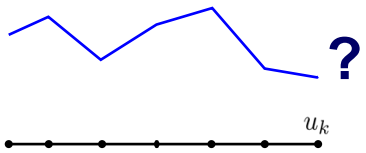
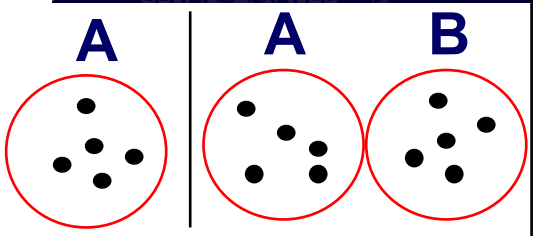
Statistica Descrittiva ed Inferenziale

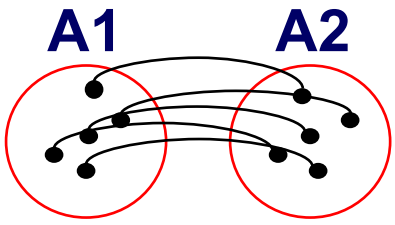
Why Statistics?



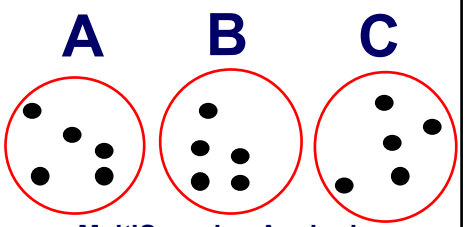
Description and Prediction



Samples Analysis



Paired Samples Analysis



MultiSamples Analysis

Why Statistics?

Formal definition of Probability

A probability measure P on the countable sample space Ω is a set function

$$P : \mathcal{F} \rightarrow [0, 1],$$

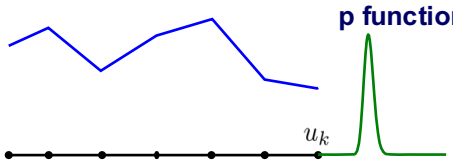
σ-field

satisfying the following conditions

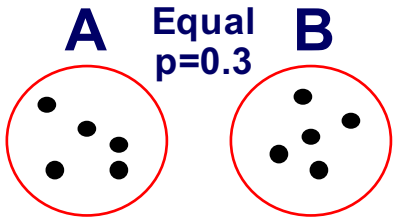
- $P(\Omega) = 1.$
- $P(\omega_i) = p_i.$
- *If* $A_1, A_2, A_3, \ldots \in \mathcal{F}$ *are mutually disjoint, then*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

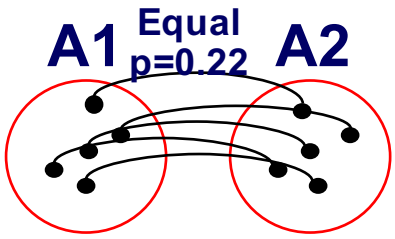
Why Statistics?



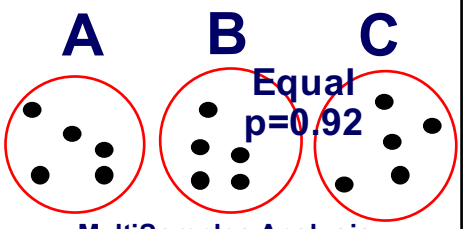
Description and Prediction



Samples Analysis



Samples Analysis



MultiSamples Analysis

Introduzione (1)

Di cosa si occupa la statistica?

Oggetto della statistica sono ***fenomeni collettivi*** che presentano carattere di ***variabilità***

Per fenomeno collettivo si intende un fenomeno che riguarda una grande collezione di elementi = **POPOLAZIONE**

Elementi della popolazione = **UNITÀ STATISTICHE**

Introduzione (2)

La popolazione è troppo vasta per poter essere studiata nella sua globalità → dalla popolazione viene estratto un **CAMPIONE** di n elementi

Sul campione vengono rilevate/misurate alcune caratteristiche. I risultati di questa misura costituiscono i **DATI**

La statistica permette di trarre conclusioni sull'intera popolazione a partire dai dati ottenuti sul campione

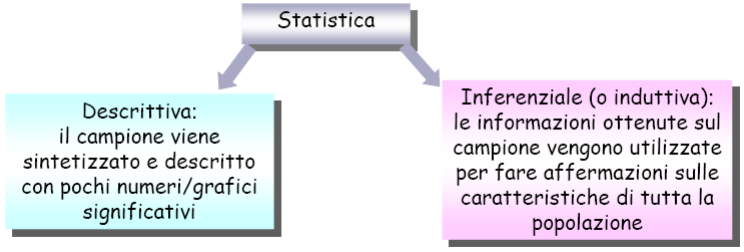
A causa della casistica ridotta non possiamo essere certi delle nostre conclusioni→ si specifica il grado di certezza in termini di probabilità

Statistica Descrittiva e Statistica Inferenziale (1)

Scopo principale della statistica consiste nel compiere un'inferenza circa l'intera popolazione a partire dal campione

Per fare questo, per prima cosa, bisogna descrivere e sintetizzare i dati, con pochi numeri o grafici significativi

Si distinguono due grandi rami della statistica:



Statistica Descrittiva e Statistica Inferenziale (2)

Nelle applicazioni, Statistica Descrittiva e Statistica Inferenziale non possono essere completamente separate.

Infatti i problemi di inferenza statistica vengono affrontati secondo il seguente schema:

Statistica Descrittiva di un CAMPIONE (SAMPLE) di dati

Statistica Descrittiva

Scopo: descrivere il campione (dati) in modo sintetico ed efficace mediante tabelle, grafici, numeri

Premessa: Le caratteristiche che osserviamo sul campione variano da un'unità di osservazione all'altra → variabili

Le variabili possono essere **discrete** o **continue**

Variabili discrete: assumono un numero finito o un'infinità numerabile di valori

Variabili continue: possono assumere qualsiasi valore

Tabelle e Grafici di Frequenza (1)

Un primo utile sistema per riassumere i dati è la costruzione di tabelle e grafici di frequenza

Esempio nel discreto: lancio di un dado

Il risultato del lancio è una variabile discreta (può assumere uno dei seguenti valori 1 2 3 4 5 6)

50 lanci → ottengo una sequenza di 50 (n) numeri

Sintetizzo i dati costruendo una tabella

Tabelle e Grafici di Frequenza (2)

Esempio nel discreto *(continua)*

risultato del lancio	frequenza (f)	frequenza relativa (f/n)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.20
6	5	0.10
	$\sum f = n$	$\sum (f/n) = 1.00$

prima colonna: possibili risultati del lancio

seconda colonna: numero totale di volte in cui è stato ottenuto quel risultato (**frequenza assoluta**)

terza colonna: frequenza assoluta del risultato divisa per il numero totale (n) di osservazioni (**frequenza relativa**)

Tabelle e Grafici di Frequenza (3)

Esempio nel discreto *(continua)*

risultato del lancio	frequenza (f)	frequenza relativa (f/n)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.20
6	5	0.10

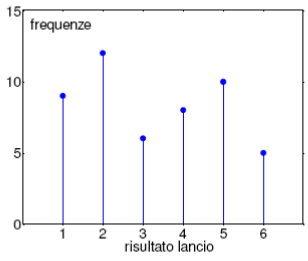
seconda colonna: **distribuzione di frequenze**

terza colonna: **distribuzione di frequenze relative**

Tabelle e Grafici di Frequenza (4)

Esempio nel discreto *(continua)*

La distribuzione di frequenza e la distribuzione di frequenza relativa possono essere rappresentate graficamente mediante **diagrammi a bastoncino**



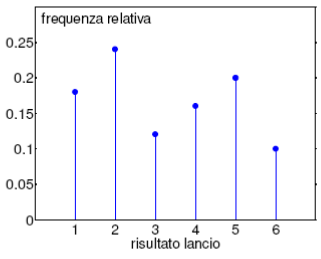


Tabelle e Grafici di Frequenza (5)

Esempio nel continuo

Campione di 200 uomini (n=200) estratto da una certa popolazione. Rileviamo l'altezza in cm.

La variabile osservata è continua. Non ha senso parlare di frequenza del singolo valore poiché non c'è alcuna possibilità di osservare due stature esattamente uguali.

L'intervallo che contiene tutti i valori osservati viene suddiviso in un certo numero di sottointervalli (classi) e si contano quante osservazioni cadono nei diversi sottointervalli.

Tabelle e Grafici di Frequenza (6)

Esempio nel continuo (continua)

Limiti intervallo (cm)	Valore centrale (cm)	frequenza (*f*)	frequenza relativa (*f*/*n*)
141.5-148.5	145	2	0.01
148.5-155.5	152	7	0.035
155.5-162.5	159	22	0.11
162.5-169.5	166	13	0.065
169.5-176.5	173	44	0.22
176.5-183.5	180	36	0.18
183.5-190.5	187	32	0.16
190.5-197.5	194	13	0.065
197.5-204.5	201	21	0.105
204.5-211.5	208	10	0.05

$\sum f = n$ $\quad$ $\sum (f/n) = 1.00$

Tabelle e Grafici di Frequenza (7)

Quante classi (sottointervalli) vi devono essere?

- compromesso ragionevole tra una distribuzione troppo dettagliata ed una troppo sintetica

Le classi vengono in genere scelte in modo che il valore centrale sia un numero intero

In quale classe viene posizionata una osservazione che cade al limite tra due classi?

In genere la si pone nella classe superiore

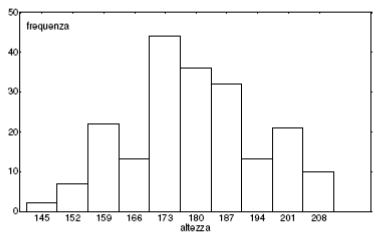
Ad esempio il valore 162.5 viene posto nella classe 162.5-169.5.

Si tratta quindi di sottointervalli del tipo [ )

Tabelle e Grafici di Frequenza (8)

Esempio nel continuo (continua)

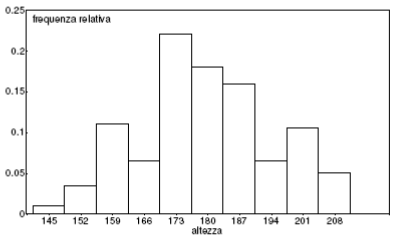
I dati raggruppati in sottointervalli possono essere rappresentati graficamente mediante **istogrammi**



istogramma della distribuzione di frequenze

base = ampiezza della classe

altezza = frequenza assoluta

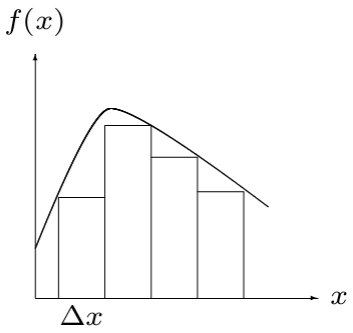


istogramma della distribuzione di frequenze relative

base = ampiezza della classe

altezza = frequenza relativa

Obiettivo: Descrizione di un Istogramma



Indici di descrizione statistica di un campione

INDICI

- di posizione
  - MODA
  - MEDIANA
  - MEDIA
- di dispersione
  - SCARTO QUADRATICO MEDIO
  - VARIANZA
  - RANGE
- di forma
  - ASIMMETRIA (SKEWNESS)
  - CURTOSI (KURTOSIS)

Misure di posizione : La Media Aritmetica (1)

La media aritmetica (m) è la più comune misura di posizione

Le osservazioni (x_1, x_2,...x_n) vengono sommate tra di loro, quindi la somma divisa per n (cioè per il numero di osservazioni):

$$m = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Per calcolare l'altezza media del nostro campione di 200 individui dobbiamo sommare le 200 osservazioni e dividere la somma per 200

Centro di una distribuzione

dato un insieme di n elementi $\{x_1, x_2, \ldots x_N\}$

- Si dice **media aritmetica semplice** di N numeri il numero che si ottiene dividendo la loro somma per N.

$$\bar{x} = \frac{x_1 + x_2 + \mathrm{L} \; + x_N}{N}$$

dato un insieme di m elementi $\{x_1, x_2, \ldots x_m\}$, e

dato un insieme di m di numeri reali $\{p_1, p_2, \ldots p_m\}$

- Si dice **media aritmetica pesata**

$$\bar{x} = \frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_m \cdot p_m}{p_1 + p_2 + \ldots + p_m}$$

che utilizza un peso p_j o la frequenza di ogni dato x_j per j=1,…,m

Misure di posizione : La Media Aritmetica (2)

Se è nota solo la distribuzione di frequenza per sottointervalli (non le singole osservazioni)

→ si calcola una **media approssimata**

Sia f_1 il numero di osservazioni che cadono nel sottointervallo 1;

tali osservazioni vengono approssimate dal valore centrale del sottointervallo (x_1)

analogamente per tutti gli altri sottointervalli

$$m \cong \frac{1}{n} \cdot \sum_{i=1}^{k} f_i x_i = \sum_{i=1}^{k} x_i \left(\frac{f_i}{n} \right)$$

f_i / n: frequenza relativa del sottointervallo *i*-esimo

k : numero dei sottointervalli

Esempio di media pesata

La media della lunghezza di un gruppo di $\mathbf{f_1}$= 7 neonati ⇨ $\mathbf{m_1}$=48.0 cm
e di altri $\mathbf{f_2}$= 3 neonati ⇨ $\mathbf{m_2}$=49.5 cm.

Per calcolare la media delle lunghezze dell'insieme totale di **10 neonati** pur senza avere la conoscenza dei valori delle lunghezze individuali, si utilizzano le proprietà della media aritmetica :

la somma delle lunghezze dei primi 7 è	48.0 × 7 = 336.0
la somma delle lunghezze dei secondi 3 è	49.5 × 3 = 148.5
la somma delle lunghezze di tutti i 10 è	= 484 .5

La media della lunghezza di tutti i 10 neonati è = 484.5/10 = 48.45

Ovvero **Media = $(f_1 \times m_1 + f_2 \times m_2)/(f_1 + f_2)$**
Media = $(7 \times 48.0 + 3 \times 49.5)/(7+3)$

esempio di media aritmetica

Lunghezza(cm) in un campione di 60 neonati.

51.0	49.4	49.0	52.5	51.5	51.8
46.5	47.8	49.7	44.5	49.8	53.0
48.7	50.0	52.9	50.8	46.2	48.9
54.5	48.2	48.9	51.2	49.5	56.3
46.0	52.2	47.0	50.8	50.0	52.5
51.2	51.1	54.7	52.3	48.2	50.8
55.0	50.2	50.3	47.7	48.5	53.8
50.2	53.4	47.4	50.5	51.7	49.5
44.4	49.2	50.5	49.5	52.9	50.5
54.0	46.5	51.5	50.9	51.6	52.7

la media aritmetica dei primi 6 valori di lunghezza di 6 neonati è:
$\bar{x}$ = (51.0+49.4+49.0+52.5+51.5+51.8)/6 = 305.2/6 = 50.87

la media aritmetica di tutti i 60 valori di lunghezza è:
= (55.9+51.3+53.0+50.5+54.9+53.4+…+53.8)/60 = 3021.8 /60
$\bar{x}$ = 50.363

La media aritmetica di N dati distinti è …

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

MEDIA per dati raggruppati in classi

ALTEZZA(cm) di un campione di 60 neonati.

limiti di classe	x_i	$f(x_j)$	$x_i f(x_j)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
Σ		60	3022.5

Nell'*esempio* del campione di 60 misure di lunghezza dei neonati:

$$\bar{x} = \frac{45.0\times 2 + 46.5\times 5 + \ \ldots\ 57.0\times 1}{2+3+\ \ldots\ 1} = \frac{3022.5}{60} = 50.375$$

La media per dati raggruppati in m classi è ...

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j \cdot f(x_j)}{\sum_{j=1}^{m} f(x_j)}$$

dove **m** è il numero di classi e ,

$\sum_{j=1}^{m} f(x_j) = N$ se $f(x_i)$ indica le frequenze assolute,

oppure $\sum_{j=1}^{m} f(x_j) = 1$ se $f(x_i)$ indica le frequenze relative.

media aritmetica e mediana

Si consideri un campione di valori di VES (*velocità di eritrosedimen-tazione*, mm/ora) misurati in 7 pazienti

{8, 5, 7, 6, 35, 5, 4}

In questo caso, **la media** ($\bar{x}$ = 10 mm/ora) **non è** un valore **tipico** della distribuzione: soltanto un valore su 7 è superiore alla media!

Conviene usare come indice del centro **la mediana**, definita come quel valore che divide a metà la distribuzione, sicché **l'insieme dei valori è per metà minore e per metà maggiore della mediana**.

Per **calcolare la mediana** si dispongono i dati in ordine crescente:
ordine originale: {8, 5, 7, 6, 35, 5, 4}
ordine crescente: {4, 5, 5, 6, 7, 8, 35}

mediana

Se n è **dispari**, la mediana è il valore che occupa la posizione (n+1)/2 nell'insieme ordinato.

Nell'*esempio*, poiché (n+1)/2=4, la mediana è 6 mm/ora, ed è tipica nel senso che si avvicina a buona parte dei valori del campione.

Se n è **pari**, la mediana è la media dei valori che occupano le posizioni (n/2) ed [(n/2)+1] nell'insieme ordinato dei numeri.

Se, nell'***esempio***, si esclude il valore più alto, si ottiene l'insieme ordinato {4, 5, 5, 6, 7, 8},

(n/2)=3 e [(n/2)+1]=4,

e la mediana vale (5+6)/2=5.5.