

Laboratorio di Informatica

Dr Carlo Meneghini

Dip. di Fisica "E. Amaldi"
via della Vasca Navale 84
st. - 83 - 1 piano

meneghini@fis.uniroma3.it

tel.: 06 55177217

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

Esercitazione V Statistica descrittiva e istogrammi

Statistica descrittiva

ORDINATI
Possono essere ordinati naturalmente (ad esempio si possono ordinare gradualmente il **carattere** "Titolo di studio": licenza elementare, di scuola secondaria, diploma, ecc...)

CONTINUE
Possono assumere qualunque valore numerico compreso nell'intervallo di variazione; non è possibile elencare tutte le modalità che può assumere la variabile ma occorrono limitarsi a contare quante unità manifestano la variabile con modalità compresa in un certo intervallo di valori reali

Unità statistica:
oggetto dell'osservazione individuale che costituisce il fenomeno collettivo in esame

QUALITATIVI
Caratteri dell'unità statistica che identificano qualità o categorie non misurabili, ma soltanto classificabili secondo modalità diverse

QUANTITATIVI
Caratteri dell'unità statistica che possono essere misurati o espressi mediante un numero e che possono essere di natura discreta o continua. Anche, si dicono *non trasferibili* se non possono essere ceduti, del tutto o in parte, ad un'altra unità statistica (età, peso)

CARATTERI
Le diverse caratteristiche di ciascuna unità statistica; possono distinguersi in *qualitativi* e *quantitativi*

SCONNESSI
Non possono essere ordinati naturalmente (ad esempio: la religione professata)

DISCRETI
Possono assumere soltanto un numero finito intero entro l'intervallo di variazione (ad esempio: numero di persone residenti in una città, num. di vani in una abitazione, ecc.); è sempre possibile elencare tutte le modalità che può assumere la variabile.

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

- **semplificare i confronti**
- **sintetizzare i risultati**

Esp. 1		Esp. 2	
età	tipo	età	tipo
1	M	1	M
2	F	2	M
3	F	3	F
4	F	4	F
5	M	5	F
6	M	6	M
7	F	7	M
8	M	8	M
9	M	9	M
10	F	10	F
11	F	11	F
12	F	12	M
13	M	13	M
14	M	14	M
15	F	15	F
16	F	16	F
17	M	17	M
18	M	18	M
19	M	19	M
20	F	20	F
21	M	21	M
22	F	22	F
23	F	23	F
24	M	24	M
25	M	25	M
26	F	26	F

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

aleatori:
pur ripetuti nelle medesime condizioni possono produrre risultati differenti



Caduta dei gravi

$$F = mg$$

$$v = mgt$$

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

x, v = variabili deterministiche



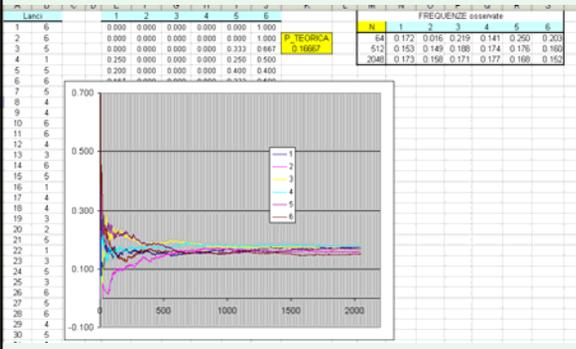
Quale numero?

Lancio di dadi

N = variabile aleatoria



Lancio di un dado



Frequenza: Variabili discrete

X = variabili aleatoria, n = numero di osservazioni

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

n_1, n_2, \dots, 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

$$\sum_{i=1}^V f_i = 1$$

$$f_i \geq 0$$

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 $[x, x+dx]$

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

$f(x_i)$: 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, \dots, x_i : valori discreti assunti dalla variabile X
 n_1, n_2, \dots, n_i : numero di volte che si osserva il valore i-esimo x_i

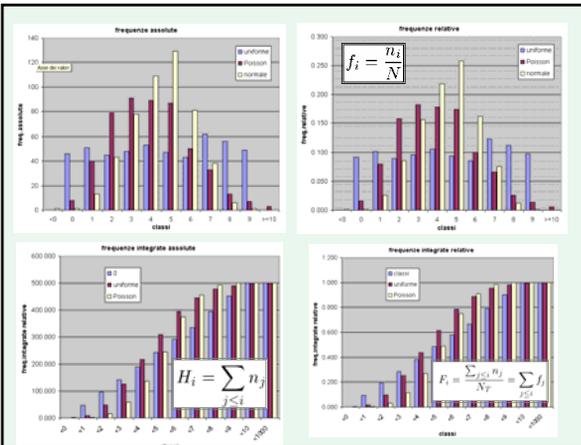
$$f_i = \frac{n_i}{N} \quad \sum_{i=1}^V f_i = 1$$

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

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	$\frac{1}{N} \sum_{i=1}^N x_i = \bar{x} = \langle x \rangle$	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

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

Studenti	Altezza (cm)	Media (cm)
marco	175	175.2
antonella	170	
luca	166	
marina	164	
gianna	165	
lugi	175	
francesco	182	
michele	178	
stefania	176	
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)	Moda (cm)
marco	171	170
antonella	169	
luca	166	
marina	170	
gianna	165	
lugi	190	
francesco	190	
michele	172	
stefania	174	
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)	Mediana (cm)
marco	171	172
antonella	169	
luca	166	
marina	170	
gianna	165	
lugi	190	
francesco	190	
michele	172	
stefania	174	
claudia	173	

Come si fa:

=MEDIANA (B2:B11).

Quartili

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

I **quartili** 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)

Studenti	Altezza (cm)	1° quartile (cm)	3° quartile (cm)
marco	171	170.25	182
antonella	169		
luca	166		
marina	170		
gianna	165		
lugi	190		
francesco	190		
michele	172		
stefania	174		
claudia	173		

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)

Studenti	Altezza (cm)	1° quartile (cm)	Percentile 85%
marco	171	170.25	182
antonella	169		
luca	166		
marina	170		
gianna	165		
lugi	190		
francesco	190		
michele	172		
stefania	174		
claudia	173		

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

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

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

	varianza (cm²)
=VAR(B2:B11)	8,12222222

Come si fa:
=VAR(B2:B11)

Deviazione standard

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

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

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

	devianza standard
=DEV.ST(B2:B11)	2,8500725

=DEV.ST(B2:B11)

Ampiezza del campione

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

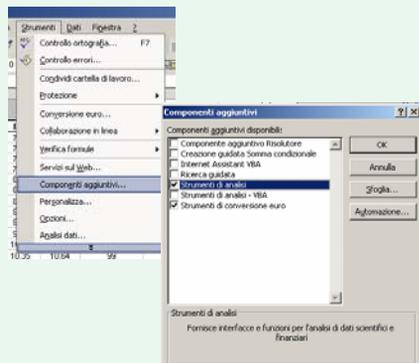
= MAX(dati) - MIN(dati).

Ampiezza interquartile

si ottiene come differenza tra il terzo e il primo quartile

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

Componenti Aggiuntivi di Excel



Strumenti di analisi	Statistica descrittiva
Analisi varianza ad un fattore	Media
Analisi varianza a due fattori con replica	Errore standard =dev.st/n ^{0,5}
Analisi varianza a due fattori senza replica	Mediana
Correlazione	Moda
Covarianza	Deviazione standard
Stima crescita esponenziale	Varianza campionaria
Test F a due campioni per varianza	Curtosi
Analisi di Fourier	Asimmetria
Diagrammi	Intervallo
	Minimo
	Massimo
	Somma
	Conteggio
	Livello di confidenza(95.0%)

Istogrammi di frequenza e indici statistici

Tabella di frequenze

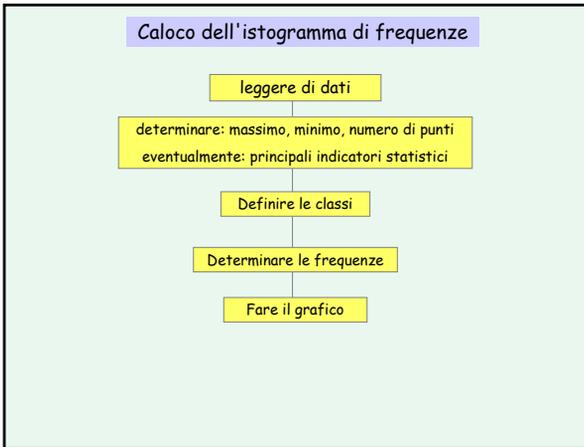
Classe	Min	Max	Freq. Abs.	Freq. Rel.	Media	Dev. St.
1	0	10	1	0.01	5	3
2	10	20	2	0.02	15	4
3	20	30	3	0.03	25	5
4	30	40	4	0.04	35	6
5	40	50	5	0.05	45	7
6	50	60	6	0.06	55	8
7	60	70	7	0.07	65	9
8	70	80	8	0.08	75	10
9	80	90	9	0.09	85	11
10	90	100	10	0.10	95	12

Indici di posizione e dispersione

Indice	Valore
Media	50.000
Deviazione standard	10.000
Varianza	100.000
Coeficiente di variazione	0.200
Skewness	0.000
Kurtosis	0.000

Istogramma

Esempio_istogramma.xls



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

leggere di dati
determinare: massimo, minimo, numero di punti
eventualmente: principali indicatori statistici

Definire le classi
dal momento che i valori sono i numeri interi tra 0 e 9, è naturale definire le classi come i numeri interi tra 0 e 9!

Determinare le frequenze
Fare il grafico

Riepilogo statistico

variabile discreta	dati
min	0
max	9
N	500
media	4.626
mediana	5
varianza	8.2667
dev. st.	2.8752
moda	7
I quartile	2
II quartile	7

Conti
La funzione **CONTA.SE(dati,criteri)** conta quante volte i dati sono in accordo con i criteri dati

Le frequenze relative si ottengono dividendo le frequenze assolute per il numero di dati.

Le distribuzioni integrate si ottengono modificando i criteri nella funzione ContA.Se

file Esempio_istogramma.xls foglio distribuzioni_discrete_1

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

es.: 5 valori per classe

Conviene calcolare prima le frequenze integrate sempre usando la funzione ContA.Se

Si calcolano le frequenze usando: $N_i = H_i - H_{i-1}$

file Esempio_istogramma.xls foglio distribuzioni_discrete_2

La funzione FREQUENZA(Dati, classe) calcola il numero di ricorrenze di dati con valore minore o uguale 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_distribuzioni.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

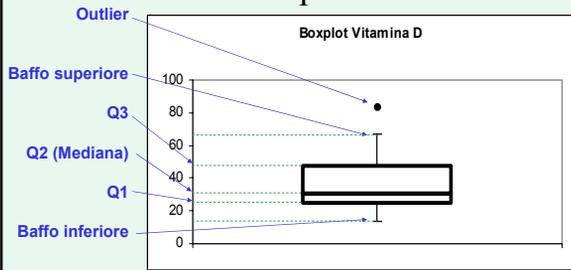
file Statistica_distribuzioni.xls foglio distribuzioni continue

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 $Q3 + 1.5(Q3 - Q1)$ $Q3 - Q1 = \text{interq.}$
 - 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



Gruppo A		Gruppo B	
19	21	Riepilogo statistiche	
16	19	media	19,07 20,92
20	22	il3 quart	21,5 22
23	24	massimo	27 25
23	24	mediana	18,50 21,00
25	25	minimo	13 18
13	18	il quartile	17 20
19	21	interQuart	4,5 2
16	20	varianza	13,03 3,91
17	20	dev.st	3,61 1,98
14	18		
14	18		
17	20		
17	20		
13	18		
18	20		
18	21		
20	22		
18	21		
19	21		
19	21		
24	24		
19	21		
18	20		
25	25		
22	27		
18			

