Ulteriori Conoscenze di Informatica e Statistica

Carlo Meneghini

Dip. di fisica – via della Vasca Navale 84,
st. 83 (I piano) tel.: 06 55 17 72 17

meneghini@fis.uniroma3.it

Comportamento in laboratorio

Durante le esercitazioni in laboratorio evitare di:

- utilizzare le macchine del laboratorio di calcolo per navigare in internet consultando pagine e siti non pertinenti al corso o alle esercitazioni
- scaricare, istallare e utilizzare programmi non interenti al corso o alle esercitazioni
- giocare al computer
- cambiare le impostazioni del sistema (es. home page, password di accesso)

Non utilizzare il **desktop** per salvare il proprio lavoro ma **creare la propria directory** e salvare i lavori in opportune sottodirectories.

*Alla fine della lezione effettuare correttamente lo shutdown del sistema e **spegnere** il PC*

Note

I PC a disposizione nel centro di calcolo del Dip. di Fisica sono macchine *dual-boot*, possono cioé utilizzare piú di un sistema operativo. Nel nostro caso W2K e Linux. Dopo il caricamento del BIOS il computer aspetta per alcuni secondi in attesa che voi scegliate il sistema da usare, **Scegliete windows**.

Dopo il caricamento del sistema un finestra vi chiederà il nome utente e la password. Digitando il vostro nome utente e password entrate nel sul desktop di Windows 2000.

Piano della lezione

Statistica descrittiva:

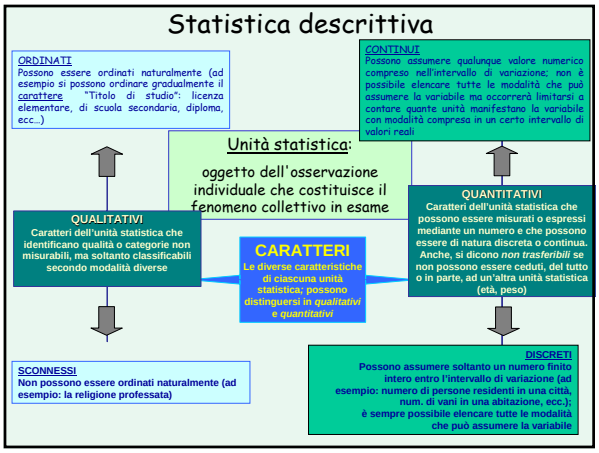
- Strumenti software: il foglio elettronico*

- Le variabili aleatorie: caratteristiche

- distribuzioni di probabilità

***Nota**: programmi freeware (come OpenOffice della Sun) sono del tutto equivalenti e quasi totalmente compatibili con il pacchetto Excel della Microsoft. La versione integrale del pacchetto OpenOffice è gratuita per studenti e istituzioni accademiche.

Statistica descrittiva



La rappresentazione dei dati statistici deve essere organizzata in modo da:

- semplificare i confronti
- sintetizzare i risultati

Esp. 1		Esp. 2	
M	F	M	F
7	8	14	12

Freq. Assoluta

Freq. Relativa

Esp. 1		Esp. 2	
M %	F %	M	F
46.7	53.3	53.8	46.2

Fenomeni

deterministici:

se ripetuti nelle medesime condizioni producono gli stessi risultati

Caduta dei gravi

F = mg

v = mgt

$x = \frac{1}{2} gt^2$

x, v = variabili deterministiche

aleatori:

pur ripetuti nelle medesime condizioni possono produrre risultati differenti

Lancio di dadi

Quale numero?

N = variabile aleatoria

Testa o croce?

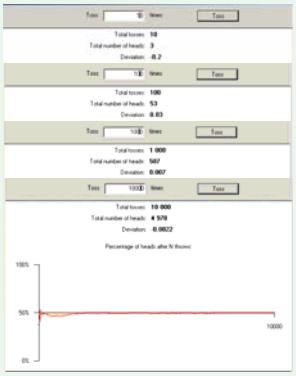
Lancia la moneta

(moneta.exe)

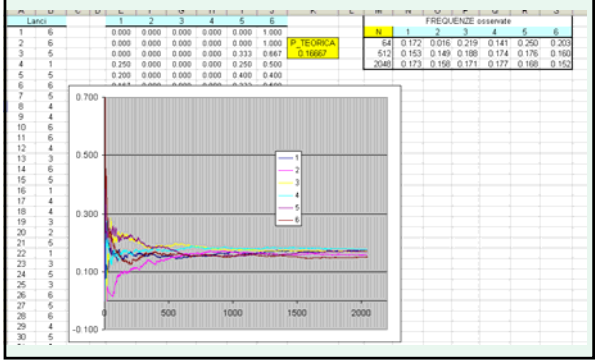
$$p_T = \lim_{N\to\infty} \frac{n_T}{N}$$

$$p_+ = \lim_{N\to\infty} \frac{n_T}{N}$$

$p_T = 0.5 = p_+$



Lancio di un dado



Frequenza:
Variabili discrete

X = variabili aleatoria, N = numero di osservazioni

$x_1, x_2, \ldots x_i$: valori assunti dalla variabile X

$n_1, n_2, \ldots n_i$: numero di volte che si osserva il valore i-esimo x_i

n_i: frequenza assoluta della variabile x_i con: $\sum_{i=1}^{V} n_i = n$

$f_i = \frac{n_i}{N}$ frequenza relativa della variabile x_i

$$\sum_{i=1}^{V} f_i = 1 \qquad f_i \geq 0$$

Frequenza:
Variabili continue

X = variabili aleatoria

x: valori assunti dalla variabile X,

f(x): densità di frequenza della variabile aleatoria X

$$f(x) \geq 0 \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

$f(x_i)$: frequenza relativa nell'intervallo $x_i < x < x_{i+1}$

$$f(x_i) = \int_{x_i}^{x_{i+1}} f(x)dx$$

$$f(x_i) = \frac{\text{numero di osservazioni tra } x_i \text{ e } x_{i+1}}{N}$$

La statistica descrittiva sintetizza l'informazione contenuta nell'insieme dei valori assunti da una variabile aleatoria (distribuzione) utilizzando:

- indici di posizione
- indici di dispersione (variabilità)
- indici di forma
- istogrammi di frequenza
- box plot

Indici di "posizione" (indici di tendenza)

indice	definizione	funzione EXCEL
Media	$\frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} =< x >$ $\quad \frac{\int_A^B f x dx}{\int_A^B dx} = \bar{x} =< x >$	MEDIA(*dati*)
Moda	Valore della variabile cui corrisponde la massima frequenza	MODA(*dati*)
Mediana	Valore della variabile che permette di dividere la distribuzione delle osservazioni in due parti uguali	MEDIANA(*dati*)
Quartili		QUARTILE(*dati;q*)

Indici di "dispersione"

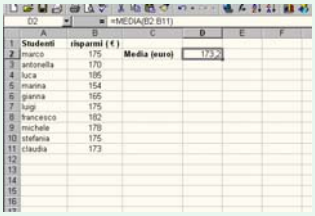
indice	definizione	funzione EXCEL
Varianza	$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} = \sigma^2$	VAR(*dati*)
Deviazione Standard	$\sqrt{\sigma^2} = \sigma$	DEV.ST(*dati*)
Interquartile	Q3-Q1	

Media

La **media** si determina attraverso la funzione **MEDIA** [AVERAGE].

Il risultato di questa funzione è la media aritmetica

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



	Studenti	risparmi (€)		
2	marco	175	Media (euro)	173,2
3	antonella	170		
4	luca	186		
5	marina	154		
6	gianna	165		
7	luigi	175		
8	francesco	182		
9	michele	178		
10	stefania	175		
11	claudia	173		

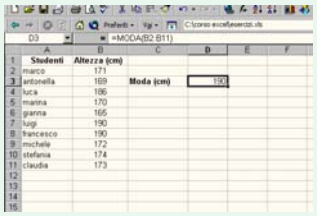
Come si fa:

=MEDIA(B2:B11).

Moda

La **moda** di un collettivo, distribuito secondo un carattere, è la modalità prevalente del carattere ossia quella a cui è associata la massima frequenza.

Si determina mediante la funzione **MODA** [MODE].



	Studenti	Altezza (cm)		
2	marco	171		
3	antonella	169	Moda (cm)	190
4	luca	186		
5	marina	170		
6	gianna	165		
7	luigi	190		
8	francesco	190		
9	michele	172		
10	stefania	174		
11	claudia	173		

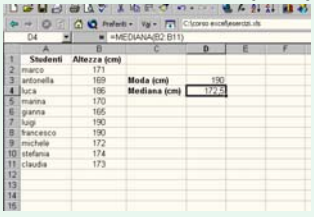
Come si fa:

=MODA (B2:B11)

Mediana

La **mediana** suddivide ogni distribuzione ordinata in due distribuzioni aventi ciascuna una numerosità (o una quantità) che è il 50% della numerosità (o della quantità) della distribuzione totale.

Si determina mediante la funzione **MEDIANA** *[MEDIAN].*



	Studenti	Altezza (cm)		
2	marco	171		
3	antonella	169	Moda (cm)	190
4	luca	186	Mediana (cm)	172,5
5	marina	170		
6	gianna	165		
7	luigi	190		
8	francesco	190		
9	michele	172		
10	stefania	174		
11	claudia	173		

Come si fa:

=MEDIANA (B2:B11).

Quartili

Si puó dividere la distribuzione parti (percentili) contenenti ognuna la *q*-esima parte della quantità della distribuzione totale.

I quantili sono le n parti in cui è stata suddivisa una distribuzione.

per *q* = 4 (più usati) si parla di ***quartili***

I quartili dividono la distribuzione in quattro parti aventi ognuna il 1/4 (25%) della quantità totale;

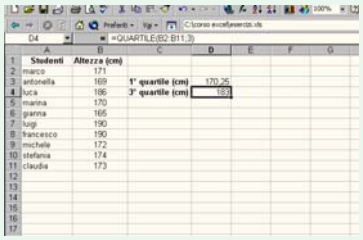
Il **I quartile (Q1)** è il limite superiore della distribuzione che ha il 25% della quantità totale;

Il **II quartile (Q2)** è il limite superiore della seconda distribuzione e quindi da solo separa nella distribuzione totale due distribuzioni che hanno ciascuna il 50% della quantità totale, il Q2 coincide con la mediana;

Il **III quartile (Q3)** è il limite superiore della distribuzione che ha il 75% dell'ammontare della distribuzione totale.

I quartili si determinano mediante le funzioni **QUARTILE** *[QUARTILE]* e **PERCENTILE** *[PERCENTILE]*.

QUARTILE (sequenza di numeri o indirizzo di cella; 0 o 1 o 2 o 3 o 4) {0 =minimo; 1 = 1° quartile; 2 = mediana; 3 = 3° quartile; 4 = massimo}



Come si fa:

Q1:

=QUARTILE (B2:B11,1).

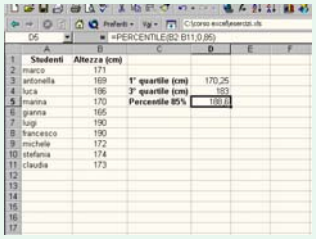
Q2:

=QUARTILE (B2:B11,2).

Q3:

=QUARTILE (B2:B11,3).

PERCENTILE (sequenza di numeri o indirizzo di cella; numero compreso tra 0 ed 1) {percentile p%: inserire il numero p}



Come si fa:

- Spostare il cursore nella cella C5 e digitare: Percentile 85% (cm)
- Spostare il cursore nella cella D5 e inserire la funzione: **=PERCENTILE (B2:B11,0.85)**.

Indicatori di variabilità (dispersione)

Misurano la dispersione dei valori di una distribuzione

Varianza

Deviazione standard

Ampiezza

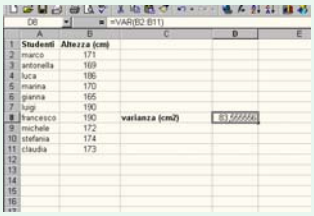
Interquartile

Varianza

La varianza si determina attraverso la funzione **VAR** *[VAR]*.

Il risultato di questa funzione è la varianza campionaria (s^2) dei valori introdotti come argomento

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



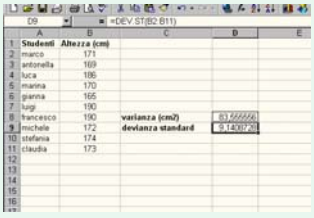
Come si fa:

=VAR(B2:B11)

Deviazione standard

La **deviazione standard** si determina attraverso la funzione **DEV.ST** *[STDEV]*.

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



=DEV.ST(B2:B11)

Ampiezza del campione

si ottiene come differenza tra l'estremo superiore e quello inferiore dei valori osservati del campione.

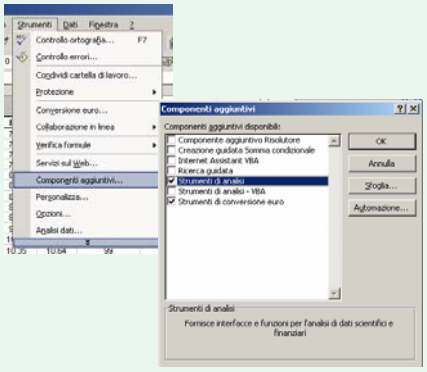
= **MAX(*dati*) - MIN(*dati*)**.

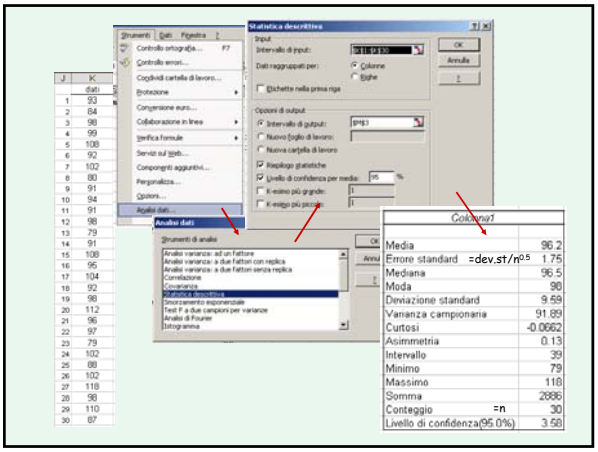
Ampiezza interquartile

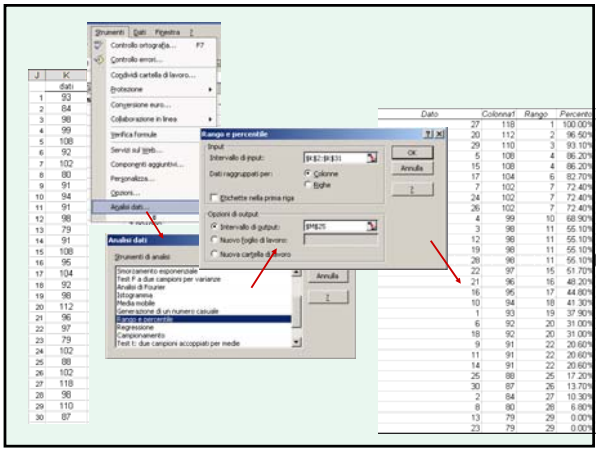
si ottiene come differenza tra il terzo e ilprimo quartile

= **quartile(*dati, 3*) - quartile(*dati, 1*)**.

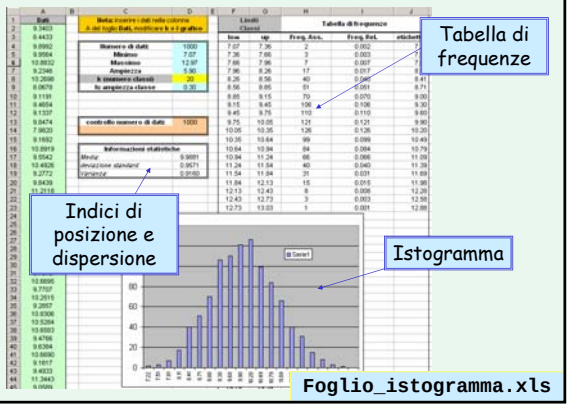
Componenti Aggiuntivi di Excel



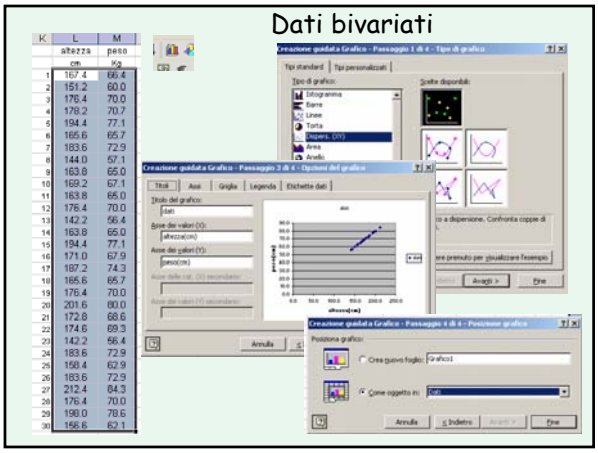


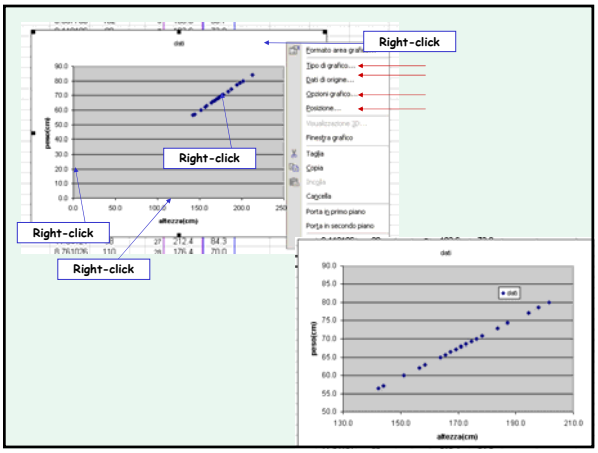


Istogrammi di frequenza e indici statistici



Dati bivariati





Coefficiente di correlazione lineare

Misura la correlazione tra due variabili. In Excel si usa la funzione **CORRELAZIONE** *[CORREL]*.
= correlazione (dati_x, dati_y)

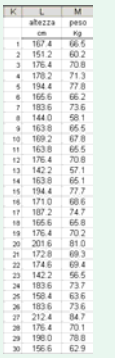
Il risultato di questa funzione è il coefficiente di correlazione (r) tra i due insiemi di valori:

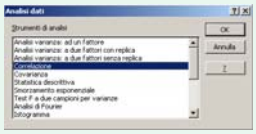
$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

dove:

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

è *n* volte la covarianza fra X e Y.





	Colonna 1	*Colonna 2*
Colonna 1	1	
Colonna 2	0.99918	1

Box plot

E' una "scatola" in cui

- I bordi corrispondono a Q1 e Q3
- Una linea fra di essi indica il valore di Q2 (mediana)
- All'esterno vengono aggiunti:
  - Un "baffo superiore" = distanza da Q3 del più grande valore inferiore a Q3+1.5(Q3-Q1)
  - Un "baffo inferiore" = distanza da Q1 del più piccolo valore minore di Q1 ma maggiore di Q1-1.5(Q3-Q1)
- I valori esterni all'intervallo compreso tra i due "baffi", detti "outliers", vengono rappresentati individualmenti

Box plot