



Aspetti regolatori, brevettuali ed economici dello sviluppo dei farmaci

Analisi Statistica



DONATA RODI

JOSEF ESCHGFAELLER

16-17 LUGLIO 2014

Programma del corso



Statistica descrittiva

Variabili (quantitative, qualitative)
Rappresentazioni dei dati
Rilevazione dei dati: misure di frequenza
Indici di posizione e di dispersione

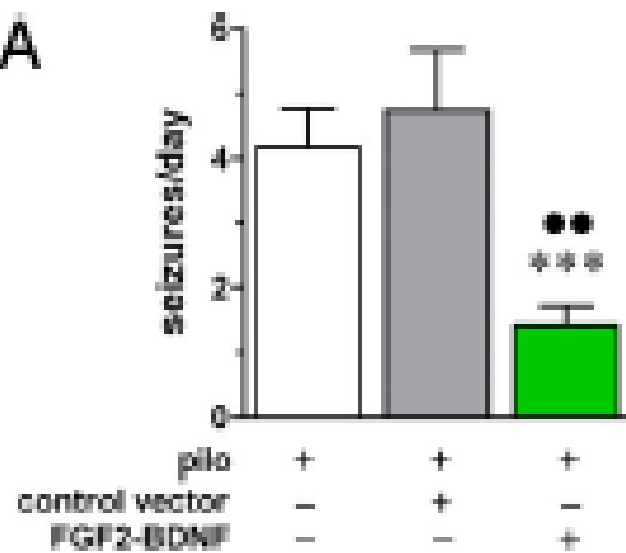
Statistica inferenziale

Campioni e popolazioni
Concetto di inferenza
Stima dei parametri
Distribuzioni campionarie di una stima e misure di incertezza
Verifica delle ipotesi
P value e significatività statistica
Errori di I e II tipo
Distribuzione binomiale
Test del χ^2
Distribuzione normale
Test t su campioni indipendenti e dipendenti
Cenni di ANOVA
Cenni di test non parametrici
La regressione

Esercitazioni

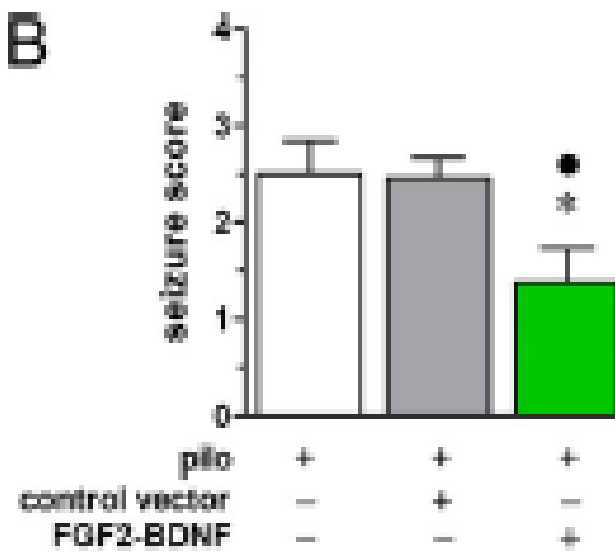
Linguaggio R
Excel
Eventuale software statistico

A



seizure frequency

B



seizure severity

STUDIO DI UN FENOMENO



1. Definizione

- Scopi, risorse, popolazione di riferimento e variabili di interesse

2. Raccolta delle informazioni

- Rilevazione delle osservazioni
- Indagine (Campionamento) / Esperimento

3. Presentazione dati: Statistica Descrittiva

- Tabelle (freq. assolute, relative, cumulate)
- Grafici (a torta, a barre, istogrammi)
- Misure di sintesi (posizione, dispersione)

4. Conclusioni generali: Statistica Inferenziale

- Stima di grandezze incognite relative alla popolazione
- Costruzione di modelli interpretativi e verifica dell'adeguatezza



Popolazione

Insieme oggetto di studio, su cui occorre produrre nuove conoscenze

Campione

Unità statistica

Elemento della popolazione

Sottoinsieme della popolazione su cui vengono rilevate le caratteristiche di interesse

Statistica descrittiva



Studio dei fenomeni aventi attitudine a variare.

- Fenomeni diversi a seconda delle circostanze in cui vengono osservati (tempo, luogo, popolazione)
- Fenomeni osservati sull'intera popolazione di interesse e, pertanto, completamente noti
- Statistica univariata
- Statistica bivariata o multivariata

Esempio: valutazione del livello di colesterolo in un determinato numero di pazienti in un determinato reparto ospedaliero

Chi, Dove, Come



Chi

Fenomeni con attitudine a variare. X , Y , A .

Dove

Unità statistiche supporto delle manifestazioni dei fenomeni (persone, enti, oggetti, etc.). L'insieme di tutte le unità statistiche $u_1, \dots, u_N$, oggetto di rilevazione, costituisce la popolazione statistica P , di numerosità N .

Come

Interpretazione delle manifestazioni del fenomeno tramite le scale di modalità

La variabile



Caratteristica propria di un fenomeno, con differenti modalità di manifestazione.

Qualitativa



Nominale

Ordinale

Dicotoma e non

Es: Lingua parlata,
colore occhi, gruppo
sanguigno, sesso,
status

Es. Grado istruzione,
stadio del tumore

Quantitativa



Discreta

Continua

Es. N medio sigarette/die,
N di aa in una proteina,
frequenza di attacchi d'asma

Es. Altezza,
T, P sanguigna,
pH.

Operatori

=, ≠

>, <

+, -, x, /

Esempio di Database: Blood



BloodType	AgeGroup	WBC	RBC	Chol
AB	Young	7710	7.4	258
A	Adult	5690	7.53	184
A	Young	6330	7.72	187
A	Old	6140	3.69	142
A	Young	6550	4.78	290
O	Adult	5200	4.96	151
O	Adult	5800	5.66	311
B	Young	6430	5.62	152
O	Adult	7410	5.85	241
O	Old	5590	6.94	152
A	Old	6520	6.03	217
O	Young	7210	5.17	193
O	Adult	6410	6.02	224
A	Old	6360	3.74	211
A	Adult	7580	5.13	179
A	Old	7150	6.35	200
A	Adult	8710	5.12	211
AB	Adult	7480	4.7	183
O	Young	8320	4.74	186
A	Adult	8020	5.03	182

La rilevazione: distribuzione di frequenza

Frequenze assolute

$$\sum_{i=1}^h f_i = N.$$

m_i	f_i
m_1	f_1
m_2	f_2
$\dots$	$\dots$
m_h	f_h
	N

m_i	f_i
A	10
B	1
O	7
AB	2
N	20

Consentono di avere informazioni sulla dimensione del fenomeno

Frequenze relative

$$p_i = \frac{f_i}{N},$$

$$0 \leq p_i \leq 1, \quad \sum_{i=1}^h p_i = 1.$$

m_i	f_i	p_i
A	10	0.5
B	1	0.05
O	7	0.35
AB	2	0.1
N	20	1

Consentono di fare confronti indipendentemente dalla numerosità

La rilevazione: distribuzione di frequenza



Frequenze percentuali

$$(100 p_i)\% = 100 p_i = 100 \frac{f_i}{N},$$

Frequenze cumulate

$$F_i = \sum_{j=1}^i f_j.$$

$$P_i = \sum_{j=1}^i p_j = \frac{F_i}{N}$$

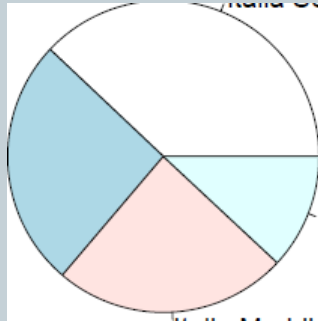
m_i	f_i	p_i	$p_i\%$	F_i	P_i
A	10	0.5	50	10	0.5
B	1	0.05	5	11	0.55
O	7	0.35	35	18	0.9
AB	2	0.1	10	20	1
N	20	1	100		

Rappresentazioni grafiche

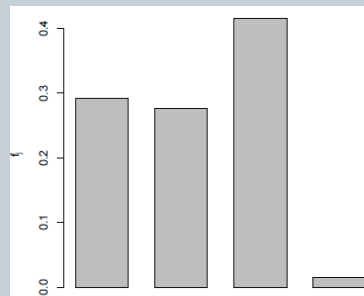
Variabili qualitative

- Grafici a torta

$$\alpha \div 360^\circ = x \div 100$$

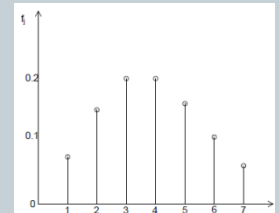


- Diagrammi a barre

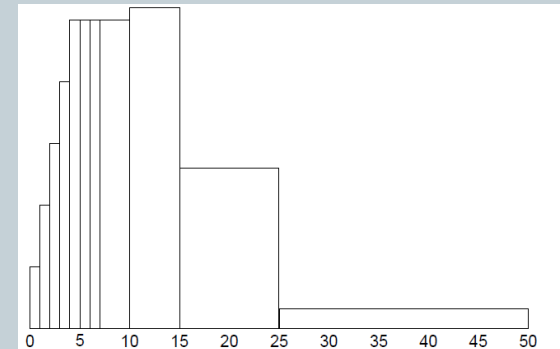


Variabili quantitative

- Discrete: diagramma a bastoncini



- Continue: istogrammi



Indici di tendenza centrale

Moda

- Valore o modalità alla quale è associata la massima frequenza
- E' calcolabile per qualunque tipo di fenomeno

m_i	f_i
A	10
B	1
O	7
AB	2
N	20

Mediana

- Valore o modalità che, ordinati in senso crescente i dati, occupa la posizione centrale separando il data-set in due parti uguali: metà delle osservazioni è $<$ della mediana, metà è $>$
- Calcolabile per fenomeni misurabili almeno con scala ordinale.

Mediana, Quartili



Calcolo della mediana

n è dispari $\frac{n+1}{2}$

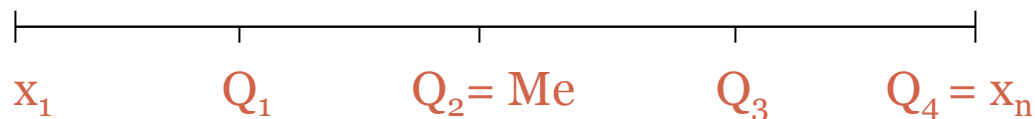
n è pari $\frac{n}{2}, \frac{n}{2} + 1$

x_i	f_i	F_i	posizioni	P_i
0	2	2	1 ^A ... 2 ^A	0.03
4	5	7	3 ^A ... 7 ^A	0.09
15	30	37	8 ^A ... 37 ^A	0.5
18	27	64	38 ^A ... 64 ^A	0.86
20	10	74	65 ^A ... 74 ^A	1
	74			

$$\frac{74}{2} = 37, \frac{n}{2} + 1 = 38$$

$$\frac{15+18}{2} = 16.5$$

Quartili: indici di posizione fondati sul concetto di divisione della distribuzione. Dividono la serie ordinata in quattro parti contenendo ciascuna lo stesso numero di dati.

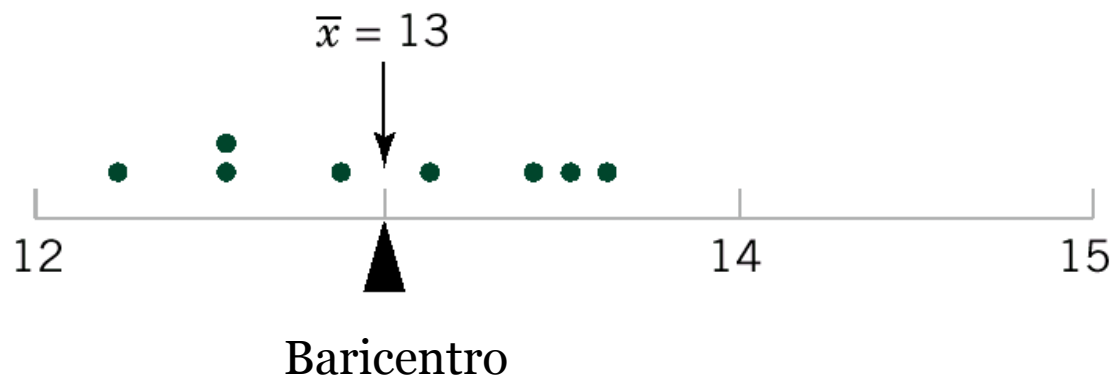


Media aritmetica

Rapporto tra
l'intensità totale
e la numerosità
n del campione

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

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$



Esempio



id	Peso (X _i)
1	70
2	65
3	54
4	71
5	70
6	67
7	65
8	45
9	55
10	70

$$\textbf{Media} = \frac{\sum_{i=1..n} x_i}{n} = (70+65+.....+55+70)/10=63,2$$

Mediana

1) Ordino i dati: 45 54 55 65 65 67 70 70 70 71

Se **n dispari**

$$X_{\text{Med}} = X_{(n+1)/2}$$

Se **n pari**

$$X_{\text{Med}} = (X_{n/2} + X_{(n)/2+1})/2$$

2) n=10

$$X_{\text{Med}} = (65+67)/2=66$$

Moda = osservazione che si verifica con maggiore frequenza

Moda=70 Kg

Indici di dispersione



Varianza

Esprime i valori di dispersione intorno alla media, elevati al quadrato

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

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

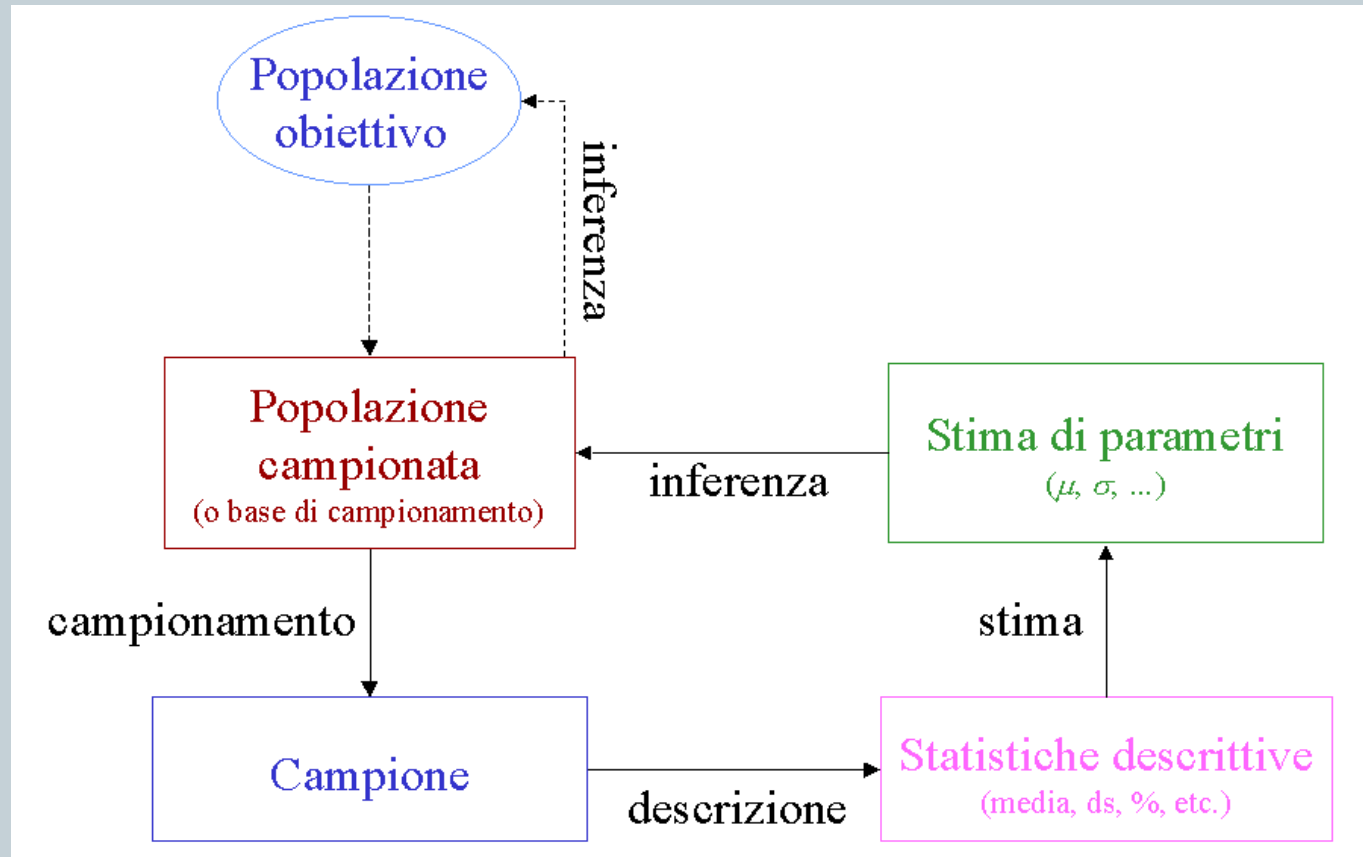
Deviazione Standard

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

$$\sum (x_i - \bar{x})^2 = \text{Devianza} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

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

Statistica inferenziale



Statistica inferenziale



- Fenomeni osservati su una parte della popolazione di interesse (rilevazione parziale): risultati non certi, ma “probabili”.
- Nelle modalità di scelta della parte di popolazione da osservare (campionamento) e dell’assegnazione di un grado di fiducia/probabilità ai risultati ottenuti si avvale dell’ausilio di
 - teoria dei campioni
 - calcolo delle probabilità

Statistica Inferenziale



Stima di un parametro

- L'evidenza campionaria, eventualmente integrata da conoscenze a priori, viene utilizzata per stimare un'entità incognita relativa ad una certa popolazione

Test delle ipotesi

- L'evidenza campionaria, eventualmente integrata da conoscenze a priori, viene utilizzata per verificare statisticamente la validità di una certa assunzione (ipotesi) concernente una specifica entità incognita



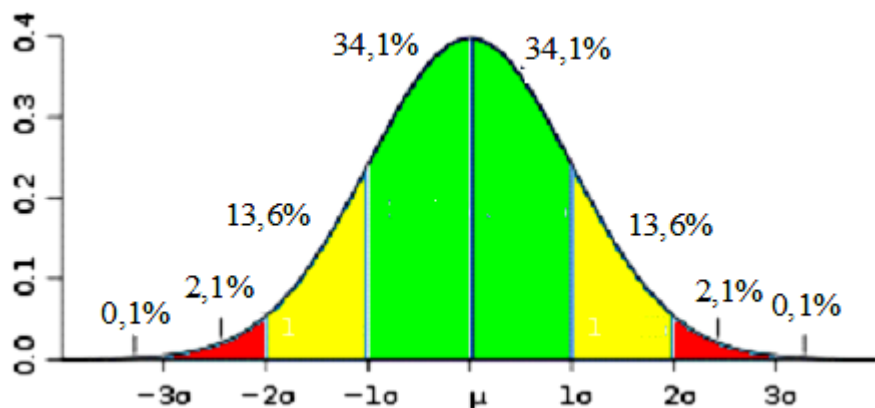
Distribuzioni campionarie

$$\bar{X}_n \sim N\left(\mu; \frac{\sigma^2}{n}\right)$$

$\bar{X}$ è stimatore
corretto per la μ

- **P**= popolazione sulle unità della quale è presente il carattere x
- **X**= variabile casuale (v.c.), cioè il modello probabilistico idoneo per interpretare il carattere x .
- $(x_1, \dots, x_n)$ n-upla campionaria
- **Distribuzione campionaria:** si ottiene se si ipotizza di estrarre tutti i possibili campioni di n elementi

La distribuzione Normale



$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right) \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- L'area tra $X=\mu$ ed $X=1\sigma$ è il 34.1% dell'area totale (68.2% tra -1σ e 1σ)
- L'area tra $X=\mu$ ed $X=2\sigma$ è il 47.7% dell'area totale (95.4% tra -2σ e 2σ)
- L'area tra $X=\mu$ ed $X=3\sigma$ è il 49.9% dell'area totale (99.8% tra -3σ e 3σ)

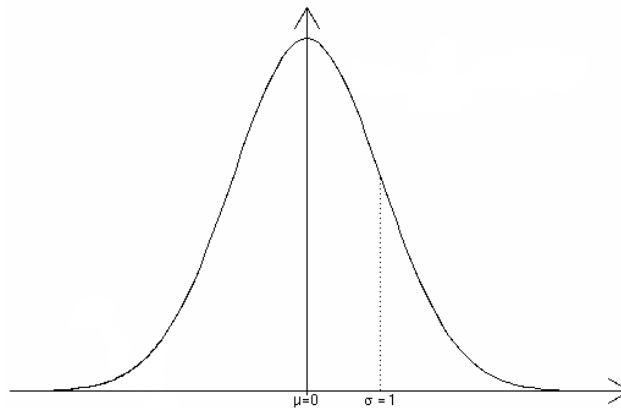
Proprietà della Normale

1. Media, moda e mediana coincidenti
2. Curva simmetrica rispetto alla media μ
3. Curva asintotica all'asse delle X
4. Area sottesa alla curva uguale a 1.

La distribuzione Normale Standardizzata



$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \longrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$



Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	<u>0,08</u>	0,09
0,0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0,1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0,2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0,3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0,4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0,5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0,6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0,7	0.242	0.239	0.236	0.233	0.230	0.227	0.224'	0.221	0.218	0.215
0,8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0,9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1,0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1,1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1,2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1,3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1,4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
<u>1,5</u>	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	<u>0.057</u>	0.056
1,6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.048	0.046	0.046
1,7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.037	0.037
1,8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.030	0.029	0.029
1,9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2,0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2,1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2,2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2,3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2,4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2,5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2,6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2,7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2,8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2,9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3,0	0.001									

Tavola z

Stima di un parametro



Stima puntuale

- La stima è data da un singolo valore

Stima per intervallo

- La stima fornisce gli estremi di un intervallo nel quale, con un certo grado di fiducia, si può supporre che il parametro sia compreso

Parametri più frequentemente stimati:

Media μ , varianza σ^2 , proporzione p , differenza $\mu_1 - \mu_2$

ES e Intervallo di confidenza



- **Errore standard:** descrive l'incertezza nella stima di un *valore statistico* (es. media) e può essere utilizzato per calcolare l'intervallo di confidenza

$$ES = \sigma / \sqrt{n}$$

- **Intervallo di confidenza** è l'intervallo di valori entro i quali si stima che cada, con un livello di probabilità scelto a piacere, il valore vero della popolazione (0.95 o 0.99).

Stima della media, varianza nota



$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \longrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$P(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq +z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq +z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

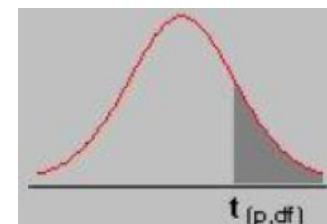
Intervallo di confidenza
per μ con grado di fiducia
 $1 - \alpha$

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Stima della media, varianza ignota



$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \left[\bar{x} - \frac{s}{\sqrt{n}} t_{n-1, \alpha/2}, \bar{x} + \frac{s}{\sqrt{n}} t_{n-1, \alpha/2} \right]$$



Distribuzione dei valori critici per test unilaterali

df	Probabilità p nella coda superiore											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

Test delle Ipotesi Statistiche



Ipotesi statistica

- Ipotesi formulata in termini di parametri di una popolazione
- L'affermazione può riferirsi sia alla forma funzionale della legge di distribuzione che ai parametri caratteristici o ai solli parametri caratteristici quando si assuma nota la forma della distribuzione stessa.

Test d'ipotesi

- Regola con la quale si decide se accettare o meno l'ipotesi formulata sulla base delle risultanze campionarie.
- Tali dati si riferiscono alla variabile casuale sulla cui legge di distribuzione è stata formulata l'ipotesi.



Test di ipotesi

1. Ipotesi: definizione di H_0 (ipotesi nulla) contro H_1 (ipotesi alternativa)
2. Scelta di α : si decide con quale livello di significatività si vuole eseguire il test
3. Scelta del test: in funzione di h_p , scala, dati
4. Determinazione dei valori critici (dipende se è a una coda o a due code)
5. Calcolo della statistica del test (p value)
6. Decisione: accetto o rifiuto H_0



Ipotesi nulla H_0

- Ipotesi di non differenza

$$H_0 : \mu = \mu_0$$

- E' posta con lo scopo di essere screditata
- Se viene rifiutata, i dati non forniscono sufficiente evidenza per sostenere H_1

Ipotesi alternativa H_1

- Tutte le altre ipotesi (generica)
- Unilaterali, bilaterali

$$H_1 : \mu \neq \mu_0$$

$$H_1 : \mu < \mu_0$$

- E' ciò che si spera di poter concludere come risultato del test

Esempio: un campione



$$H_0 : \mu = 30$$

$$\alpha = 0,01$$

$$H_1 : \mu < 30$$

Dati
campionari

$$\bar{x} = 26$$

$$s^2 = 36$$

$$n=25$$

$$t = \frac{\bar{X} - \mu}{S_x}$$

$$t = \frac{26 - 30}{\sigma / \sqrt{25}} = -3,33 < -2,492$$



Rifiuto H_0

Tavola t

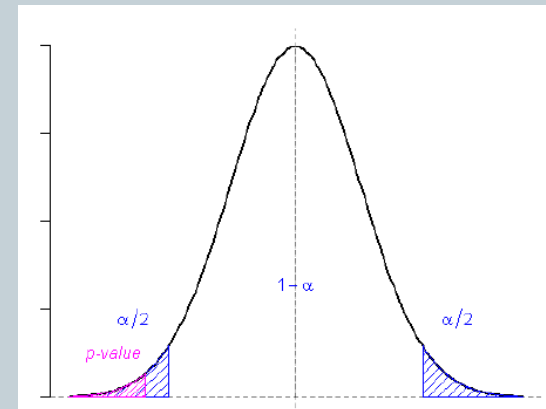
GL	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	1.292	1.664	1.990	2.374	2.639	3.195	3.416
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
infinity	1.282	1.645	1.960	2.326	2.576	3.091	3.291

P value e Significatività statistica

- P è la probabilità che, data H_0 (cioè assumendo H_0 vera), si abbia un campione che si allontani dai valori ideali, descritti da H_0 , almeno tanto quanto il campione esaminato
- Esempio: ho meno del 5% di P di sbagliare affermando che c'è differenza tra i due gruppi sperimentali e che tale differenza è dovuta non al caso ma al trattamento.
- Significatività statistica: Soggettiva!

$$P < 0.05$$

$$P [(X \geq x) | H_0]$$



LIVELLO DI PROBABILITA'	SIGNIFICATIVITA' DEL TEST	SIMBOLO
$P < 0.05$ (5%)	significativo	*
$P < 0.01$ (1%)	molto significativo	**
$P < 0.001$ (0,1%)	altamente significativo	***

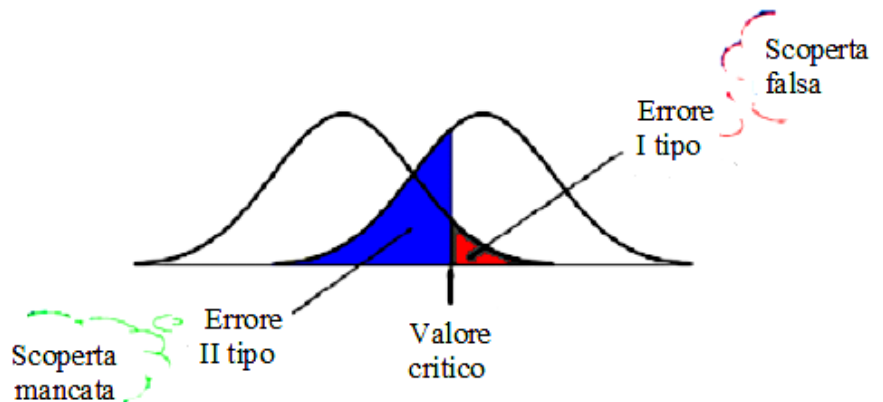
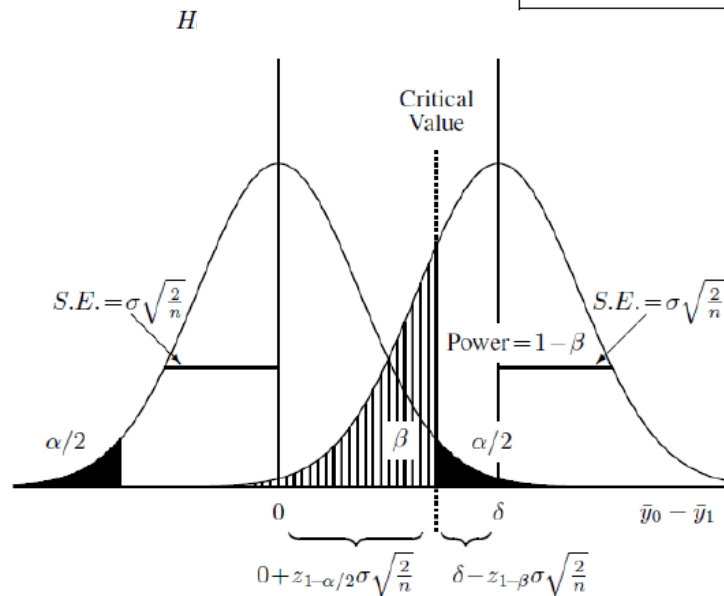
Errori nella verifica delle ipotesi



REALTÀ'

CONCLUSIONE
DEL
TEST

	H_0 vera	H_0 falsa
H_0 vera (non significativo)	Esatto $P = 1 - \alpha$ <i>Protezione</i>	Errore Tipo II $P = \beta$
H_0 falsa (significativo)	Errore Tipo I $P = \alpha$ <i>Significatività</i>	Esatto $P = 1 - \beta$ <i>Potenza</i>



Confronto tra due campioni: t di Student



	n	$\bar{x}$	s
Gruppo 1	12	4,5 mg/100 ml	1 mg/100 ml
Gruppo 2	15	3,4 mg/100 ml	1,22 mg/100 ml

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$t = \frac{(4,5 - 3,4)}{\sqrt{1,273504 \left(\frac{1}{12} + \frac{1}{15} \right)}} = 2,52$$

Due campioni dipendenti



Nuovo training autogeno su 10 soggetti depressi. La depressione è misurata con opportuna scala e $>$ è il punteggio tanto maggiore è lo stato di depressione. Il livello di depressione viene misurato sugli stessi 10 soggetti prima e dopo la somministrazione del training autogeno. Avendo fissato $\alpha = 0.10$ si può ritenere che il training non abbia avuto effetto sulla depressione?

<i>Prima</i>	13.48	15.30	13.56	12.01	10.11	17.20	13.11	14.12	15.00	16.00
<i>Dopo</i>	13.20	15.00	13.50	12.20	10.10	16.10	13.20	14.00	15.00	15.90
e_i	-0.28	-0.30	-0.06	0.19	-0.01	-1.1	0.09	-0.12	0	-0.1

Sotto H_0 , cioè se il trattamento non ha effetto, il valore atteso di E è 0
Divario $d = e - o = e$ dove e è la media delle n differenze. L'ipotesi sarà rifiutata se il divario risulta troppo "grande" dal momento che in tal caso la media e risulta troppo distante da quella ipotizzata che è pari a 0.

$$t = \frac{-0.17}{\sqrt{\frac{0.13}{10}}} = -1.49 \quad t_{\alpha/2} = 1.83,$$

H_0 non rifiutata

Test Chi quadro di indipendenza



	Nuovo farmaco	Placebo	
Vivo	11	6	17
Morto	11	16	27
	22	22	44

Ho: non esiste associazione tra trattamento ed esito

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$E_1 = \frac{22 \times 17}{44} = 8,5 \quad \text{Frequenze attese}$$

$$\begin{aligned} \chi^2 &= \frac{(11 - 8,5)^2}{8,5} + \frac{(6 - 8,5)^2}{8,5} + \frac{(11 - 13,5)^2}{13,5} + \frac{(16 - 13,5)^2}{13,5} = \\ &= 0,74 + 0,74 + 0,46 + 0,46 = 2,4 \end{aligned}$$

$$2,4 < 3,84$$



Non rifiuto Ho

ANOVA



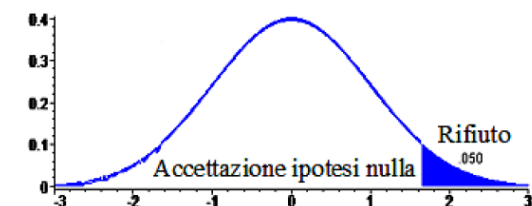
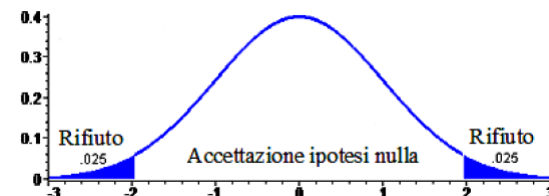
Fonte di variazione	Somma dei quadrati	Gradi di libertà	Media quadratica	Rapporto di varianza
Fra gruppi	$SSB = \sum_{j=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2$	$k - 1$	$MSB = \frac{SSB}{k - 1}$	$R.V. = F = \frac{MSB}{MSW}$
All'interno dei gruppi	$SSW = \sum_{i=1}^k \sum_{j=1}^{nk} (y_{ij} - \bar{y}_{.j})^2$	$N - k$	$MSW = \frac{SSW}{N - k}$	
Totale	$SST = \sum_{i=1}^k \sum_{j=1}^{nk} (y_{ij} - \bar{y}_{..})^2$	$N - 1$		

ANOVA



Fonte di variazione	df	Somma dei quadrati	Media dei quadrati	F	P-value
Gruppi	k-1	SS_{gruppi}	$MS_{\text{gruppi}} = SS_{\text{gruppi}} / (k-1)$	$F_{\text{calc}} = MS_{\text{gruppi}} / MS_{\text{errore}}$	$P(F > F_{\text{calc}})$
Errore	N-k	SS_{errore}	$MS_{\text{errore}} = SS_{\text{errore}} / (N-k)$		
Total	N-1	SS_{totale}			

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0,1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0,2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0,3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0,4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0,5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0,6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0,7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0,8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0,9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1,0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1,1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1,2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1,3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1,4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1,5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1,6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.048	0.046	0.046
1,7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.037	0.037
1,8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.030	0.029	0.029
1,9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023
2,0	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018
2,1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014
2,2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011
2,3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008
2,4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
2,5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
2,6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
2,7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2,8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2,9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3,0	0.001									



Test parametrici vs non parametrici



Parametric	Non-Parametric
1 Sample T-test	Sign Test/Wilcoxon Signed Rank test
Paired T-test	Sign Test/Wilcoxon Signed Rank test
2 Sample T-test	Mann Whitney U test/Wilcoxon Sum Rank test
ANOVA	Kruskal Wallis test

One sample t test



- Indica se la media di una variabile differisce o no da una costante specificata

Pz: dolore toracico acuto
N=500

Table III. Descriptive statistics for the systolic & diastolic BP.

	Mean	Std Deviation	Minimum	Maximum	Median
systolic blood pressure	140.51	24.08	79.00	200.00	139.00
diastolic blood pressure	78.65	12.91	48.00	110.00	79.00

One-Sample Test

Test Value = 120

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
systolic blood pressure	19.046	499	.000	20.5080	18.3925	22.6235

Assunzione:
distribuzione normale

Se la distribuzione non è
normale: test dei segni o
Wilcoxon

Paired t-test



- Dati appaiati: within group comparison

P sistolica prima e dopo
un intervento chirurgico

	Mean	Std Deviation	Minimum	Maximum	Median
systolic BP before	142.31	22.38	90.00	200.00	139.00
systolic BP after	137.14	24.87	90.00	199.00	137.00

Paired Differences							
			95% Confidence Interval of the Difference				
		Std. Error					Sig. (2-tailed)
Mean	Deviation	Mean	Lower	Upper	t	df	
5.17	34.11	2.64	-.04	10.38	1.958	166	.052

Assunzione:

La differenza tra prima e
dopo ha una distribuzione
normale

Se la distribuzione non è
normale: Wilcoxon

Two sample t-test



- Due campioni indipendenti: between group comparison

Assunzione:
obs distribuite
normalmente

Se la distribuzione
non è normale: Mann
Whitney o Wilcoxon

	Mean	Std Deviation	Minimum	Maximum	Median
over-weight	141.65	23.06	90.00	200.00	138.00
normal-weight	97.12	10.82	80.00	132.00	100.00

Independent Samples Test									
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	131.183	.000	27.638	498	.000	44.5280	1.61111	41.36258	47.69342

ANOVA



- Più di due gruppi indipendenti

	Mean	Std Deviation	Minimum	Maximum	Median
over-weight	140.89	24.65	90.00	195.00	137.50
under-weight	104.72	21.58	80.00	186.00	100.00
normal-weight	112.14	26.79	80.00	194.00	100.00

Assunzione:

Le obs hanno una
distribuzione normale

Se la distribuzione non è
normale: Kruskal-Wallis

ANOVA

systolic BP

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	72943.542	2	36471.771	61.126	.000
Within Groups	176613.970	296	596.669		
Total	249557.512	298			

ANOVA



Multiple comparisons

Dependent variable: systolic BP
Bonferroni

(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	95% Confidence Interval		
				Sig.	Lower Bound	Upper Bound
over-weight	over-weight					
	under-weight	36.1700*	3.45447	.000	27.8528	44.4872
	normal-weight	28.7486*	3.46318	.000	20.4104	37.0868
under-weight	over-weight	-36.1700*	3.45447	.000	-44.4872	-27.8528
	under-weight					
	normal-weight	-7.4214	3.46318	.099	-15.7596	.9168
normal-weight	over-weight	-28.7486*	3.46318	.000	-37.0868	-20.4104
	under-weight	7.4214	3.46318	.099	-.9168	15.7596
	normal-weight					

• Confronti ripetuti

Number of
comparisons

1 2 3 4 5 6 7 8 9 10

Probability of
false positive

5% 10% 14% 19% 23% 27% 30% 34% 37% 40%