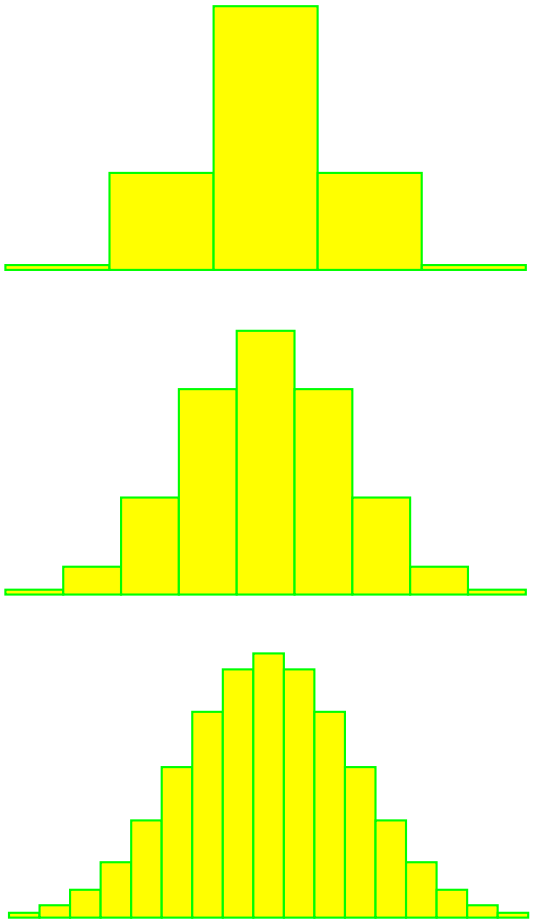
Distribuzione Normale

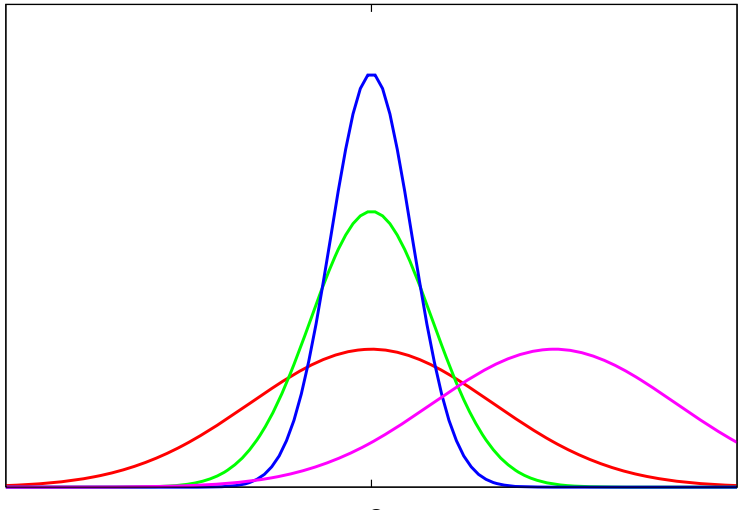


- istogramma delle frequenze di un insieme di misure relative a una grandezza che può variare con continuità
- popolazione molto numerosa, costituita da una quantità praticamente illimitata di individui (popolazione infinita)
- area dell'istogramma uguale a 1 (normalizzata)
- aumentando il numero di intervallini $n = 5, 9, 17, \ldots$ l'istogramma tende a stabilizzarsi intorno a una forma limite: la curva di distribuzione delle frequenze
- nel caso in figura: $y = ae^{-b(x-c)^2}$ distribuzione normale o gaussiana

Distribuzione Normale

Curve Gaussiane

$$y = ae^{-b(x-c)^2}$$



Se la distribuzione è di tipo gaussiano con

- *media aritmetica* μ
- *deviazione standard* σ

si ha

$$a = \frac{1}{\sigma\sqrt{2\pi}} \qquad b = \frac{1}{2\sigma^2} \qquad c = \mu$$

La corrispondente curva normale sarà

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

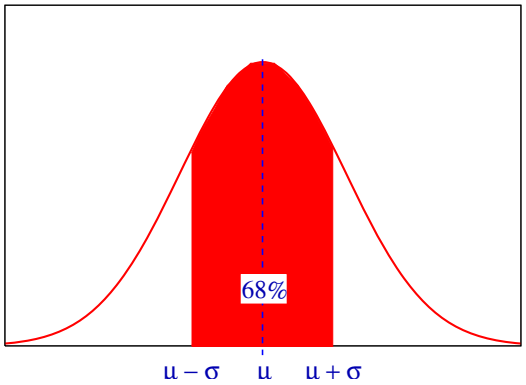
Curva normale standardizzata:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad \mu = 0 \ , \sigma = 1$$

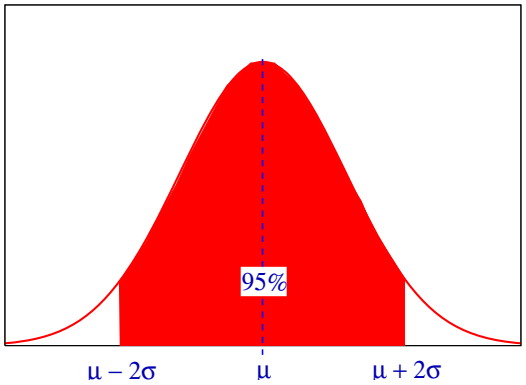
Distribuzione Normale

valori di u	Nell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Fuori dell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Nell'intervallo $[\mu + u\sigma, +\infty)$
0	0	1	$0,5$
$0,2$	$0,1586$	$0,8414$	$0,4207$
$0,4$	$0,3108$	$0,6892$	$0,3446$
$0,6$	$0,4514$	$0,5486$	$0,2743$
$0,8$	$0,5762$	$0,4238$	$0,2119$
1	$0,6826$	$0,3174$	$0,1587$
$1,2$	$0,7698$	$0,2302$	$0,1151$
$1,4$	$0,8384$	$0,1616$	$0,0808$
$1,6$	$0,8904$	$0,1096$	$0,0548$
$1,8$	$0,9282$	$0,0718$	$0,0359$
2	$0,9544$	$0,0456$	$0,0228$
$2,2$	$0,9722$	$0,0278$	$0,0139$
$2,4$	$0,9836$	$0,0164$	$0,0082$
$2,6$	$0,9906$	$0,0094$	$0,0047$
$2,8$	$0,9950$	$0,0050$	$0,0025$
3	$0,9974$	$0,0026$	$0,0013$
$3,2$	$0,9986$	$0,0014$	$0,0007$

Fissati due valori x_0, x_1 sull'asse delle ascisse, l'area sottesa dal grafico sull'intervallo $[x_0, x_1]$ rappresenta la porzione di misure che cadono nell'intervallo considerato.



Nell'intervallo $[\mu - \sigma, \mu + \sigma]$ cade circa il 68% delle misure



Nell'intervallo $[\mu - 2\sigma, \mu + 2\sigma]$ cade circa il 95% delle misure

Distribuzione Normale – Esercizi

Esercizio 1. Supponendo che la distribuzione dei pesi degli individui di una popolazione sia gaussiana con media $\mu = 61\,\text{Kg}$ e deviazione standard *(scarto quadratico medio)* $\sigma = 5\,\text{Kg}$,

(a) scrivere l'equazione della gaussiana relativa ai pesi di tale popolazione;

(b) calcolare la percentuale di individui il cui peso è compreso tra $59\,\text{Kg}$ e $63\,\text{Kg}$.

Distribuzione Normale – Esercizi

Esercizio 2. Le altezze h di un gruppo di reclute sono distribuite con buona approssimazione secondo una curva gaussiana con media $\mu = 170\,\text{cm}$ e deviazione standard *(scarto quadratico)* $\sigma = 5\,\text{cm}$. Le divise sono disponibili in 5 taglie:

1. per individui di altezza $\leq 161\,\text{cm}$
2. per individui di altezza compresa tra 161 e 167 cm
3. per individui di altezza compresa tra 167 e 173 cm
4. per individui di altezza compresa tra 173 e 179 cm
5. per individui di altezza $> 179\,\text{cm}$.

Stimare il numero delle divise delle varie taglie sapendo che le reclute sono 750.

Distribuzione Normale – Esercizi

Soluzione: si tratta di stimare la percentuale di reclute che cade in ciascuna delle quattro differenti classi di altezza:

1. per $h \leq 161 = 170 - 1.8\,\sigma \;\Rightarrow\;$ 3.6% (27 reclute)
2. per $161 < h \leq 167 \;\Rightarrow\; h \in (170 - 1.8\,\sigma\,,\, 170 - 0.6\,\sigma] \;\Rightarrow\;$ 23.8% ($\simeq$ 179 reclute)
3. per $167 < h \leq 173 \;\Rightarrow\; h \in (170 - 0.6\,\sigma\,,\, 170 + 0.6\,\sigma] \;\Rightarrow\;$ 45.1% ($\simeq$ 338 reclute)
4. per $173 < h \leq 179 \;\Rightarrow\; h \in (170 + 0.6\,\sigma\,,\, 170 + 1.8\,\sigma] \;\Rightarrow\;$ 23.8% ($\simeq$ 179 reclute)
5. per $h > 179 = 170 + 1.8\,\sigma \;\Rightarrow\;$ 3.6% (27 reclute)

Tabella Curva Gaussiana

valori di u	Nell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Fuori dell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Nell'intervallo $[\mu + u\sigma, +\infty)$
0	0	1	$0,5$
$0,2$	$0,1586$	$0,8414$	$0,4207$
$0,4$	$0,3108$	$0,6892$	$0,3446$
$0,6$	$0,4514$	$0,5486$	$0,2743$
$0,8$	$0,5762$	$0,4238$	$0,2119$
1	$0,6826$	$0,3174$	$0,1587$
$1,2$	$0,7698$	$0,2302$	$0,1151$
$1,4$	$0,8384$	$0,1616$	$0,0808$
$1,6$	$0,8904$	$0,1096$	$0,0548$
$1,8$	$0,9282$	$0,0718$	$0,0359$
2	$0,9544$	$0,0456$	$0,0228$
$2,2$	$0,9722$	$0,0278$	$0,0139$
$2,4$	$0,9836$	$0,0164$	$0,0082$
$2,6$	$0,9906$	$0,0094$	$0,0047$
$2,8$	$0,9950$	$0,0050$	$0,0025$
3	$0,9974$	$0,0026$	$0,0013$
$3,2$	$0,9986$	$0,0014$	$0,0007$

Tabella Curva Gaussiana

aree sottese dalla curva gaussiana sull'intervallo [μ , μ + z σ]

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,10	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,20	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,30	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,40	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,50	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,60	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,70	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,80	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,90	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,00	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,10	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,20	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,30	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,40	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,50	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,60	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,70	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,80	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,90	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	**0,4750**	0,4756	0,4761	0,4767
2,00	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,10	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,20	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,30	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,40	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,50	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	**0,4951**	0,4952
2,60	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,70	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,80	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,90	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,00	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Esercizi

Esercizio 3. Il diametro di certe biglie di acciaio segue una distribuzione gaussiana di media $\mu = 6.2$mm e deviazione standard $\sigma = 0.05$mm. Dire quale è la percentuale di biglie con diametro compreso tra 6.3mm e 6.35mm.

Soluzione: $[6.3\,,\,6.35] = [\mu + 2\sigma, \mu + 3\sigma] \;\Rightarrow\; 2.15\%$

Distribuzione Normale – Esercizi

Esercizio 4. Sapendo che una certa famiglia di dati segue una distribuzione gaussiana di media $\mu = 8$ e deviazione standard $\sigma = 5$, determinare:

(a) la percentuale di dati che cadono fuori dall'intervallo $[-2, 18]$;

(b) la percentuale di dati che cadono nell'intervallo $[3, 18]$;

(c) la percentuale di dati maggiori di 10.

Soluzione: (a) 4.56% (b) 81.85% (c) 34.46%

Teorema del Limite Centrale

Problema. Determinare come la media campionaria $\overline{x}$ e la deviazione standard campionaria s misurano la media μ e la deviazione standard σ della popolazione.

È data una popolazione numerica di media μ e deviazione standard σ. Da essa estraiamo dei campioni casuali $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_i$, ..., ciascuno formato da n individui, con $n > 30$. Possiamo calcolare la media campionaria $\overline{x}_i$ di ciascun campione $\mathcal{C}_i$ ed ottenere così un nuovo insieme numerico, quello delle medie campionarie.

Come si distribuiscono le medie campionarie?

Manifestano una tendenza in un certo senso *universale*, seguendo una *legge generale*, oppure il loro comportamento dipende dalla distribuzione della popolazione?

Teorema del Limite Centrale

Teorema. Sia data una popolazione numerica infinita di media μ e deviazione standard σ da cui vengono estratti dei campioni casuali formati ciascuno da n individui, con n abbastanza grande. La distribuzione delle medie campionarie è vicina a una distribuzione gaussiana di media $\mu_{\overline{x}} = \mu$ e deviazione standard $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$.

In altre parole, anche in una popolazione che non segue il modello gaussiano, le medie campionarie, se calcolate su campioni abbastanza grandi, tendono a distribuirsi secondo una legge gaussiana.

Tabella Curva Gaussiana

aree sottese dalla curva gaussiana sull'intervallo [μ , μ + z σ]

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,10	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,20	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,30	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,40	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,50	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,60	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,70	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,80	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,90	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,00	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,10	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,20	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,30	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,40	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,50	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,60	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,70	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,80	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,90	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	**0,4750**	0,4756	0,4761	0,4767
2,00	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,10	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,20	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,30	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,40	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,50	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	**0,4951**	0,4952
2,60	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,70	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,80	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,90	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,00	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Intervalli di Confidenza

Come si utilizza il teorema del limite centrale?

Supponiamo di avere un campione casuale abbastanza grande.
Calcoliamo la *media campionaria* $\overline{x}$.

La distribuzione delle medie campionarie è gaussiana, quindi:

- il 99% dei dati cade nell'intervallo $[\mu - 2.58\,\sigma_{\overline{x}}\,,\ \mu + 2.58\,\sigma_{\overline{x}}]$, cioè per il 99% dei campioni:

$$\mu - 2.58\,\sigma_{\overline{x}} \leq \overline{x} \leq \mu + 2.58\,\sigma_{\overline{x}}$$

- il 95% dei dati cade nell'intervallo $[\mu - 1.96\,\sigma_{\overline{x}}\,,\ \mu + 1.96\,\sigma_{\overline{x}}]$, cioè per il 95% dei campioni:

$$\mu - 1.96\,\sigma_{\overline{x}} \leq \overline{x} \leq \mu + 1.96\,\sigma_{\overline{x}}$$

- ...

Intervalli di Confidenza

Lette in termini di μ, le disuguaglianze precedenti definiscono gli intervalli di confidenza per la media μ della popolazione:

- intervallo di confidenza al 99%: $\overline{x} - 2.58\,\sigma_{\overline{x}} \leq \mu \leq \overline{x} + 2.58\,\sigma_{\overline{x}}$
- intervallo di confidenza al 95%: $\overline{x} - 1.96\,\sigma_{\overline{x}} \leq \mu \leq \overline{x} + 1.96\,\sigma_{\overline{x}}$
- ...

L'ampiezza degli intervalli di confidenza è espressa in funzione di

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

che dipende dalla deviazione standard, incognita, della popolazione.

Intervalli di Confidenza

Si può dimostrare che la *deviazione standard campionaria*

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

approssima bene la deviazione standard σ della popolazione.

Quindi, possiamo scrivere gli intervalli di confidenza nella forma:

- al 99%, $\overline{x} - 2.58 \dfrac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + 2.58 \dfrac{s}{\sqrt{n}}$
- al 95%, $\overline{x} - 1.96 \dfrac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + 1.96 \dfrac{s}{\sqrt{n}}$
- ...

che dipende solo dai dati campionari $(\overline{x}\,,\, s\,,\, n)$.

Esercizi

Esercizio 1. Si vuole stimare l'età media degli utenti di una biblioteca civica. A questo scopo si seleziona un campione casuale composto da $n = 100$ persone avente media $\overline{x} = 29$ anni e deviazione standard $s = 8$ anni. Trovare intervalli di confidenza per l'età media μ al 95% ed al 99%. Scrivere i risultati arrotondati alla seconda cifra decimale.

Esercizi

Soluzione: poiché il campione è composto da $n = 100 > 30$ individui, possiamo applicare il teorema del limite centrale.

- Nel 95% dei casi la media μ appartiene all'intervallo

$$29 - 1.96 \cdot \frac{s}{\sqrt{n}} \leq \mu \leq 29 + 1.96 \cdot \frac{s}{\sqrt{n}}$$

  Inserendo i dati, concludiamo che $27.43 \leq \mu \leq 30.57$ con un grado di fiducia pari al 95%.

- Nel 99% dei casi la media μ appartiene all'intervallo

$$29 - 2.58 \cdot \frac{s}{\sqrt{n}} \leq \mu \leq 29 + 2.58 \cdot \frac{s}{\sqrt{n}}$$

  Inserendo i dati, concludiamo che $26.94 \leq \mu \leq 31.06$ con un grado di fiducia pari al 99%.

Esercizi

Esercizio 2. Nell'esercizio precedente si supponga che i dati $\overline{x} = 29$ anni e deviazione standard $s = 8$ anni siano stati ottenuti da un campione casuale composto da $n = 400$ persone. Trovare i nuovi intervalli di confidenza per l'età media μ al 95% ed al 99%.

Esercizi

Soluzione: l'unico cambiamento riguarda il fatto che

$$\frac{s}{\sqrt{n}} = \frac{8}{20} = \frac{2}{5}.$$

L'intervallo di confidenza al 95% è $28.22 \leq \mu \leq 29.78$.

L'intervallo di confidenza al 99% è $27.97 \leq \mu \leq 30.03$.

Rispetto all'esercizio precedente gli intervalli si sono ridotti di ampiezza, dunque la stima è più precisa. Il maggior grado di precisione è dovuto al fatto che i dati provengono da un campione più ampio.

Esercizi

Esercizio 3. Si vuole stimare l'età media μ di una popolazione di pazienti affetti da una certa malattia. Su un campione casuale composto da 576 pazienti affetti dalla malattia risulta un'età media $\bar{x} = 12$ anni e una deviazione standard campionaria $s = 4$ anni. Trovare l'intervallo di confidenza all'89% per l'età media μ dei malati.

Soluzione: $\left[12 - 1.6 \cdot \frac{4}{\sqrt{576}}, 12 + 1.6 \cdot \frac{4}{\sqrt{576}}\right] \cong [11.73, 12.27]$

Come cambia la stima se gli stessi dati $\bar{x}, s$ sono ottenuti a partire da un campione composto da 900 pazienti?

Soluzione: $\left[12 - 1.6 \cdot \frac{4}{\sqrt{900}}, 12 + 1.6 \cdot \frac{4}{\sqrt{900}}\right] \cong [11.79, 12.21]$

Tabella Curva Gaussiana

valori di u	Nell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Fuori dell'intervallo $[\mu - u\sigma, \mu + u\sigma]$	Nell'intervallo $[\mu + u\sigma, +\infty)$
0	0	1	$0,5$
$0,2$	$0,1586$	$0,8414$	$0,4207$
$0,4$	$0,3108$	$0,6892$	$0,3446$
$0,6$	$0,4514$	$0,5486$	$0,2743$
$0,8$	$0,5762$	$0,4238$	$0,2119$
1	$0,6826$	$0,3174$	$0,1587$
$1,2$	$0,7698$	$0,2302$	$0,1151$
$1,4$	$0,8384$	$0,1616$	$0,0808$
$1,6$	$0,8904$	$0,1096$	$0,0548$
$1,8$	$0,9282$	$0,0718$	$0,0359$
2	$0,9544$	$0,0456$	$0,0228$
$2,2$	$0,9722$	$0,0278$	$0,0139$
$2,4$	$0,9836$	$0,0164$	$0,0082$
$2,6$	$0,9906$	$0,0094$	$0,0047$
$2,8$	$0,9950$	$0,0050$	$0,0025$
3	$0,9974$	$0,0026$	$0,0013$
$3,2$	$0,9986$	$0,0014$	$0,0007$

Esercizi

Esercizio 4. Nella seguente tabella sono riportati, raggruppati in classi, i dati relativi al diametro x delle fibre del pelo di un campione di 50 pecore affette da dermatofilosi.

diametro (10^{-1} mm)	f_i
17.75 – 19.75	4
19.75 – 21.75	10
21.75 – 23.75	20
23.75 – 25.75	11
25.75 – 27.75	4
27.75 – 29.75	1
	50

(a) Calcolare media e deviazione standard campionarie, utilizzando la trasformazione $y = \frac{1}{2}(x - 18.75)$.

(b) Rappresentare graficamente la distribuzione delle frequenze.

(c) Rappresentare graficamente la distribuzione delle frequenze cumulate, indicando sull'asse x la posizione della mediana.

(d) Costruire l'intervallo di confidenza al 95% del diametro medio μ nella popolazione.

Esercizi

La trasformazione suggerita $y = \frac{1}{2}(x - 18.75)$ semplifica il calcolo della media e della deviazione standard campionarie:

diametro (10^{-1} mm)	x	f	y	$f \cdot y$	y^2	$f \cdot y^2$
17.75 – 19.75	18.75	4	0	0	0	0
19.75 – 21.75	20.75	10	1	10	1	10
21.75 – 23.75	22.75	20	2	40	4	80
23.75 – 25.75	24.75	11	3	33	9	99
25.75 – 27.75	26.75	4	4	16	16	64
27.75 – 29.75	28.75	1	5	5	25	25
		50		104		278

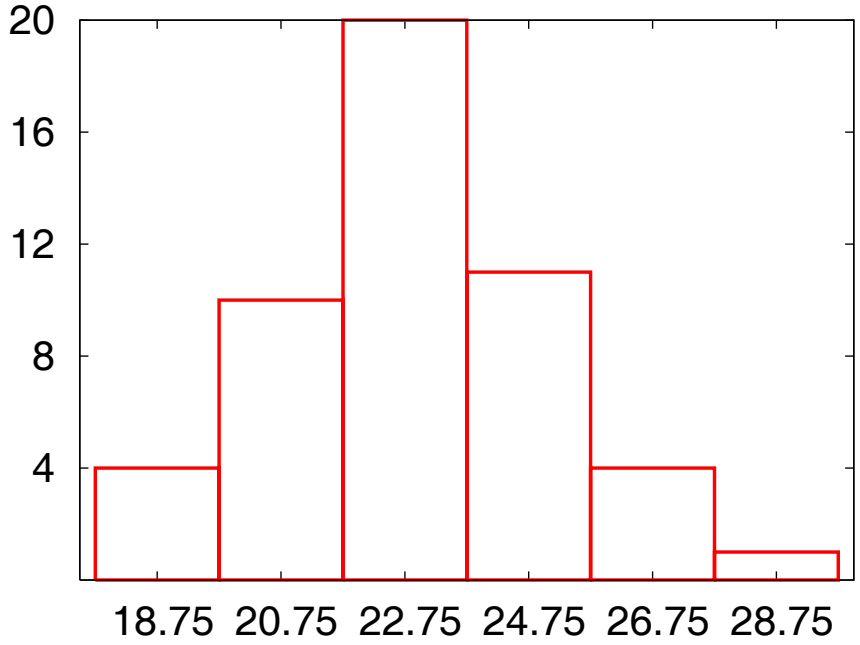
Quindi,

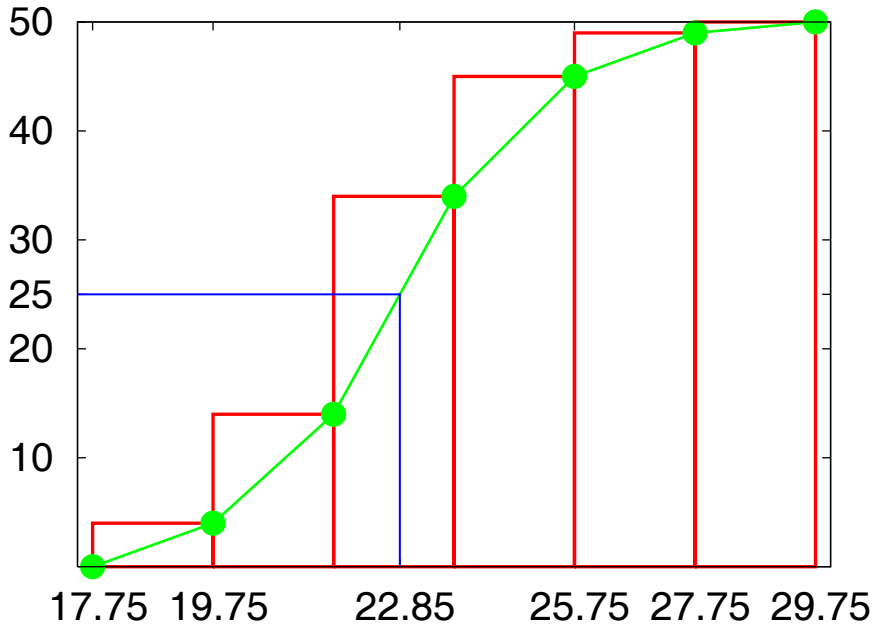
$$\bar{y} = \frac{104}{50} = 2.08 \qquad s_y^2 = \frac{1}{49}\left[278 - 50 \cdot 2.08^2\right] = 1.2588 \qquad s_y = \sqrt{s_y^2} = 1.12$$

Tornando a x mediante la trasformazione inversa $x = 2y + 18.75$,

$$\bar{x} = 2 \cdot (2.08) + 18.75 = 22.91\,(\cdot 10^{-1}\,\text{mm}) \qquad s_x = 2 \cdot (1.12) = 2.24\,(\cdot 10^{-1}\,\text{mm})$$

Esercizi


20
16
12
8
4
18.75
20.75
22.75
24.75
26.75
28.75


50
40
30
25
20
10
17.75
19.75
22.85
25.75
27.75
29.75

Esercizi

Intervallo di confidenza al 95%

Utilizzando la tavola si trova

$$\bar{x} \pm 1.96 \cdot \frac{s_x}{\sqrt{n}} = 22.91 \pm 1.96 \cdot \frac{2.24}{\sqrt{50}} \simeq 22.91 \pm 0.62\,(10^{-1}\,\mathrm{mm})$$

L'intervallo di confidenza al 95% del diametro medio μ (espresso in 10^{-1} mm) è

$$22.91 - 0.62 \leq \mu \leq 22.91 + 0.62$$

Test d'Ipotesi – Introduzione

Uno degli scopi più importanti di un'analisi statistica è quello di utilizzare i dati provenienti da un campione per fare inferenza sulla popolazione da cui è stato estratto il campione.

Si è visto come, utilizzando la media campionaria, si può stimare il corrispondente parametro della popolazione.

Un'altra applicazione del teorema del limite centrale sono i test d'ipotesi. Si vuole sottoporre a test un'ipotesi su un parametro di una popolazione, con lo scopo di decidere, esaminando un campione estratto dalla popolazione, se l'affermazione (cioè, l'ipotesi) riguardante il parametro è vera o falsa.

Test d'Ipotesi – Introduzione

Ad esempio:

- il responsabile della produzione di un'azienda può affermare che le confezioni prodotte hanno un peso medio di 250 g
- un medico può ipotizzare che un farmaco sia efficace nel 90% dei casi in cui viene somministrato

Con la verifica delle ipotesi si può determinare se tali congetture sono compatibili con i dati disponibili dal campione.

Test d'Ipotesi – Esempi

Esempio 1. Si vuole verificare se le lattine di caffè, confezionate automaticamente da una ditta, contengono in media il peso dichiarato di $\mu = 250\,\mathrm{g}$.

A tale scopo si prende un campione di 50 *lattine, se ne pesa il contenuto e si calcola il peso medio, per stabilire se il peso medio del campione differisce da* 250 *g.*

Esempio 2. Si vuole sottoporre a test l'affermazione di un produttore di vernici secondo cui il tempo medio di asciugatura di una nuova vernice è non superiore a 30 minuti.

A tale scopo si prende un campione di 40 *lattine di vernice, si effettuano* 40 *prove di verniciatura con la vernice delle diverse confezioni e si calcola il tempo medio di asciugatura. L'intenzione è rifiutare l'affermazione del produttore se la media osservata supera il valore di* 30 *minuti, o accettarla in caso contrario.*

Test d'Ipotesi – Definizioni

Un'ipotesi formulata in termini di parametri di una popolazione, come media o varianza, è detta **ipotesi statistica**.

Il procedimento che consente di rigettare o accettare un'ipotesi statistica, utilizzando i dati di un campione abbastanza numeroso, viene chiamato **test d'ipotesi**.

Le possibili conclusioni di un test d'ipotesi sono:

- l'ipotesi statistica è rifiutata
- l'ipotesi statistica non è rifiutata

Test d'Ipotesi – Ipotesi Zero

La distribuzione gaussiana delle medie consente di sottoporre ad esame critico ipotesi effettuate su una popolazione.

1. Supponiamo venga fatta un'affermazione che localizza la media μ della popolazione (ipotesi zero).
2. Per verificare l'attendibilità dell'ipotesi, si seleziona un campione casuale sufficientemente grande ($n > 30$) di cui si calcola la media campionaria $\overline{x}$ e la deviazione standard campionaria s.
3. Si misura la distanza, in termini di deviazioni standard, di μ dalla media osservata "sul campo" $\overline{x}$.
4. Quanto più $\overline{x}$ si allontana da μ, tanto più diventiamo sospettosi circa la validità dell'ipotesi riguardante la media e siamo condotti a rigettare l'ipotesi.

Test d'Ipotesi – Livello di Significatività

Così facendo ci assumiamo un rischio, cioè quello che il campione scelto avesse media $\overline{x}$ realmente molto lontana da μ e che la media μ fosse accettabile.

Il livello di rischio di prendere una decisione sbagliata, che siamo disposti a correre, dipende dalle circostanze.

Solitamente si accetta un rischio dell'1% o del 5%. Il rischio di prendere la decisione sbagliata sulla scorta dei dati del campione è detto livello di significatività del test.

Test d'Ipotesi – Esempi

Esempio 1. È stato affermato che il peso medio degli individui adulti di una certa nazione è $\mu = 68.5$ Kg. Volendo sottoporre questa ipotesi a verifica, si considera un campione casuale di 625 individui che vengono pesati. Si ottiene un valor medio campionario $\overline{x} = 69.1$ Kg con una deviazione standard campionaria $s = 7$ Kg.

Con livello di significatività del 5%, *qual è l'esito del test?*

Test d'Ipotesi – Esempi

Misuriamo la distanza $\overline{x} - \mu = 0.6\,\mathrm{Kg}$ della media campionaria dalla stima di μ, contenuta nell'ipotesi zero, in termini di deviazioni standard. Poiché

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \simeq \frac{s}{\sqrt{n}} = \frac{7}{25},$$

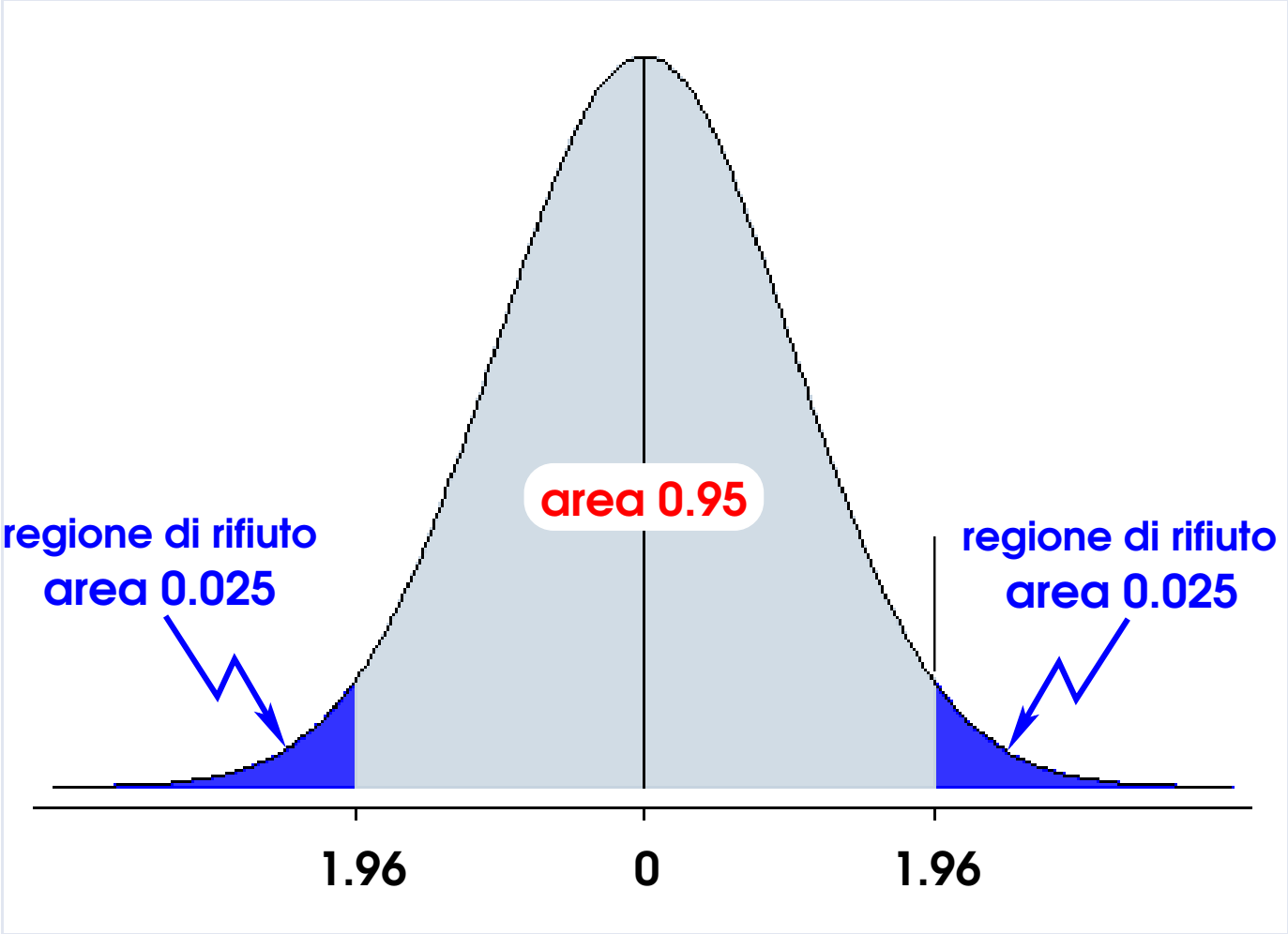
il numero u di deviazioni standard di cui $\overline{x}$ si allontana da μ soddisfa la relazione

$$u\,\frac{s}{\sqrt{n}} = \frac{7}{25}\,u = 0.6 \qquad \Rightarrow \qquad u = 2.14\,.$$

Il 95% delle medie campionarie cade nell'intervallo $\left[\mu - 1.96\,\frac{s}{\sqrt{n}}\,,\,\mu + 1.96\,\frac{s}{\sqrt{n}}\right]$.

La media campionaria analizzata cade fuori da questo intervallo (poiché $2.14 > 1.96$). Siamo quindi autorizzati a rigettare l'ipotesi su μ, assumendoci un rischio del 5%.

Test a Due Code



Test d'Ipotesi – Esempi

*Con livello di significatività dell'*1%*, qual è l'esito del test?*

Se siamo disposti ad assumerci solo l'1% di rischio di prendere la decisione sbagliata, l'esito del test è diverso.

Il 99% delle medie campionarie cade nell'intervallo

$$\left[\mu - 2.58\,\frac{s}{\sqrt{n}}\,,\ \mu + 2.58\,\frac{s}{\sqrt{n}}\right].$$

La media campionaria $\overline{x}$ appartiene a questo intervallo (poiché $2.14 < 2.58$). L'ipotesi su μ è quindi compatibile con il risultato ottenuto dal campione casuale, con un livello di significatività dell'1%.

Test d'Ipotesi – Esempi

Esempio 2. Una compagnia aerea afferma che il peso medio del bagaglio dei passeggeri dei suoi voli di linea è $\mu = 19.8\,\text{Kg}$.

Per sottoporre a verifica tale ipotesi si considera un campione casuale di 324 passeggeri. Si ottiene un peso medio campionario $\overline{x} = 20.3\,\text{Kg}$, con scarto quadratico medio campionario $s = 3.6\,\text{Kg}$.

*Con livello di significatività dell'*1%*, qual è l'esito del test?*

Test d'Ipotesi – Esempi

Si ha $\dfrac{s}{\sqrt{n}} = \dfrac{3.6}{18} = 0.2$, da cui

$$\overline{x} - \mu = u\,\frac{s}{\sqrt{n}} \quad \Rightarrow \quad 0.5 = 0.2 \cdot u \quad \Rightarrow \quad u = 2.5\,.$$

La media campionaria dista, da quella contenuta nell'ipotesi zero, 2.5 scarti quadratici medi e dunque è ancora interna all'intervallo

$$\left[\mu - 2.58\frac{s}{\sqrt{n}}\,,\ \mu + 2.58\frac{s}{\sqrt{n}}\right] \qquad \text{(poiché } 2.50 < 2.58\text{)}$$

entro cui cade il 99% dei dati.

L'esito del test è: non esistono elementi sufficienti per rigettare l'ipotesi su μ, al livello di significatività richiesto.

Test d'Ipotesi – Esempi

Se i valori di $\overline{x}$ e s sono ottenuti da un campione di 400 *passeggeri, come cambia l'esito del test?*

Se il campione è formato da 400 passeggeri: $\dfrac{s}{\sqrt{n}} = \dfrac{3.6}{20} = 0.18$ che corrisponde a una distanza tra $\overline{x}$ e μ pari a 2.78 scarti quadratici medi.

Dunque, $\overline{x}$ cade fuori dall'intervallo

$$\left[\mu - 2.58\frac{s}{\sqrt{n}}\,,\ \mu + 2.58\frac{s}{\sqrt{n}}\right] \qquad \text{(poiché } 2.78 > 2.58\text{)}$$

La conclusione del test è: rigettiamo l'ipotesi formulata su μ, con un margine di rischio dell'1%.

L'esempio mostra il ruolo della grandezza del campione nell'esito del test.

Test d'Ipotesi – Errori

Se l'ipotesi zero è vera, ma erroneamente viene rigettata, si commette un errore di primo tipo.

Se l'ipotesi zero è falsa, ma erroneamente non viene rigettata, si commette un errore di secondo tipo.

Test d'Ipotesi – Esempi

Esempio 3. Il reddito medio μ di una famiglia che abita in una certa regione *non supera* i 12500 Euro ($\mu \leq 12500$).

Sottoporre a verifica tale ipotesi.

L'ipotesi sulla media è più sfumata rispetto agli esempi precedenti in quanto si limita a imporre un limite superiore alla media, non un valore preciso.

Si può adottare una procedura simile a quella vista in precedenza.

Test d'Ipotesi – Esempi

1. Si considera la media $\overline{x}$ ottenuta da un campione sufficientemente ampio e la si confronta con il limite superiore per μ di 12500 Euro.
2. Se $\overline{x}$ risulta inferiore a 12500 Euro, non si ha alcun elemento per rigettare l'ipotesi.
3. I dubbi sulla bontà dell'ipotesi cominciano ad affiorare se $\overline{x}$ supera i 12500 Euro.

Test d'Ipotesi – Esempi

Come si procede se $\overline{x}$ supera i 12500 Euro?

- Se si vuole un livello di significatività del 5%, si avranno elementi per rigettare l'ipotesi quando $\overline{x}$ si lascia *a sinistra* almeno il 95% dei dati. Questo accade non appena si esce dall'intervallo $(-\infty\,,\,\mu + 1.64\,\frac{s}{\sqrt{n}}\,]$.
- Se si vuole un livello di significatività dell'1%, l'ipotesi diventa rigettabile (con un margine di rischio dell'1%) non appena $\overline{x}$ esce dall'intervallo $(-\infty\,,\,\mu + 2.33\,\frac{s}{\sqrt{n}}\,]$, entro il quale cade il 99% dei dati.

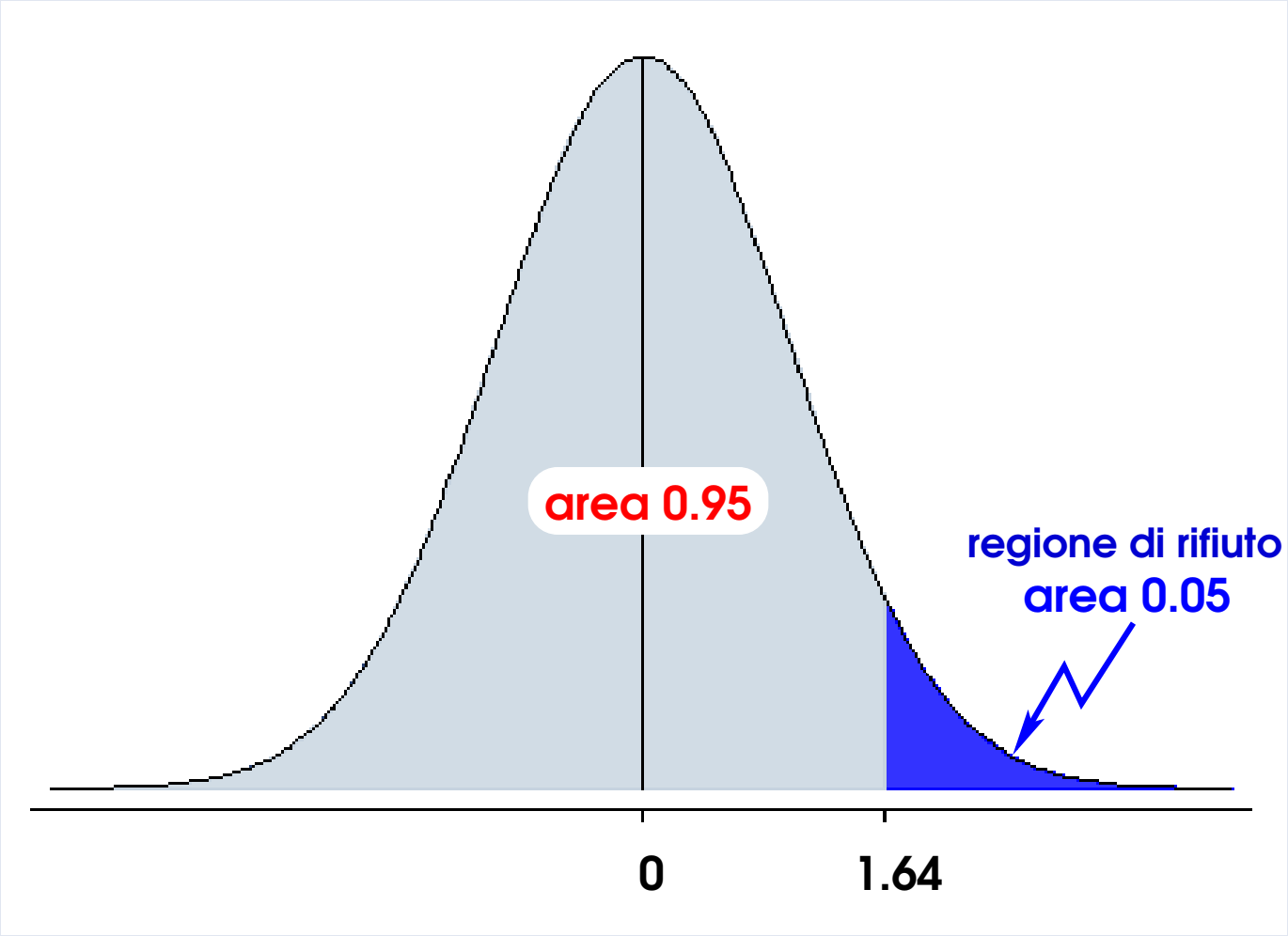
n è la numerosità del campione, s è la deviazione standard campionaria.

Curva Gaussiana

aree sottese dalla curva gaussiana
sull' intervallo [μ , μ + z σ]

z	0.00	0.01	0.02	0.03	0.04	0.05
0.00	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199
0.10	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596
0.20	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987
0.30	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368
0.40	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736
0.50	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088
0.60	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422
0.70	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734
0.80	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023
0.90	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289
1.00	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531
1.10	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749
1.20	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944
1.30	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115
1.40	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265
1.50	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394
1.60	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505
1.70	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599
1.80	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678
1.90	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744
2.00	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798
2.10	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842
2.20	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878
2.30	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906
2.40	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929
2.50	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946
2.60	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960
2.70	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970
2.80	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978
2.90	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984
3.00	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989

Test a Una Coda



Test d'Ipotesi – Esercizi

Esercizio 1. Un ricercatore intende saggiare, con livello di significatività del 5%, l'affermazione di una ditta farmaceutica secondo la quale il tempo che intercorre tra l'assunzione di un farmaco e la manifestazione dei primi effetti è al più di 4 minuti.

A questo scopo considera un campione casuale di 100 pazienti e trova che in media il tempo necessario per riscontrare efficacia nel farmaco è di 4 minuti e 6 secondi, con scarto quadratico medio $s = 0.6$ minuti.

Quali sono le conclusioni del test? Cambia qualcosa se i dati sperimentali hanno scarto quadratico medio $s = 0.64$ minuti?

Test d'Ipotesi – Esercizi

Convertiamo il tempo medio campionario di reazione del farmaco in forma decimale: $\overline{x} = 4$ minuti e 6 secondi $= 4.1$ minuti.

Ipotesi zero: $\mu \leq 4$ minuti. Si tratta di un *test a una coda*.

Avremo ragione di dubitare dell'affermazione della ditta solo se si misurano, come in questo caso, tempi di reazione maggiori al limite superiore dichiarato dal produttore.

Calcoliamo il valore di u che risolve l'equazione:

$$\overline{x} - \mu = 0.1 = u \frac{s}{\sqrt{n}} = 0.06\, u \quad \Rightarrow \quad u = 1.67$$

Dunque, $\overline{x}$ si trova fuori dall'intervallo $\left(-\infty\,,\, \mu + 1.64 \frac{s}{\sqrt{n}}\right]$, entro il quale cade il 95% dei dati.

Al livello di significatività del 5%, l'ipotesi zero è da respingere.

Test d'Ipotesi – Esercizi

Se $s = 0.64$,

$$\overline{x} - \mu = 0.1 = u\,\frac{s}{\sqrt{n}} = 0.064\,u \qquad \Rightarrow \qquad u = 1.56$$

Dunque, $\overline{x}$ cade nell'intervallo $\left(-\infty\,,\,\mu + 1.64\,\frac{s}{\sqrt{n}}\right]$.

Al livello di significatività richiesto l'ipotesi non è rigettabile.

L'esempio evidenzia l'importanza dell'accuratezza della verifica sperimentale. Lo scarto quadratico medio misura infatti la dispersione dei dati attorno al valor medio e indica il grado di imprecisione nelle misure. Le conclusioni tratte dall'esame dei dati sono difformi nei due casi.

Test d'Ipotesi – Esercizi

Esercizio 2. Si vuole sottoporre a verifica la seguente affermazione: la spesa media per le vacanze degli italiani è inferiore a 800 Euro a testa. A questo scopo si considera un campione di 100 italiani e si osserva che la spesa media per le vacanze di questo campione è stata di 808 Euro a testa con uno scarto quadratico medio $s = 40$ Euro.

Dopo aver precisato se il test debba essere a una o due code, trarre le conclusioni se il livello di significatività è del 5%. Cosa cambia se il livello di significatività del test è dell'1%?

Test d'Ipotesi – Esercizi

Esercizio 3. Si vuole sottoporre a verifica la seguente affermazione: il peso medio degli abitanti adulti di una certa nazione è $\mu = 72\,\text{Kg}$. A questo scopo si considera un campione casuale di 100 individui, che vengono pesati. Si ottiene un peso medio campionario $\bar{x} = 73.8\,\text{Kg}$ con deviazione standard campionaria $s = 8\,\text{Kg}$.

Dopo aver precisato se il test debba essere a una o due code, trarre le conclusioni se il livello di significatività è del 5%. Cosa cambia se il livello di significatività del test è dell'1%? E se il campione fosse stato composto da 400 individui?