata da valori estremi.
  - Esempio: Il prezzo mediano delle case può essere riportato per una regione – meno sensibile agli outlier

Forma della Distribuzione

- Descrive come i dati sono distribuiti
- Misure della forma
  - Simmetrica o asimmetrica

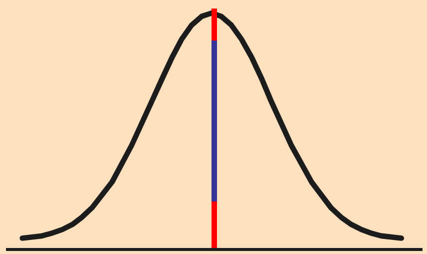
Obliqua a sinistra

Media < Mediana



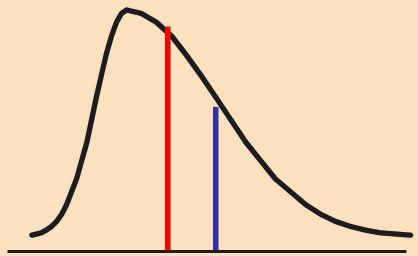
Simmetrica

Media = Mediana



Obliqua a destra

Media > Mediana

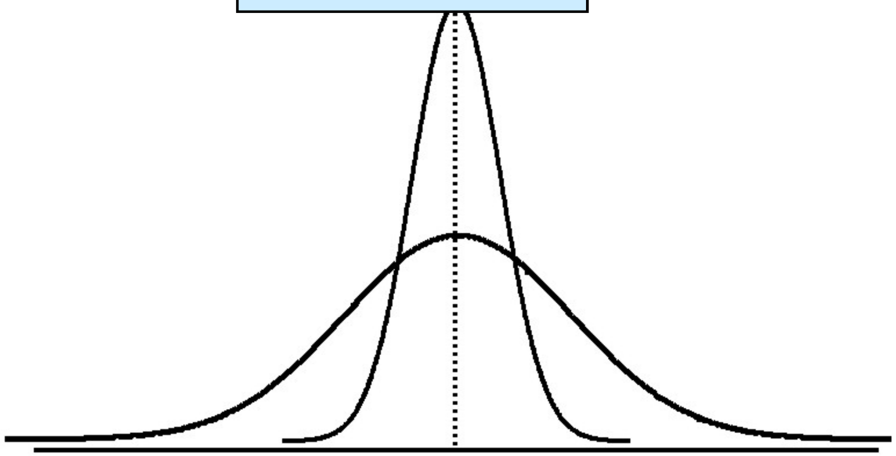


Misure di Variabilità

Variabilità

- **Campo di Variazione**
- **Differenza Interquartile**
- **Varianza**
- **Scarto Quadratico Medio**
- **Coefficiente di Variazione**

- Le misure di variabilità forniscono informazioni sulla **dispersione** o **variabilità** dei valori.



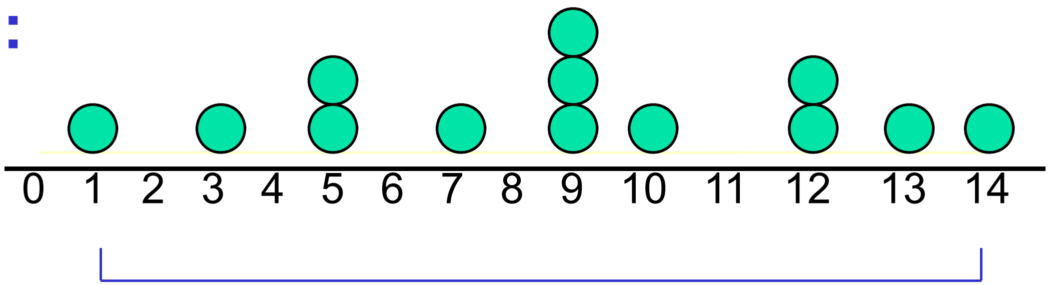
Stesso centro, diversa variabilità

Campo di Variazione

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

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

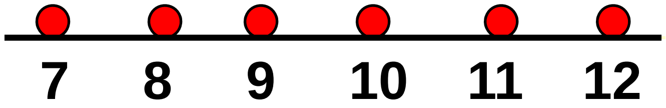
Esempio:



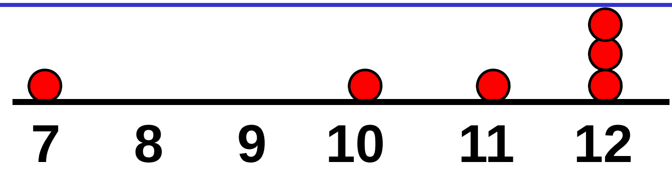
Campo di Variazione = 14 - 1 = 13

Svantaggi del Campo di Variazione

- Ignora il modo in cui i dati sono distribuiti



Campo di Var. = 12 - 7 = 5



Campo di Var. = 12 - 7 = 5

- Sensibile agli outlier

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,5

Campo di Var. = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,120

Campo di Var = 120 - 1 = 119

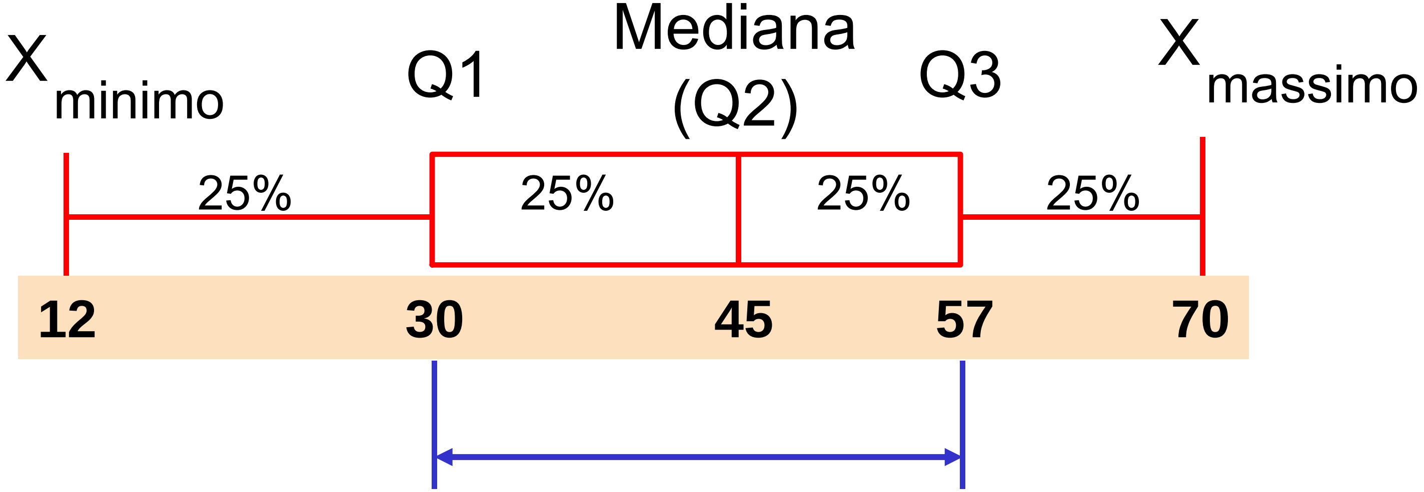
Differenza Interquartile

- 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 = 3^{zo} quartile – 1mo quartile
  Si noti come 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)

$$IQR = Q_3 - Q_1$$

Differenza Interquartile

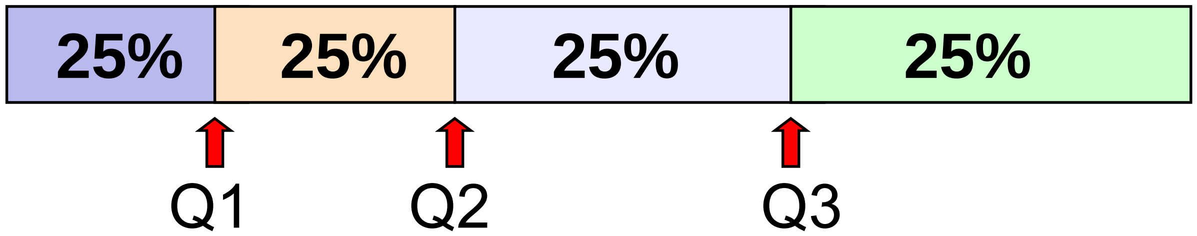
Esempio:



Differenza Interquartile
= 57 – 30 = 27

Quartili

- I Quartili dividono la sequenza ordinata dei dati in 4 segmenti contenenti lo stesso numero di valori



- Il primo quartile, Q_1, è il valore per il quale 25% delle osservazioni sono minori e 75% sono maggiori di esso
- Q_2 coincide con la mediana (50% sono minori, 50% sono maggiori)
- Solo 25% delle osservazioni sono maggiori del terzo quartile

Formule per i Quartili

Un quartile si trova determinando il valore della sua posizione nella sequenza ordinata dei dati, dove

Posizione primo quartile: $Q_1 = 0.25(n+1)$

Posizione secondo quartile: $Q_2 = 0.50(n+1)$
(la posizione della mediana)

Posizione terzo quartile: $Q_3 = 0.75(n+1)$

dove **n** è il numero di valori osservati

Quartili

- Esempio: Trova il primo quartile

Dati Campionari Ordinati:	**11**	**12**	**13**	**16**	**16**	**17**	**18**	**21**	**22**

(n = 9)

Q_1 = è nella 0.25(9+1)=2.5 posizione nella sequenza ordinata dei dati, usiamo quindi la media fra il 2do e il 3zo valore,

per cui $\boxed{Q_1 = 12.5}$

Varianza della Popolazione

- Media dei quadrati delle differenze fra ciascuna osservazione e la media

  - Varianza della Popolazione:

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

dove μ = media della popolazione

N = dimensione della popolazione

x_i = i^{imo} valore della variabile X

Varianza Campionaria

- Media (approssimativamente) dei quadrati delle differenze fra ciascuna osservazione e la media
  - Varianza campionaria: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

dove $\bar{X}$ = media aritmetica

n = dimensione del campione

X_i = i^{mo} valore della variabile X

Scarto Quadratico Medio della Popolazione

- Misura di variabilità comunemente usata
- Mostra la variabilità rispetto alla media
- Ha la stessa unità di misura dei dati originali

  - Scarto Quadratico Medio della Popolazione:

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

Scarto Quadratico Medio Campionario

- Misura di variabilità comunemente usata
- Mostra la variabilità rispetto alla media
- Ha la stessa unità di misura dei dati originali

  - Scarto Quadratico Medio Campionario:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Esempio di Calcolo: Scarto Quadratico Medio Campionario

Dati Campionari (x_i) :

10	**12**	**14**	**15**	**17**	**18**	**18**	**24**

n = 8   Media = $\bar{x}$ = 16

$$s = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{x})^2 + (14 - \bar{x})^2 + \cdots + (24 - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \cdots + (24 - 16)^2}{8-1}}$$

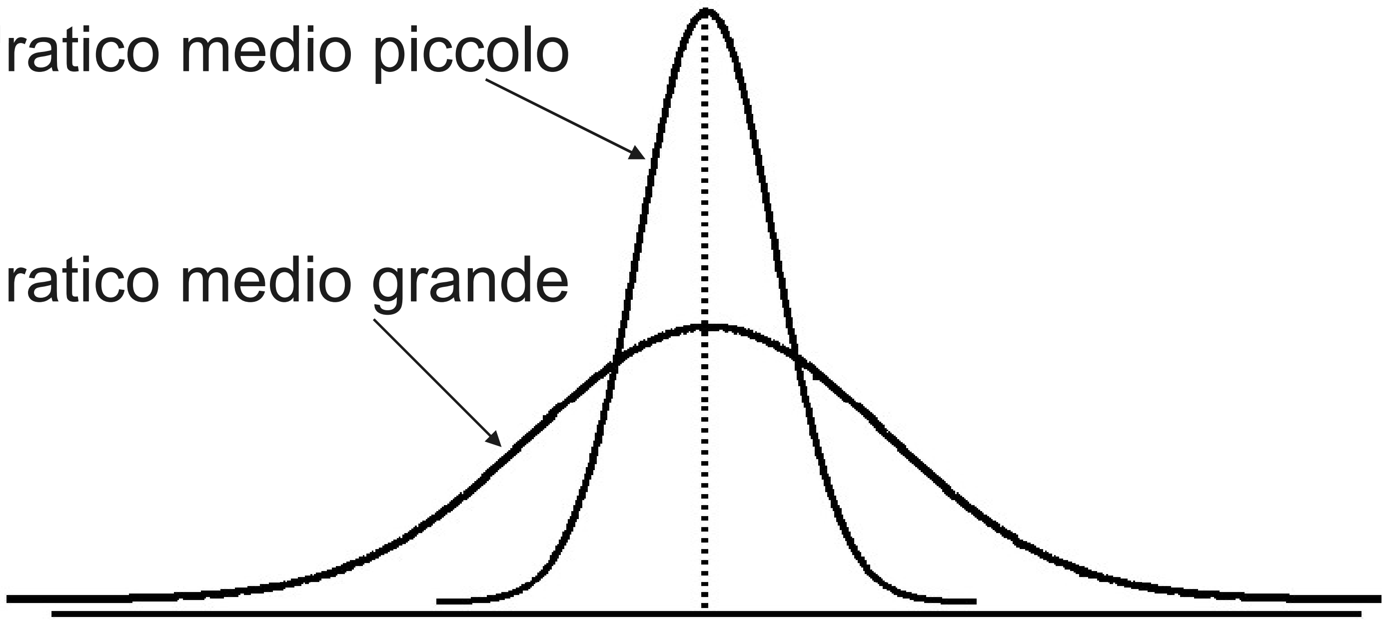
$$= \sqrt{\frac{130}{7}} = \boxed{4.3095}$$

Una misura della dispersione “media” attorno alla media

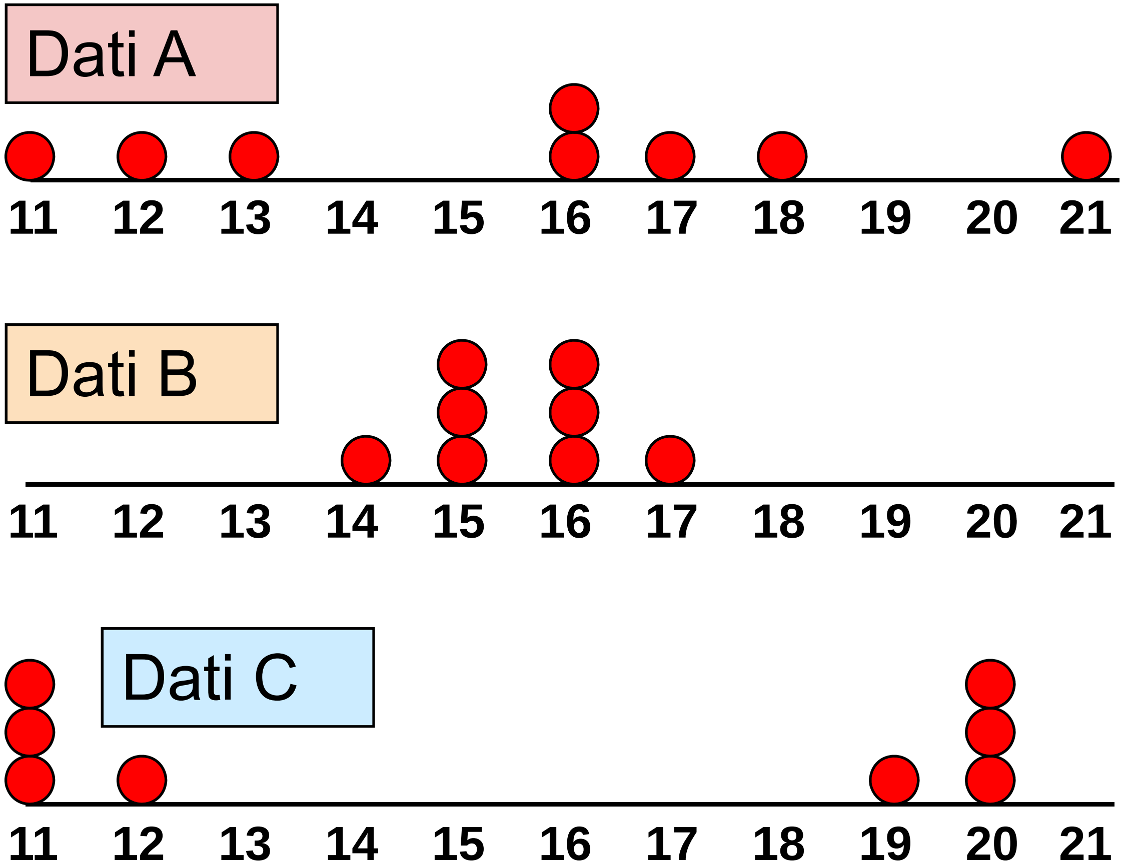
Misurando la Variabilità

Scarto quadratico medio piccolo

Scarto quadratico medio grande



Confrontando lo Scarto Quadratico Medio



Media = 15.5
s = 3.338

Media = 15.5
s = 0.926

Media = 15.5
s = 4.570

Vantaggi della Varianza e dello Scarto Quadratico Medio

- Sono calcolati usando tutti i valori nel set di dati

- Valori lontani dalla media hanno più peso
  (poichè si usa il quadrato delle deviazioni dalla media)

Teorema di Chebyshev

- Per ogni popolazione con media μ, scarto quadratico medio σ, e $k > 1$, la percentuale di osservazioni che appartengono all'intervallo

$$[\mu - k\sigma ; \mu + k\sigma]$$

è *almeno*

$$\boxed{100[1-(1/k^2)]\%}$$

Teorema di Chebyshev

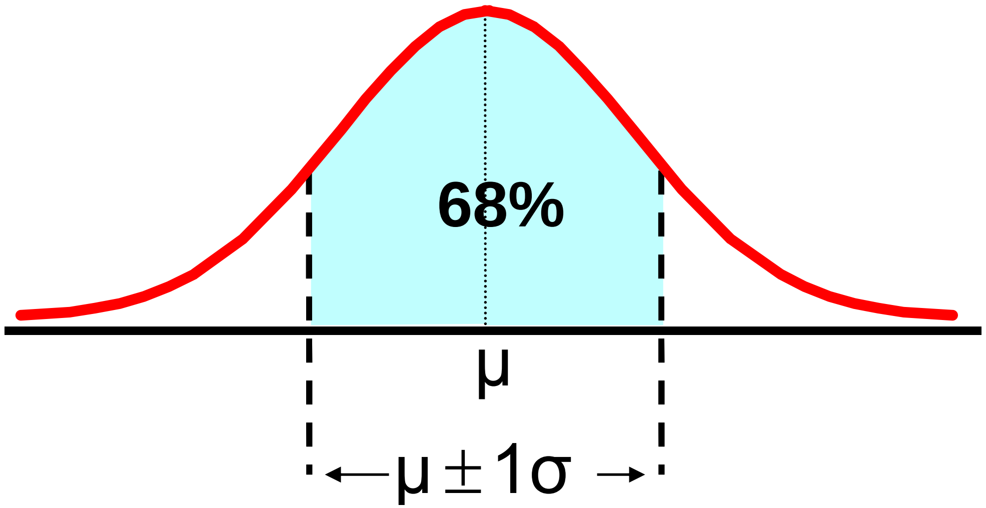
(continuazione)

- Indipendentemente da come i dati sono distribuiti, almeno $(1 - 1/k^2)$ dei valori cadranno entro k scarti quadratici medi dalla media (per $k > 1$)
  - Esempi:

Almeno		entro	
$(1 - 1/1^2) =$	0%	$k=1$	$(\mu \pm 1\sigma)$
$(1 - 1/2^2) =$	75%	$k=2$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) =$	89%	$k=3$	$(\mu \pm 3\sigma)$

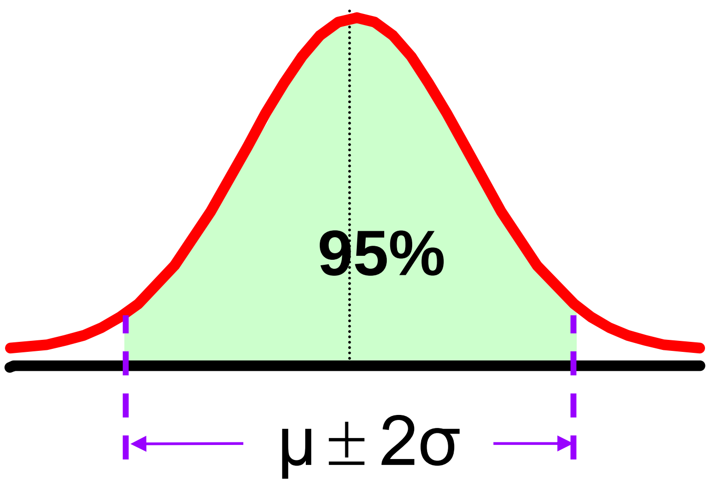
La Regola Empirica

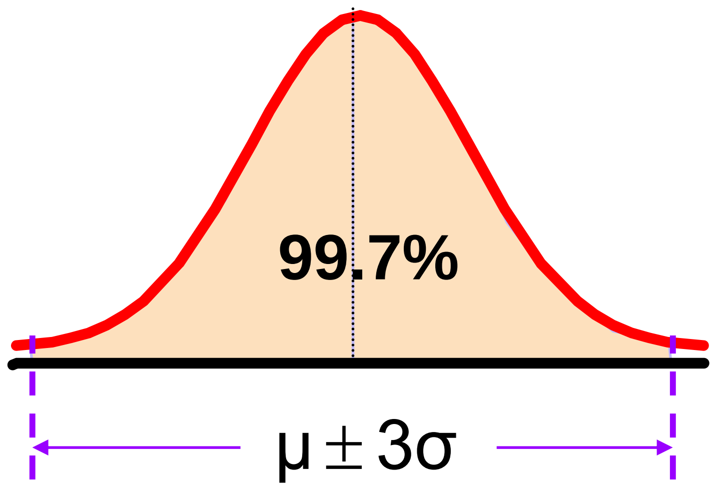
- Se la distribuzione dei dati ha una forma simmetrica e campanulare, allora l'intervallo:
- $\mu \pm 1\sigma$ contiene circa 68% dei valori della popolazione o del campione



La Regola Empirica

- $\mu \pm 2\sigma$ contiene circa 95% dei valori della popolazione o del campione
- $\mu \pm 3\sigma$ contiene circa 99.7% dei valori della popolazione o del campione





Coefficiente di Variazione

- Misura la variabilità relativa
- Sempre in percentuale (%)
- Mostra la variabilità relativa rispetto alla media
- Può essere usato per confrontare due o più set di dati misurati con unità di misura diversa

$$CV = \left(\frac{\sigma}{|\mu|} \right) \cdot 100\%$$

$$CV = \left(\frac{s}{|\bar{x}|} \right) \cdot 100\%$$

Confronto fra Coefficienti di Variazione

- Azione A:
  - Prezzo medio scorso anno = \$50
  - Scarto quadratico medio = \$5

$$CV_A = \left(\frac{s}{|\bar{x}|} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Azione B:
  - Prezzo medio scorso anno = \$100
  - Scarto quadratico medio = \$5

$$CV_B = \left(\frac{s}{|\bar{x}|} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

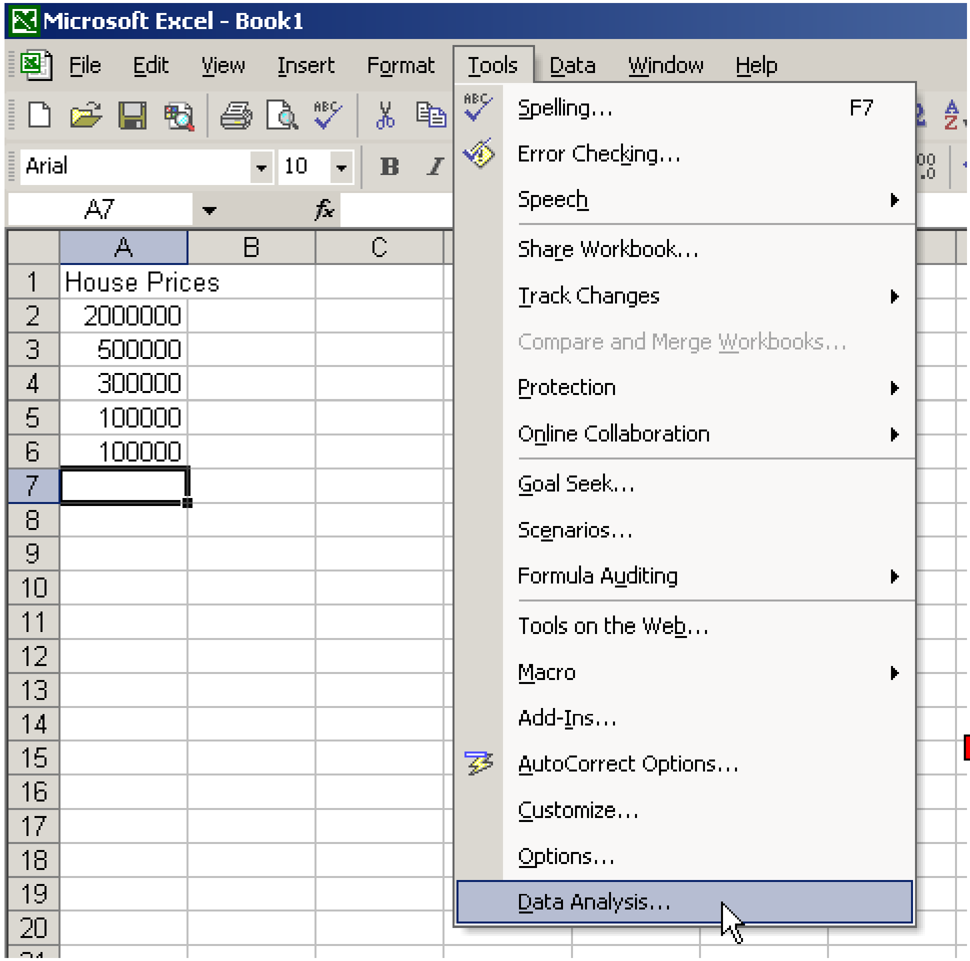
Entrambe le azioni hanno lo stesso scarto quadratico medio, ma l'azione B è meno variabile rispetto al suo prezzo medio

Usando Microsoft Excel

- Statistica Descrittiva può essere condotta usando Microsoft® Excel
  - Seleziona il menu:

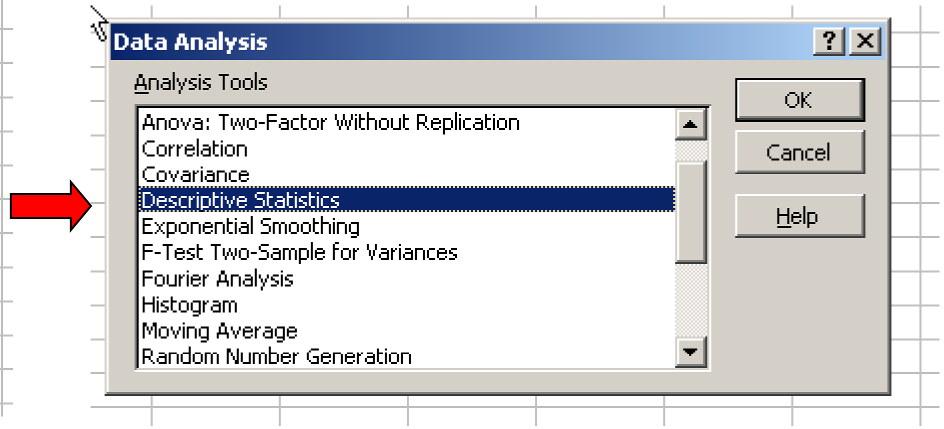
    strumenti / analisi dati / statistica descrittiva
  - Inserire i dettagli nella finestra di dialogo

Usando Excel



- Seleziona il menu:

  strumenti / analisi dati / statistica descrittiva



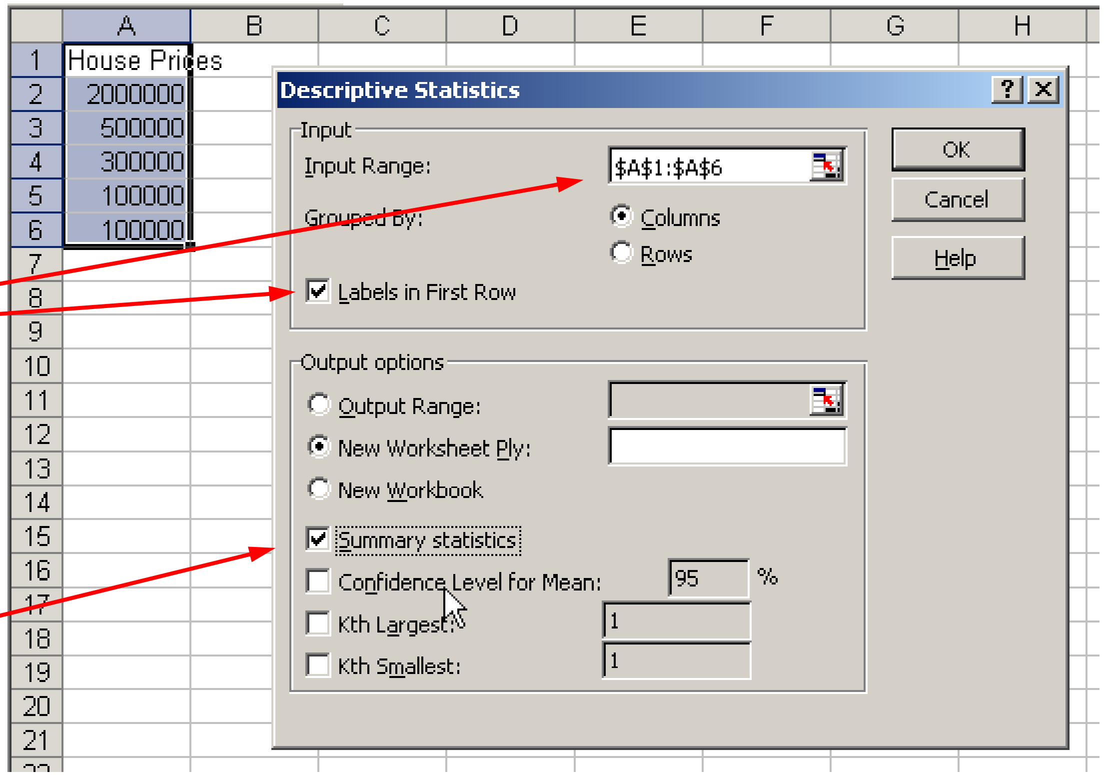
Using Excel

(continuazione)

- Inserire dettagli nella finestra di dialogo

- Seleziona l'opzione Riepilogo statistiche

- Cliccare su OK



Output di Excel

Output di Microsoft Excel di statistica descrittiva

usando i dati sul prezzo delle case:

Prezzi delle case:

$2,000,000
500,000
300,000
100,000
100,000

	A	B
1	*House Prices*	
2		
3	Mean	600000
4	Standard Error	357770.8764
5	Median	300000
6	Mode	100000
7	Standard Deviation	800000
8	Sample Variance	6.4E+11
9	Kurtosis	4.130126953
10	Skewness	2.006835938
11	Range	1900000
12	Minimum	100000
13	Maximum	2000000
14	Sum	3000000
15	Count	5
16		

Media Pesata

- La media pesata di un set di dati è

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 X_1 + w_2 X_2 + \cdots + w_n X_n}{w_1 + w_2 + \cdots + w_n}$$

  - Dove w_i è il peso assegnato alla i^{ma} osservazione

- Usata quando i dati sono già raggruppati in n classi, con w_i valori nella i^{ma} classe

Approssimazioni per Dati Raggruppati

Supponiamo un set di dati contiene i valori $m_1, m_2, \ldots, m_k$, che occorrono con frequenze $f_1, f_2, \ldots f_K$

- Per una popolazione di N osservazioni la media è

$$\boxed{\mu = \frac{\sum_{i=1}^{K} f_i m_i}{N}}$$

dove $N = \sum_{i=1}^{K} f_i$

- Per un campione di n osservazioni, la media è

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}}$$

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

Approssimazioni per Dati Raggruppati

Supponiamo un set di dati contenga i valori $m_1, m_2, \ldots, {}_k$, che occorrono con frequenze $f_1, f_2, \ldots f_K$

- Per una popolazione di N osservazioni la varianza è

$$\sigma^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \mu)^2}{N}$$

- Per un campione di n osservazioni, la varianza è

$$s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n-1}$$

La Covarianza Campionaria

- La covarianza misura la forza della relazione lineare tra **due variabili**

- La covarianza della popolazione:

$$\mathrm{Cov}(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- La covarianza campionaria:

$$\mathrm{Cov}(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

  - Riguarda solo la forza della relazione
  - Non implica un effetto casuale

Interpretazione della Covarianza

- **Covarianza** tra due variabili:

Cov(x,y) > 0 ⟶ x e y tendono a muoversi nella stessa direzione

Cov(x,y) < 0 ⟶ x e y tendono a muoversi in direzioni opposte

Cov(x,y) = 0 ⟶ x e y non mostrano una relazione lineare

Coefficiente di Correlazione

- Misura la forza relativa della relazione lineare tra due variabili

- Coefficiente di correlazione della popolazione:

$$\rho = \frac{\mathrm{Cov}(x, y)}{\sigma_X \sigma_Y}$$

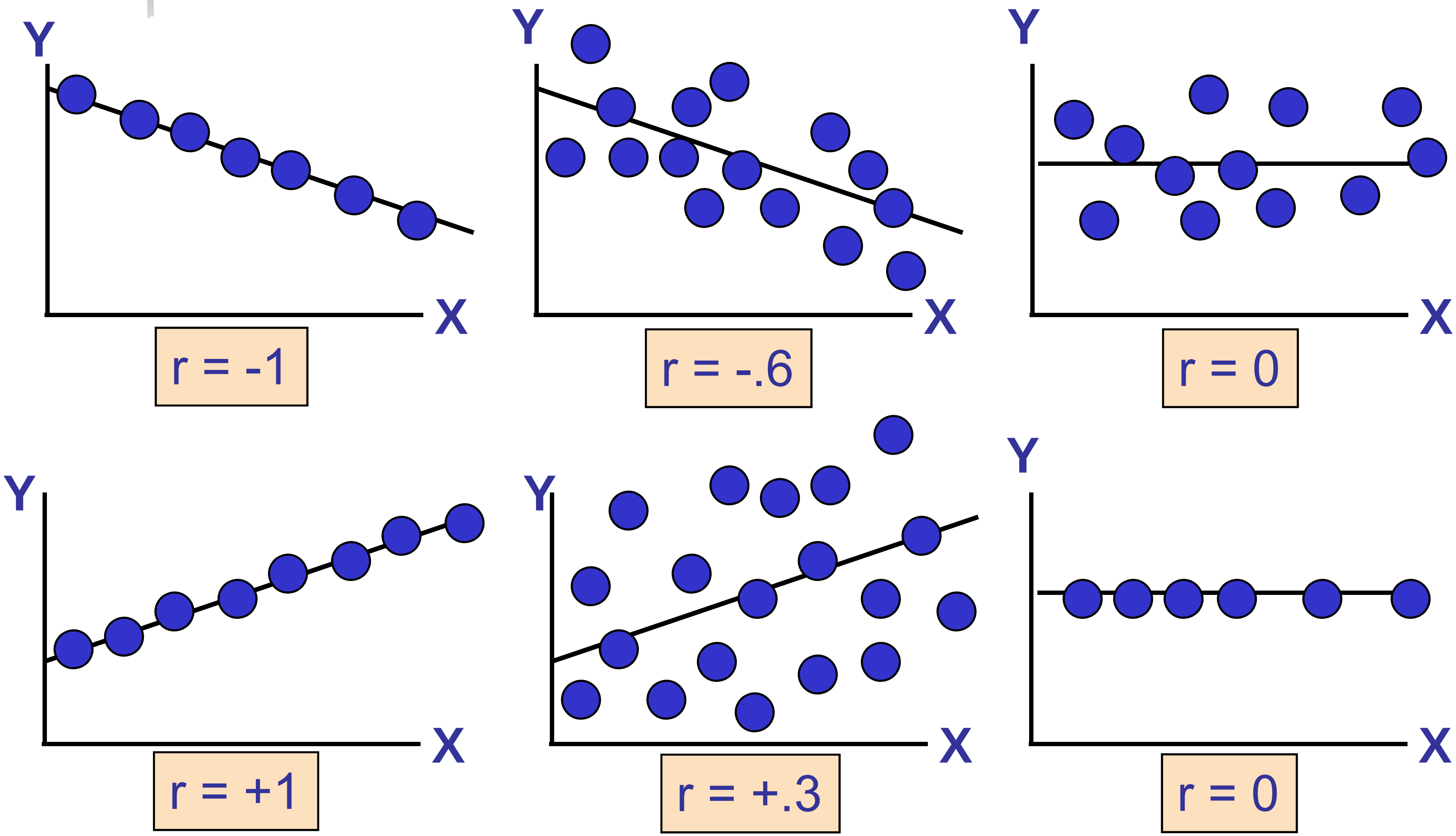
- Coefficiente di correlazione campionario:

$$r = \frac{\mathrm{Cov}(x, y)}{s_X s_Y}$$

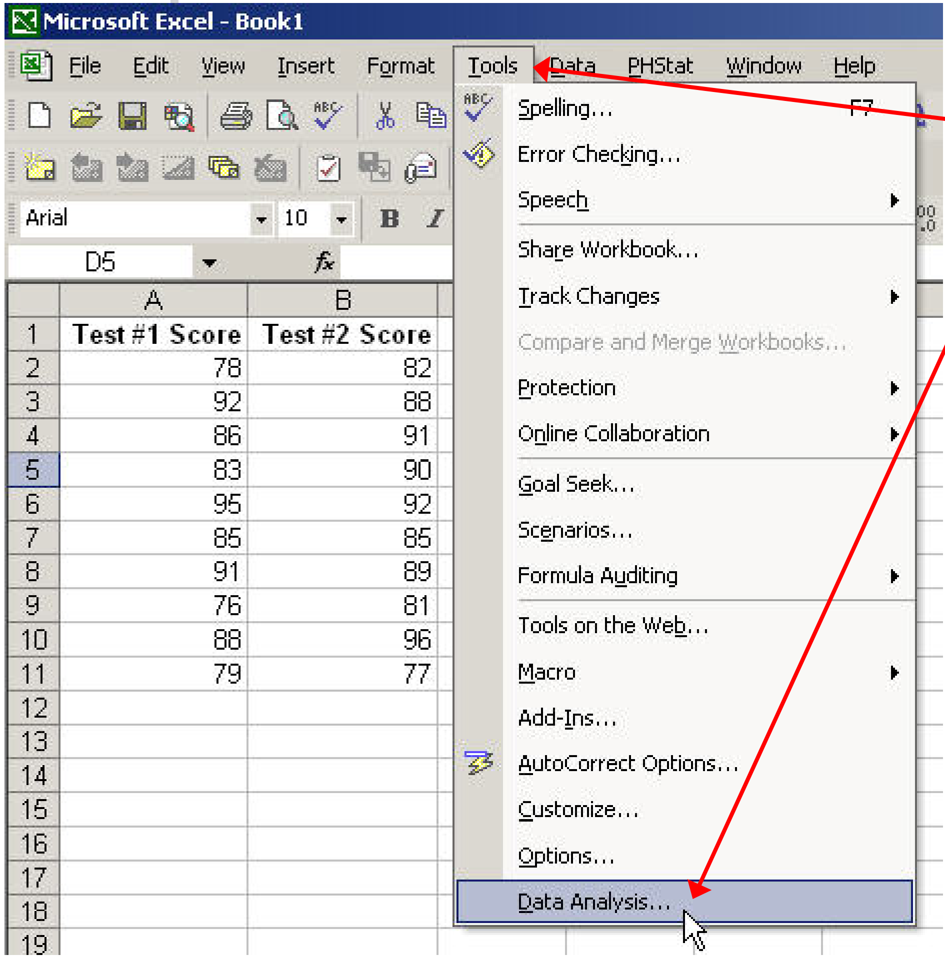
Caratteristiche del Coefficiente di Correlazione, r

- Senza unità di misura
- Campo di variazione fra –1 e 1
- Quanto più è vicino a –1, tanto più è forte la relazione lineare negativa
- Quanto più è vicino a 1, tanto più è forte la relazione lineare positiva
- Quanto più è vicino a 0, tanto più è debole la relazione lineare

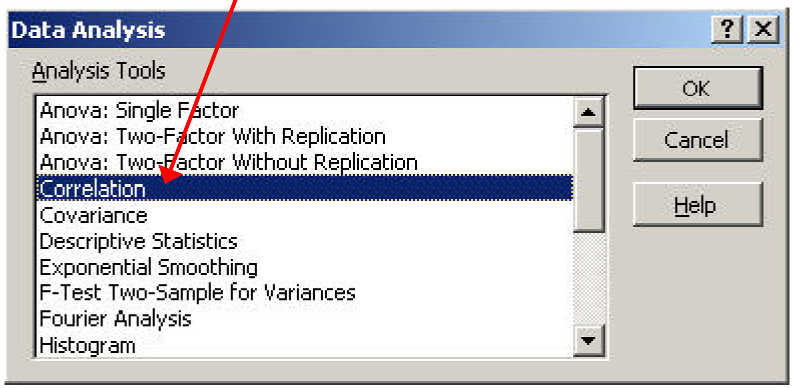
Diagrammi di Dispersione con Vari Coefficienti di Correlazione



Usando Excel per Calcolare il Coefficiente di Correlazione

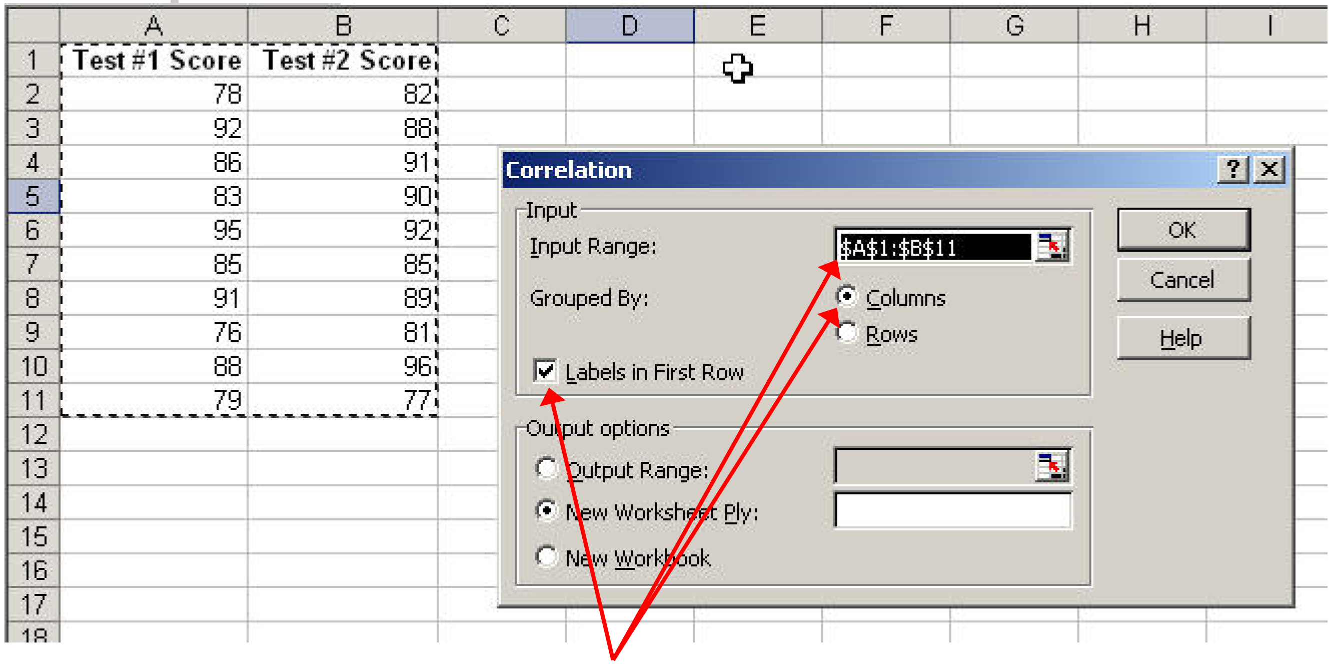


- Selezionare Strumenti/Analisi Dati
- Scegliere Correlazione dal menu a scorrimento
- Cliccare su OK . . .

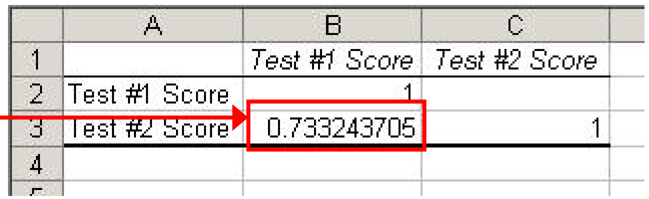


Usando Excel per Calcolare il Coefficiente di Correlazione

(continuazione)



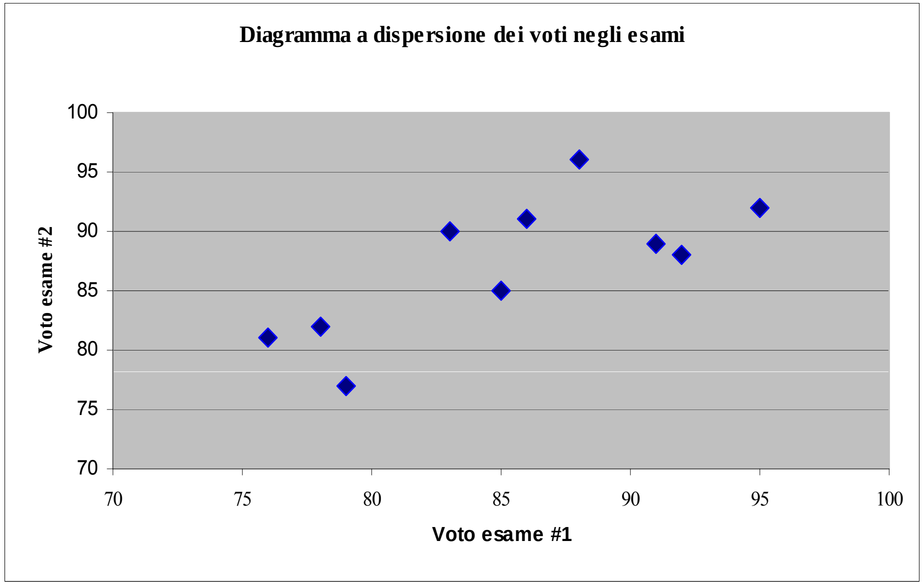
- Inserire le celle contenenti i dati e selezionare le opzioni appropriate
- Cliccare su OK per ottenere l'output



Interpretazione dei Risultati

- r = .733

- Esiste una relazione lineare positiva relativamente forte tra i voti in esame #1 e i voti in esame #2



- Studenti con voti alti nel primo esame tendono ad avere voti alti nel secondo esame

Ottenere Relazioni Lineari

- Un'equazione può essere usata per rappresentare la migliore relazione lineare tra due variabili:

$$Y = \beta_0 + \beta_1 X$$

Dove Y è la variabile dipendente e X è la variabile esplicativa

Regressione con il Metodo dei Minimi Quadrati

- Le stime dei coefficienti β_0 e β_1 vengono calcolate minimizzando la somma dei quadrati dei residui

- La regressione lineare con il metodo dei minimi quadrati, basata sui valori campionati, è

$$\hat{y} = b_0 + b_1 x$$

- Dove b_1 è la pendenza della retta e b_0 è l'ordinata all'origine:

$$b_1 = \frac{\mathrm{Cov}(x,y)}{s_x^2} = r\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Riepilogo del Capitolo

- Si sono descritte le misure di tendenza centrale
  - Media, mediana, moda
- Illustrate la forma della distribuzione
  - Simmetrica, asimmetrica
- Descritte le misure di variabilità
  - Campo di variazione, differenza interquartile, varianza e scarto quadratico medio, coefficiente di variazione
- Discusse le misure per dati raggruppati
- Calcolate le misure delle relazioni tra variabili
  - Covarianza e coefficiente di correlazione