Laboratorio di Informatica

Dr Carlo Meneghini

Dip. di Fisica "E. Amaldi"
via della Vasca Navale 84
st. – 83 – I piano

meneghini@fis.uniroma3.it

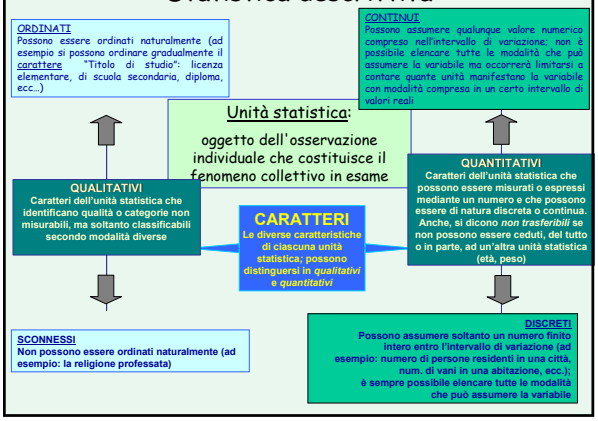
tel.: 06 55177217

http://www.fis.uniroma3.it/~meneghini

Esercitazione V

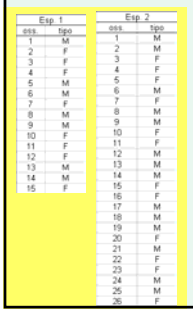
Statistica descrittiva e istogrammi

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:

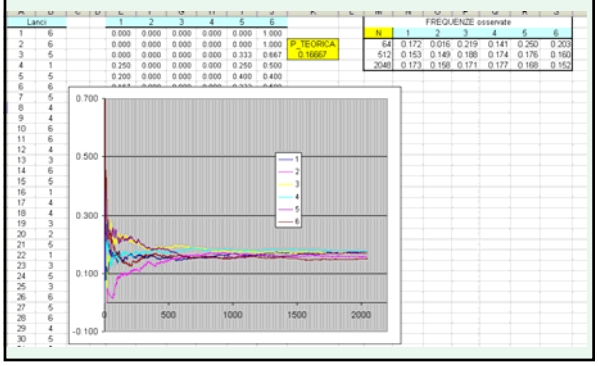
pur ripetuti nelle **medesime** condizioni possono produrre risultati differenti

Lancio di dadi

Quale numero?

N = variabile aleatoria

Lancio di un dado



Frequenza:
Variabili discrete

X = variabili aleatoria, n = numero di osservazioni

$x_1, x_2, \ldots x_i$: valori discreti 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

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

per N molto grande, f_i rappresenta una definizione operativa di probabilità di osservare il valore x_i

Frequenza:

Variabili continue

X = variabili aleatoria

x: valori (continui) assunti dalla variabile X,

$\rho(x)$: numero di osservazioni (frequenza assoluta) nell'intervallo [x , x+dx]

f(x): densità di frequenza della variabile aleatoria X, ovvero frazione di osservazioni nell'intervallo [x , x+dx]

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

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

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

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

Distribuzioni integrate:

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

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

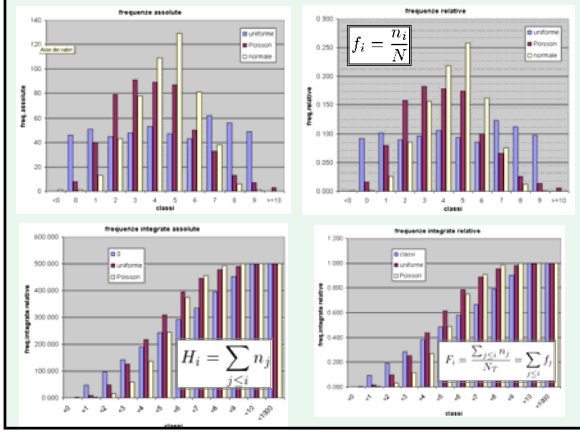
$$f_i = \frac{n_i}{N} \qquad \sum_{i=1}^{V} f_i = 1$$

$$H_i = \sum_{j \leq i} n_j$$

frequenza integrata assoluta: numero di volte in cui si osserva un valore minore o uguale a x_i

$$F_i = \frac{\sum_{j \leq i} n_j}{N_T} = \sum_{j \leq i} f_j$$

frequenza integrata relativa: frazione di volte in cui si osserva un valore minore o uguale a x_i



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	$\frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} = < x >$	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	Valori della variabile dividono la distribuzione in quarti	QUARTILE(*dati;q*)
Percentili	Valori della variabile dividono la distribuzione in percentili	Percentile(dati;k)
Frequenze integrate	$H_i = \sum_{j \leq i} n_j$ $F_i = \frac{\sum_{j \leq i} n_j}{N_T} = \sum_{j \leq i} f_j$	

Indici di "dispersione"

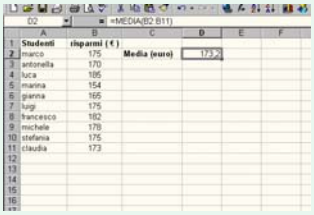
indice	definizione	funzione EXCEL
Varianza	$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} = \sigma^2$	VAR(*dati*)
Deviazione Standard	$\sqrt{\sigma^2} = \sigma$	DEV.ST(*dati*)
Interquartile	Q3-Q1	

Media

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

Il risultato di questa funzione è la media aritmetica

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$$



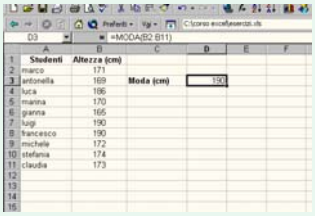
Come si fa:

=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].



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]*.

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;

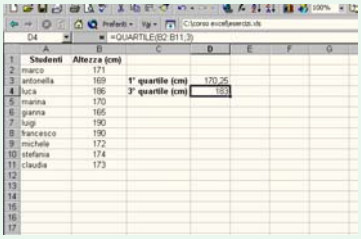
Il **I quartile (Q1)** è il limite superiore della distribuzione che ha il 25% della quantità totale;

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).

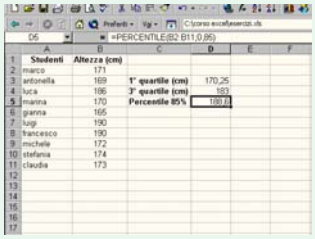
Q2:

=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

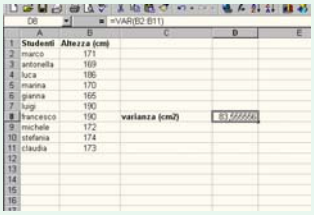
- Varianza
- Deviazione standard
- Ampiezza
- Interquartile

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

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



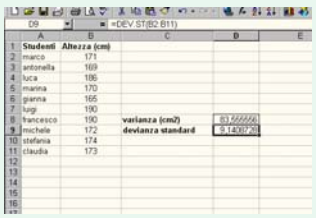
Come si fa:

=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.

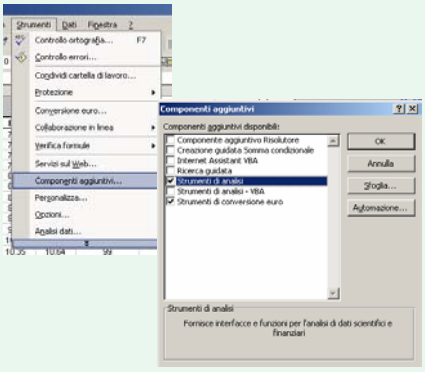
= **MAX(*dati*) - MIN(*dati*)**.

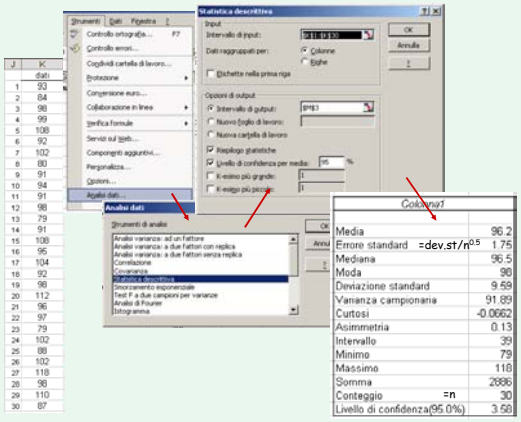
Ampiezza interquartile

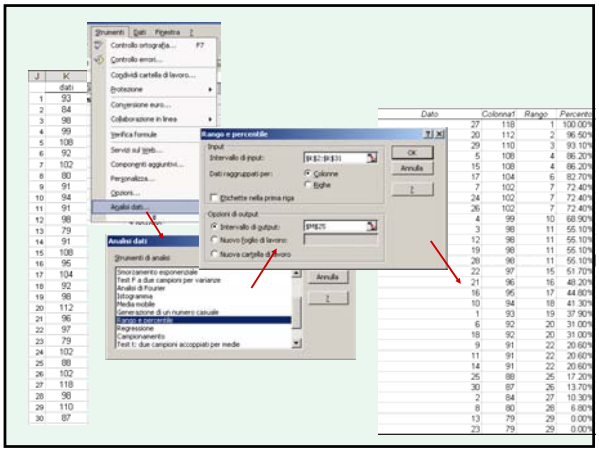
si ottiene come differenza tra il terzo e ilprimo quartile

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

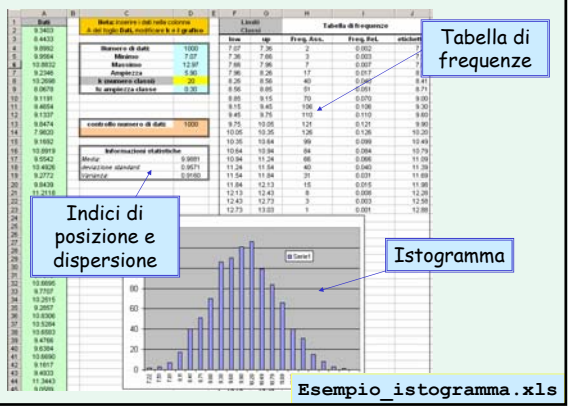
Componenti Aggiuntivi di Excel







Istogrammi di frequenza e indici statistici

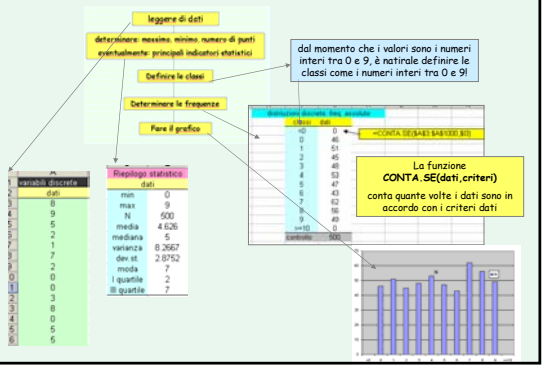


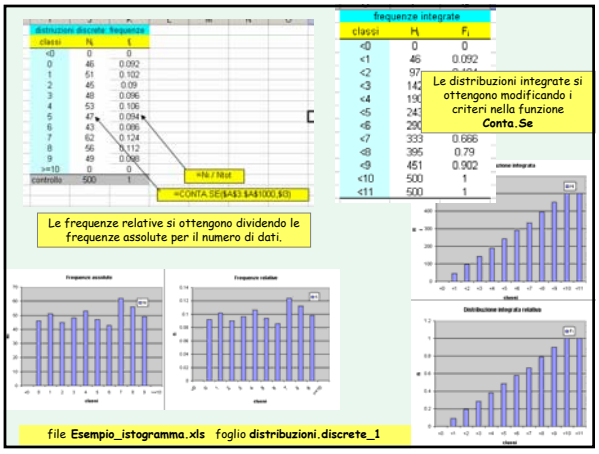
Esempio_istogramma.xls

Caloco dell'istogramma di frequenze

- leggere di dati
- determinare: massimo, minimo, numero di punti
  eventualmente: principali indicatori statistici
- Definire le classi
- Determinare le frequenze
- Fare il grafico

Caso semplice: dati discreti, una classe per ogni valore osservato

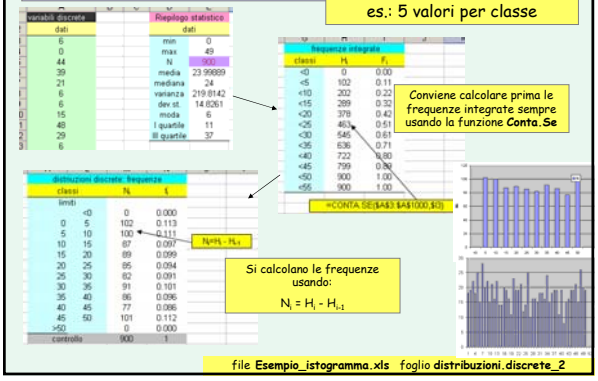




file **Esempio_istogramma.xls** foglio **distribuzioni.discrete_1**

Caso meno semplice:
dati discreti, una classe contiene più valori

es.: 5 valori per classe



file **Esempio_istogramma.xls** foglio **distribuzioni.discrete_2**

La funzione **FREQUENZA(Dati,classe)** calcola il numero di ricorrenze di dati con valore minore o eguale al valore "classe". E preferibile trattando con numeri, numeri reali in particolare.

La funzione FREQUENZA calcola le frequenze assolute integrate.

Si procede come prima per il calcolo delle frequenze: prima calcolando la distribuzione integrata e poi le frequenze per differenza.

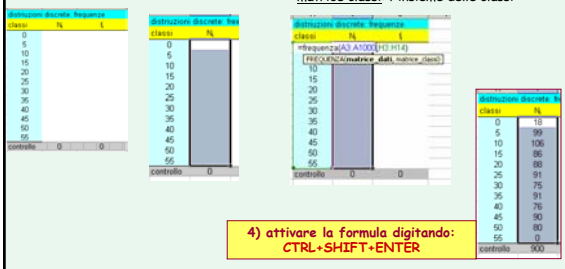
file **Esempio_istogramma.xls** foglio **distribuzioni.discrete_3**

Si possono calcolare le frequenze (assolute) senza passare per le frequenze integrate immettendo le **formule in forma di matrice**

1) definire le classi

2) selezionare il gruppo di celle in cui si vuole effettuare il calcolo

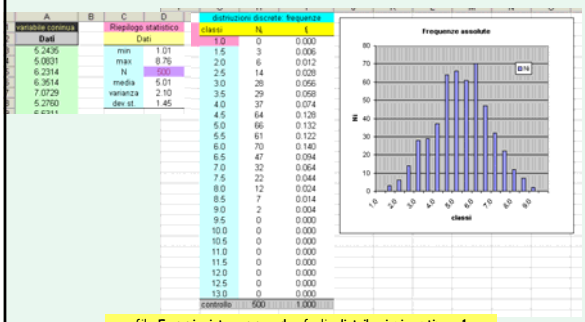
3) =frequenza(**matrice dati,matrice classi**)
matrice dati: l'insieme dei dati
matrice classi: l'insieme delle classi



4) attivare la formula digitando: CTRL+SHIFT+ENTER

file **Esempio_istogramma.xls** foglio **distribuzioni.discrete_4**

Utilizzando lo stesso schema mostrato in precedenza si possono calcolare e graficare gli istogrammi per distribuzioni continue.



file **Esempio_istogramma.xls** foglio **distribuzioni.continue_1**

Esercizio

Il file ASCII **distribuzioni_discrete.dat** è un file 3 colonne che riporta tre set di dati, a variabili discrete, rispettivamente con distribuzione Uniforme, di Poisson e Normale (Gauss).

1) riportare i principali indicatori statistici per le tre distribuzioni

2) calcolare gli istogrammi di frequenza (assolute, relative, integrate) per le tre distribuzioni

file **Statistica_distribizioni.xls** foglio **distribuzioni discrete**

Esercizio

Il file ASCII **distribuzioni_continue.dat** è un file 3 colonne che riporta tre set di dati distribuito in modo continuo rispettivamente con distribuzione Uniforme e Normali (Gauss).

1) riportare i principali indicatori statistici per le tre distribuzioni

2) calcolare gli istogrammi di frequenza (assolute, relative, integrate) per le tre distribuzioni

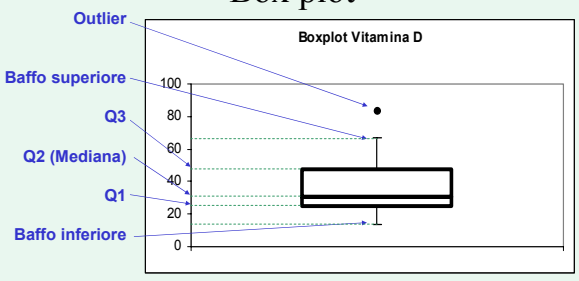
file **Statistica_distribizioni.xls** foglio **distribuzioni continue**

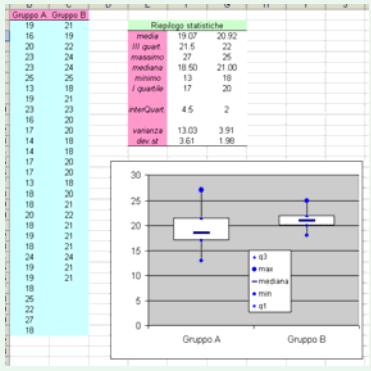
Box plot: metodo grafico per rappresentare le informazioni statistiche di un dato. Usato frequentemente in campo medico/biologico

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) *Q3-Q1= interq.*
  - 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 individualmente

Box plot


Outlier
Baffo superiore
Q3
Q2 (Mediana)
Q1
Baffo inferiore
Boxplot Vitamina D
100
80
60
40
20
0


Gruppo A
Gruppo B
Riepilogo statistiche
media 19.07 20.92
III quart 21.5 22
massimo 27 25
mediana 18.50 21.00
minimo 13 18
I quartile 17 20
interQuart 4.5 2
varianza 13.03 3.91
dev st 3.61 1.98
30
25
20
15
10
5
0
q3
max
mediana
min
q1
Gruppo A
Gruppo B