Lezione VII:

test z

Cattedra di Biostatistica – Dipartimento di Scienze Biomediche, Università degli Studi "G. d'Annunzio" di Chieti – Pescara

Prof. Enzo Ballone

Statistica inferenziale per variabili quantitative

SIMBOLOGIA CORRENTE

	campione	popolazione
numerosità	n	N
media aritmetica	$\bar{x}$	μ
deviazione standard	s	σ

70.

Statistica inferenziale per variabili quantitative

	glicemia
x_1	103
x_2	97
x_3	90
x_4	119
x_5	107
x_6	71
x_7	94
x_8	81
x_9	92
x_{10}	96

Un problema pratico

- Quesito:
  - qual è la glicemia media della popolazione degli studenti iscritti al primo anno del corso di laurea in medicina?
- Procedura
  1. dall'elenco dei 200 studenti estraggo un campione casualmente costituito da 10 unità;
  2. calcolo la media aritmetica e la deviazione standard.
  3. determino i livelli di glicemia in ciascun soggetto che costituisce il campioni

Un problema pratico

	campione	popolazione
numerosità	$n = 10$	$N = 95$
media aritmetica	$\bar{x} = 95$	$\mu = ?$
deviazione standard	$s = 13.32$	$\sigma = ?$

Un problema pratico

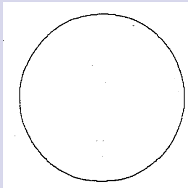
- Posso affermare che $= \bar{x} = \mu$?

NO!

- perché l'estrazione del campione comporta necessariamente una perdita di informazione che si traduce nell'introduzione di un **errore casuale**
- Se il problema si esaurisse con questa affermazione tutte le osservazioni effettuate su base campionaria non servirebbero a nulla in quanto l'interesse è sempre rivolto alla popolazione!

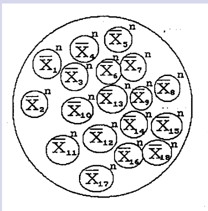
Un breve salto nella teoria

- Popolazione costituita da N elementi
- Il carattere in esame ha media aritmetica= μ deviazione standard= σ



Un breve salto nella teoria

1. si estraggano da tale popolazione infiniti campioni di numerosità n e per ognuno di essi si calcoli la media aritmetica = (= $\bar{x}_i$)



Un breve salto nella teoria

2. il risultato è un insieme infinito di medie aritmetiche = (= $\bar{X}_i$) riferite a campioni di numerosità n;

$$\bar{X}_1 , \bar{X}_2 , \bar{X}_3 , \bar{X}_4 , \dots\dots\dots, \bar{X}_\infty$$

$$n_1 = n_2 = n_3 = n_4 \dots\dots\dots = n_\infty$$

Un breve salto nella teoria

3. se ciascuna = $\bar{X}_i$ viene considerata come un'osservazione individuale è possibile costruire una distribuzione di frequenza definita:

DISTRIBUZIONE DI CAMPIONAMENTO DELLE MEDIE DEI CAMPIONI DI NUMEROSITA' n.

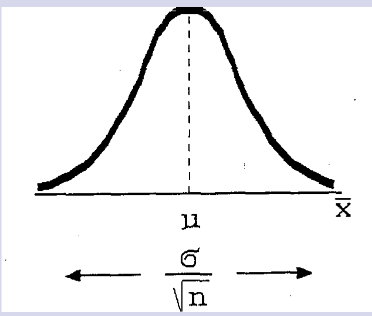
Un breve salto nella teoria

4. La distribuzione di campionamento ha le seguenti proprietà:

- A) la **media aritmetica** è uguale alla media aritmetica della popolazione $\bar{X}_{\bar{x}_i} = \mu$
- B) la deviazione standard è uguale a $\frac{\sigma}{\sqrt{n}}$

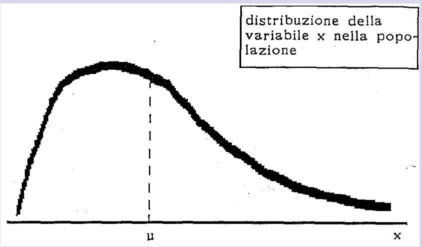
  questo parametro viene definito errore standard (E.S.) ed è una misura della precisione della stima campionaria della media aritmetica della popolazione (misura dell'errore campionario delle medie aritmetiche).
- C) la forma è approssimativamente normale, a condizione che:
  - il carattere sia distribuito normalmente nella po-polazione o;
  - n sia sufficientemente grande (>30);

Un breve salto nella teoria

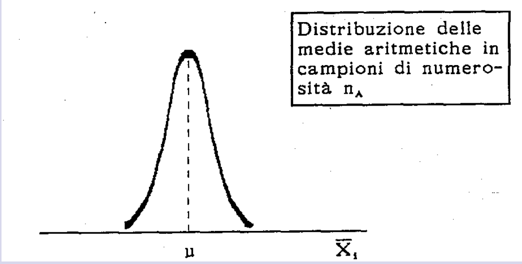


Un breve salto nella teoria

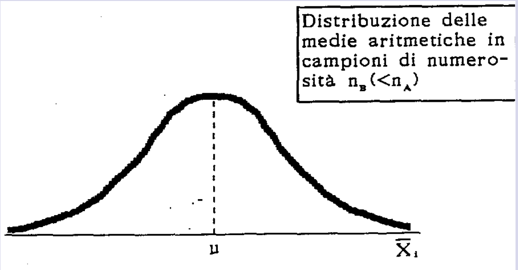
- L'errore di campionamento è tanto più evidente quanto più piccolo è il campione:



Un breve salto nella teoria

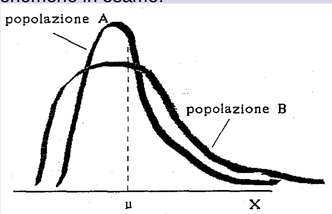


Un breve salto nella teoria



Un breve salto nella teoria

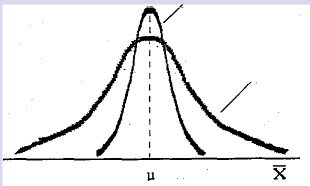
- L'errore di campionamento è tanto più evidente quanto più è disperso è il fenomeno in esame:



- distribuzione della variabile x in due popolazioni con μ uguali e σ diverse

Un breve salto nella teoria

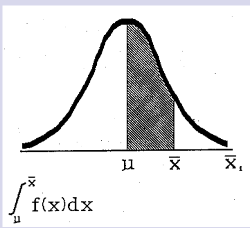
- distribuzione di campionamento delle $\bar{x}_{iA}$ e $\bar{x}_{iB}$



- Distribuzione di campionamento delle $\bar{x}_{iB}$

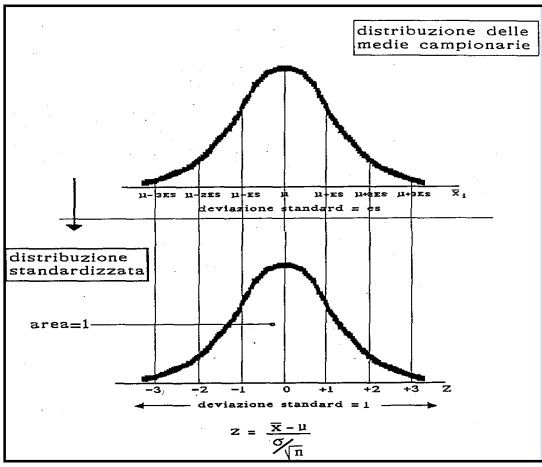
Un breve salto nella teoria

5. il fatto che le $=\bar{x}_i$ si distribuiscano secondo la legge normale consente di calcolare la proba-bilità che $=\bar{x}$ sia più o meno diversa da μ.



Un breve salto nella teoria

6. per calcolare questa probabilità si trasforma la distribuzione di campionamento nella distribuzione normale standardizzata:



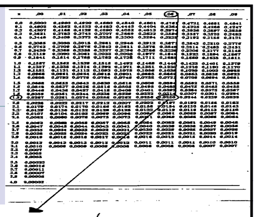
Un breve salto nella teoria

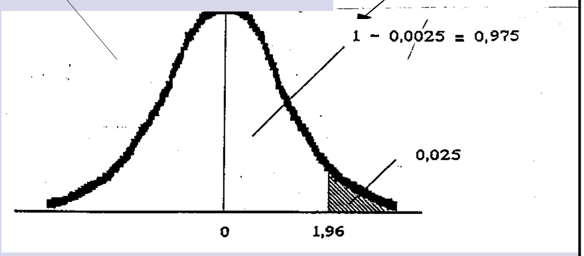
- ad ogni valore di z corrisponde un preciso valore dell'area a destra del valore stesso
  - ad **esempio** per Z= 1,96

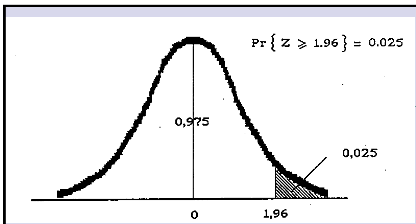
z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007									
3,3	0,0005									
3,4	0,0003									
3,5	0,00023									
3,6	0,00016									
3,7	0,00011									
3,8	0,00007									
3,9	0,00005									
4,0	0,00003									

$$\Pr\{-\infty < Z < +1{,}96\} = 0.975$$

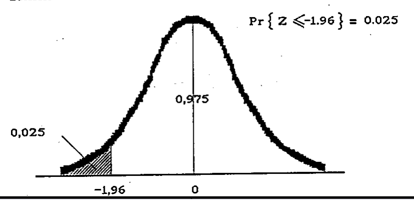
$$\mathrm{pr}\{Z \geqslant 1.96\} = 0.025$$



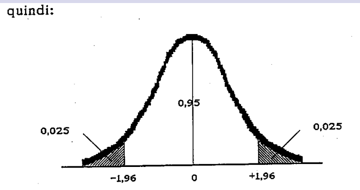




per simmetria:



Segue esempio:



$$Pr\left\{-1.96 \leqslant Z \leqslant +1.96\right\} = 0.95$$

Segue esempio:

$$Pr\left\{-1{,}96 \leqslant Z \leqslant +1{,}96\right\} = 0.95$$

$$Pr\left\{-1{,}96 \leqslant \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leqslant +1{,}96\right\} = 0.95$$

$$Pr\left\{\bar{x}-1{,}96\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+1{,}96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

intervallo di confidenza al 95%

$$\bar{x} \pm 1{,}96\frac{\sigma}{\sqrt{n}}$$

Segue esempio:

- la media aritmetica della popolazione è compresa in questo intervallo con una probabilità del 95%.
- In generale:
  - intervallo di confidenza al livello di probabilità (1-a) :

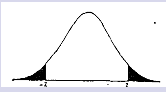
$$\boxed{\bar{x} \pm z\frac{\sigma}{\sqrt{n}}}$$

Segue esempio:

- ovvero μ è compreso nell'intervallo indicato con una probabilità (1-α); α è l'area delle code della gaussiana ed è definita dal valore di z.

Questa tabella riporta le aree tratteggiate

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	1.000	0.992	0.984	0.976	0.968	0.960	0.952	0.944	0.936	0.928
0.1	0.920	0.912	0.904	0.897	0.889	0.881	0.873	0.865	0.857	0.849
0.2	0.841	0.834	0.826	0.818	0.810	0.803	0.795	0.787	0.779	0.772
0.3	0.764	0.757	0.749	0.741	0.734	0.726	0.719	0.711	0.704	0.697
0.4	0.689	0.682	0.674	0.667	0.660	0.653	0.646	0.638	0.631	0.624
0.5	0.617	0.610	0.603	0.596	0.589	0.582	0.575	0.569	0.562	0.555
0.6	0.549	0.542	0.535	0.529	0.522	0.516	0.509	0.503	0.497	0.490
0.7	0.484	0.478	0.472	0.465	0.459	0.453	0.447	0.441	0.435	0.430
0.8	0.424	0.418	0.412	0.407	0.401	0.395	0.390	0.384	0.379	0.373
0.9	0.368	0.363	0.358	0.352	0.347	0.342	0.337	0.332	0.327	0.322
1.0	0.317	0.312	0.308	0.303	0.298	0.294	0.289	0.285	0.280	0.276
1.1	0.271	0.267	0.263	0.258	0.254	0.250	0.246	0.242	0.238	0.234
1.2	0.230	0.226	0.222	0.219	0.215	0.211	0.208	0.204	0.201	0.197
1.3	0.194	0.190	0.187	0.184	0.180	0.177	0.174	0.171	0.168	0.165
1.4	0.162	0.159	0.156	0.153	0.150	0.147	0.144	0.142	0.139	0.136
1.5	0.134	0.131	0.129	0.126	0.124	0.121	0.119	0.116	0.114	0.112
1.6	0.110	0.107	0.105	0.103	0.101	0.099	0.097	0.095	0.093	0.091
1.7	0.089	0.087	0.085	0.084	0.082	0.080	0.078	0.077	0.075	0.073
1.8	0.072	0.070	0.069	0.067	0.066	0.064	0.063	0.061	0.060	0.059
1.9	0.057	0.056	0.055	0.054	0.052	0.051	0.050	0.049	0.048	0.047
2.0	0.046	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
2.1	0.036	0.035	0.034	0.033	0.032	0.032	0.031	0.030	0.029	0.029
2.2	0.028	0.027	0.026	0.026	0.025	0.024	0.024	0.023	0.023	0.022
2.3	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018	0.17	0.17
2.4	0.016	0.016	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013
2.5	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.010
2.6	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007
2.7	0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.006	0.005	0.005
2.8	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004
2.9	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003
3.0	0.003									



Problema:

- se è sconosciuta (come spesso accade) il discorso procede come prima con due differenze:
  - si utilizza S al posto di σ (σ --▸ S) :
  - si utilizza la "distribuzione t di student", anzichè quella normale (Z --▸ t);
  - l'intervallo di confidenza al livello di probabilità (1- α)`diventa:

$$\boxed{\bar{x} \pm t_{g.l.} \frac{s}{\sqrt{n}}}$$

Tabella

- Percentili della distribuzione t (questa tabella dà i valori di t che,, per diversi g.l., forniscono una specificata proporzione in una o nelle due code della distribuzione t)

	Area nelle due code				
	0.10	0.05	0.02	0.01	0.001
	Area in una coda				
g.l.	0.05	0.025	0.01	0.005	0.0005
1	6.314	12.706	31.821	63.657	636.619
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.941
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.859
6	1.943	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.405
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.093	2.539	2.861	3.883
20	1.725	2.086	2.528	2.845	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
∞	1.645	1.960	2.326	2.576	3.291

Esempio:

- Nell'esempio delle glicemie
  - per una probabilità del 95%, con (10-1)= 9 gradi di libertà;

    t = 2,262
  - la glicemia media (= μ) della popolazione degli studenti iscritti al primo anno di medicina) è compresa tra:

    $$95 - 2{,}262\,\frac{13.32}{\sqrt{10}} \div 95 + 2{,}262\,\frac{13.32}{\sqrt{10}}$$

    85,47 ± 104.53
  - con una probabilità del 95%

Riepilogo

- RIEPILOGO SUGLI INTERVALLI DI CONFIDENZA DELLA MEDIA ARITMETICA CAMPIONARIA
  - se σ è conosciuta: $\boxed{\bar{x} \pm z\,\frac{\sigma}{\sqrt{n}}}$
  - se σ è sconosciuta: $\boxed{\bar{x} \pm t_{g.l.}\,\frac{s}{\sqrt{n}}}$

Un secondo problema

- La letteratura riporta che i pazienti affetti da cancro hanno una sopravvivenza media di 38,3 mesi e deviazione standard di 43,3 mesi:

  P: μ = 38,3 mesi σ = 43,3 mesi

- **(la distribuzione dei tempi di sopravvivenza non è sicuramente gaussiana)**
- Si saggia la capacità di un nuovo farmaco di allungare la sopravvivenza somministrandolo a 100 pazienti, ottenendo una sopravvivenza media di 46,9 mesi:
- **(La dimensione campionaria è sufficiente a garantire la normalità della distribuzione di campionamento)**

Un secondo problema

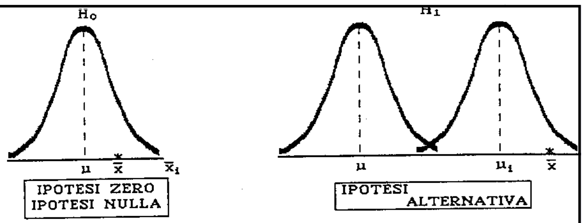
Si ipotizzi che tre possibili campioni diano i seguenti risultati:

$$H_1 = C_1 : n_1 = 100; \bar{x} = 46{,}9$$

$$H_2 = C_2 : n_2 = 100; \bar{x} = 53{,}3$$

$$H_3 = C_3 : n_3 = 25; \bar{x} = 55{,}0$$

- Formalizzo il problema avanzando due ipotesi:



- I 100 pazienti rappresentano un campione casuale estratto dalla popolazione dei pazienti affetti da cancro.
- $\bar{x}$ è diversa da μ per effetto dell'errore di campionamento
- Il farmaco non può essere considerato efficace.

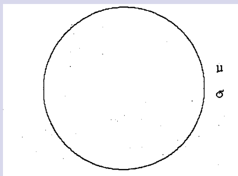
- I 100 pazienti non rappresentano un campione casuale estratto dalla popolazione dei pazienti affetti da cancro, ma di una altra popolazione di pazienti che ha tratto beneficio dal farmaco.
- La differenza tra $\bar{x}$ e μ non può essere spiegata dall'errore di campionamento.
- Il farmaco può essere considerato efficace.

Conclusione

- I test statistici saggiano la veridicità dell'ipotesi nulla.

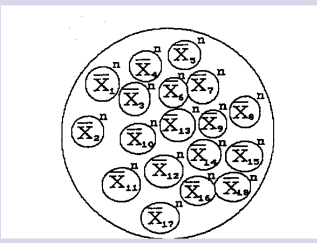
Teoria dei campioni

- Si torni alla teoria dei campioni:
  1. dalla popolazione con media μ e deviazione standard σ:



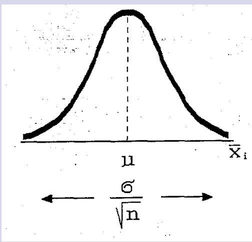
Teoria dei campioni

- Si estraggono infiniti campioni di numerosità n



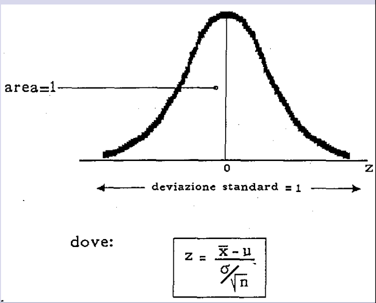
Teoria dei campioni

2. se ciascuna $\bar{x}_i$ viene considerata come un'osservazione individuale, è possibile costruire una distribuzione di campionamento delle medie dei campioni di numerosità n



Teoria dei campioni

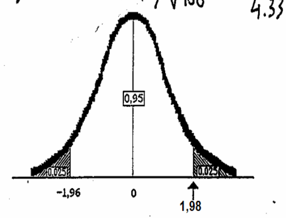
3. standardizzando:



Teoria dei campioni

- Si torni al problema:

$$1)\ Z = \frac{46.9 - 38.3}{43.3/\sqrt{100}} = \frac{8.6}{4.33} = 1.98 \text{ (n.s.)}$$



Teoria dei campioni

- si accetta l'ipotesi nulla
- ($p>0.05$) è la probabilità associata alla veridicità dell'ipotesi nulla
- il farmaco non può essere considerato efficace

$$H_2 \rightarrow Z = \frac{53.3-38.3}{43.3/\sqrt{100}} = \frac{15}{4.33} = 3.46 \; (p<0.001)$$

$$H_3 \rightarrow Z = \frac{55-38.3}{43.3/\sqrt{25}} = \frac{16.7}{8.66} = 1.92 \; (n.s.)$$