Ulteriori Conoscenze di Informatica e Statistica

Carlo Meneghini

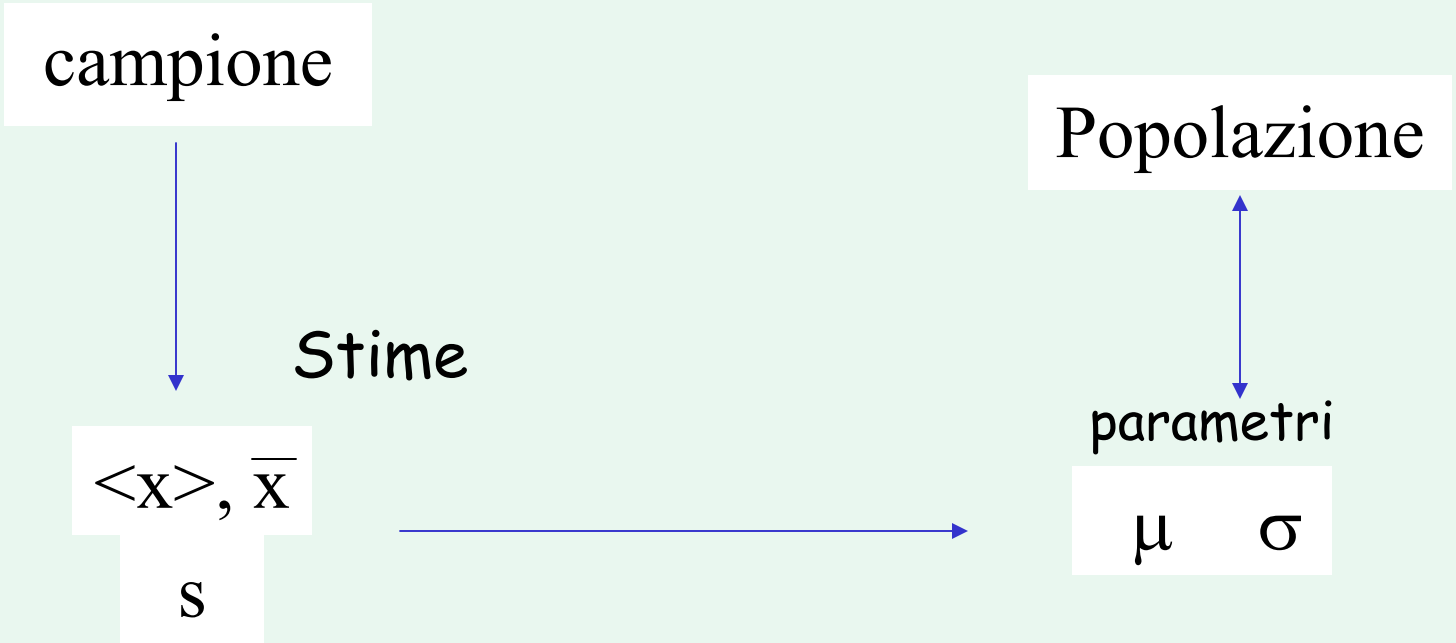
Dip. di fisica – via della Vasca Navale 84,

st. 83 (I piano) tel.: 06 55 17 72 17

meneghini@fis.uniroma3.it

I risultati di un esperimento sono variabili aleatorie.

Un esperimento non consente di esaminare ogni elemento di una popolazione o di effettuare tutte le misure possibili.



Dato un campione *n* estratto da una popolazione N è possibile fornire una stima (<x>, s) dei parametri reali della distribuzione (μ, σ).

I risultati ottenuti su un campione rappresentano una stima dei valori "veri"

I valori stimati sono variabili aleatorie

Quanto sono accurate queste stime?

Popolazione

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{k=1}^{n_c} x_k p(x_k)$$

Valore atteso (media)

Varianza

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{k=1}^{n_c} p(x_k)(x_k - \mu)^2$$

Campione

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

Media campionaria

Varianza campionaria

$$s^2 = \frac{1}{m-1}\sum_{j=1}^{m}(x_j - \bar{x})^2$$

Teorema del limite centrale

La distribuzione delle medie campionari ($\langle x \rangle_i$) segue una distribuzione normale indipendentemente dalla distribuzione della popolazione d'origine

Il valor medio della distribuzione delle media campionarie è uguale alla media della popolazione d'origine

La deviazione standard dell'insieme di tutte le medie campionarie (errore standard della media σ_x) è una funzione della deviazione standard della popolazione originaria e del numero di elementi del campione.

$$\frac{1}{\sqrt{2\pi\sigma_{\bar{x}}^2}} e^{-\frac{(\bar{x}_i - \mu)^2}{2\sigma_{\bar{x}}^2}}$$

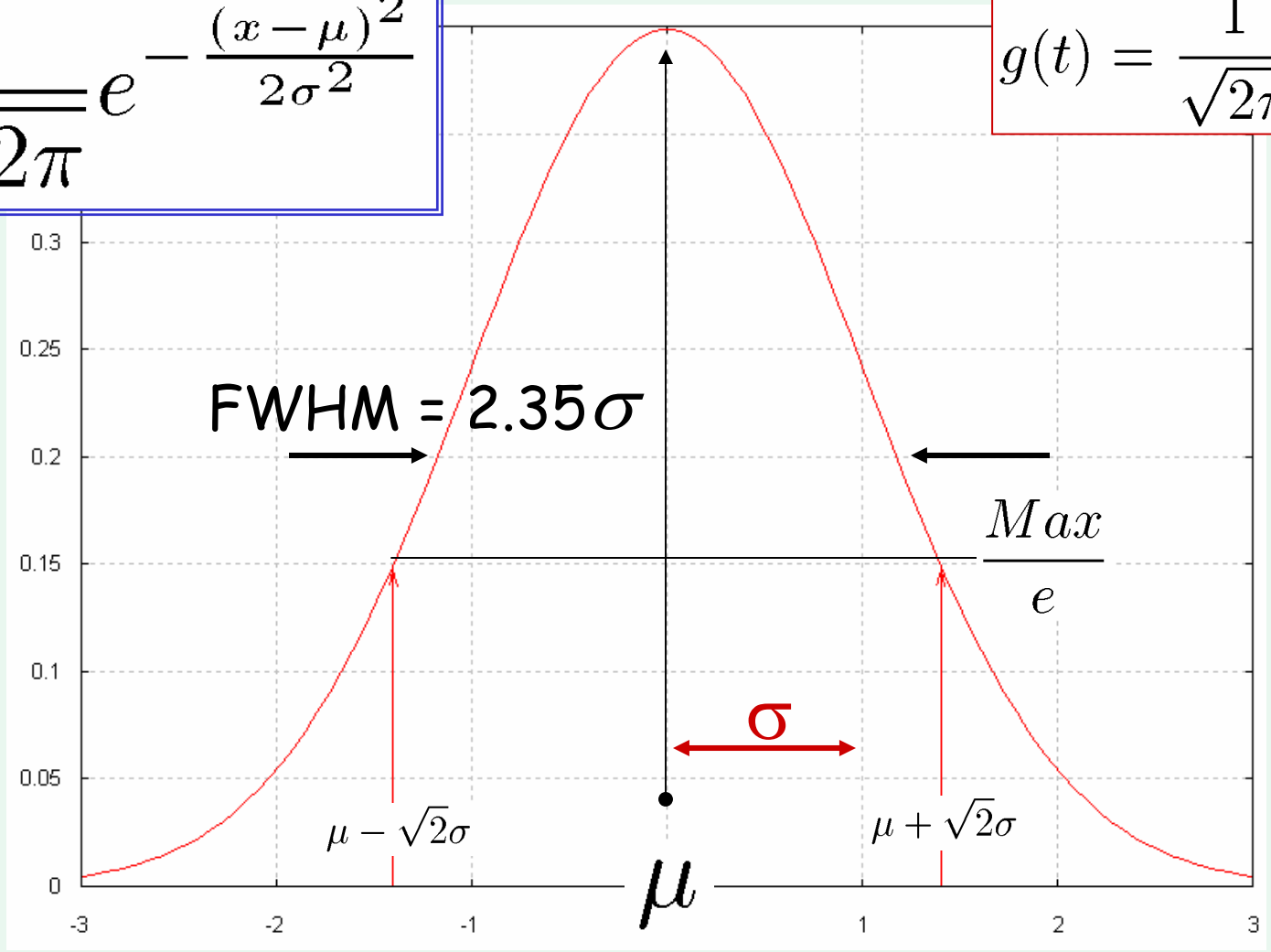
nota: dev.st. della popolazione

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}}$$

Proprietà della distribuzione di Gauss

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

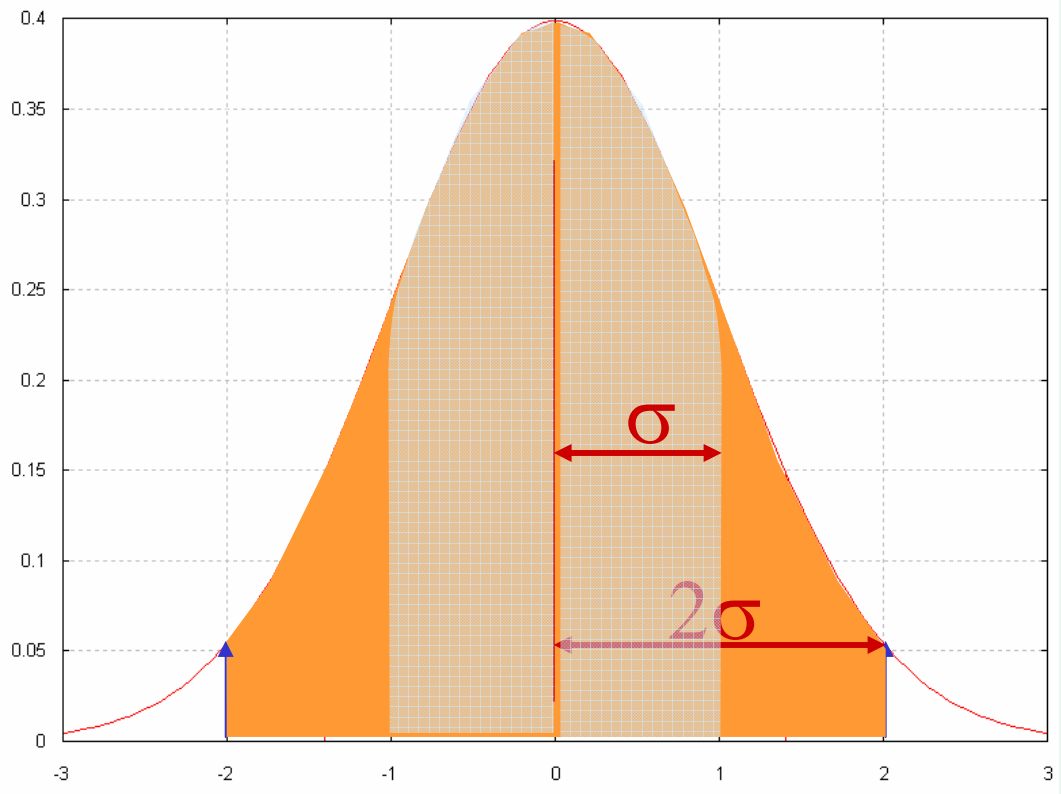
$$g(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



$$\int_{\mu-\sigma}^{\mu+\sigma} g(x)dx = 0.68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} g(x)dx = 0.95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} g(x)dx = 0.997$$



Date due variabili aleatorie indipendenti X_a, X_b caratterizzate da μ_a σ_a, μ_b σ_b, la variabile $Z = X_a + X_b$ è una variabile aleatoria con:

$$\mu_z = \mu_a + \mu_b$$

$$\sigma_z = \sigma_a + \sigma_b$$

Stima della media

$$\frac{1}{\sqrt{2\pi\sigma_{\bar{x}}^2}} e^{-\frac{(\bar{x}_i-\mu)^2}{2\sigma_{\bar{x}}^2}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}}$$

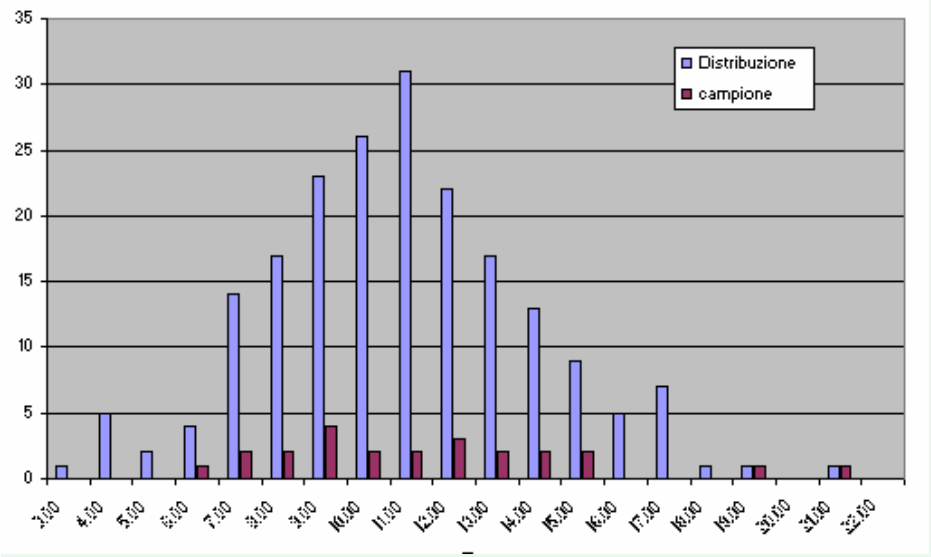
L'errore standard della media

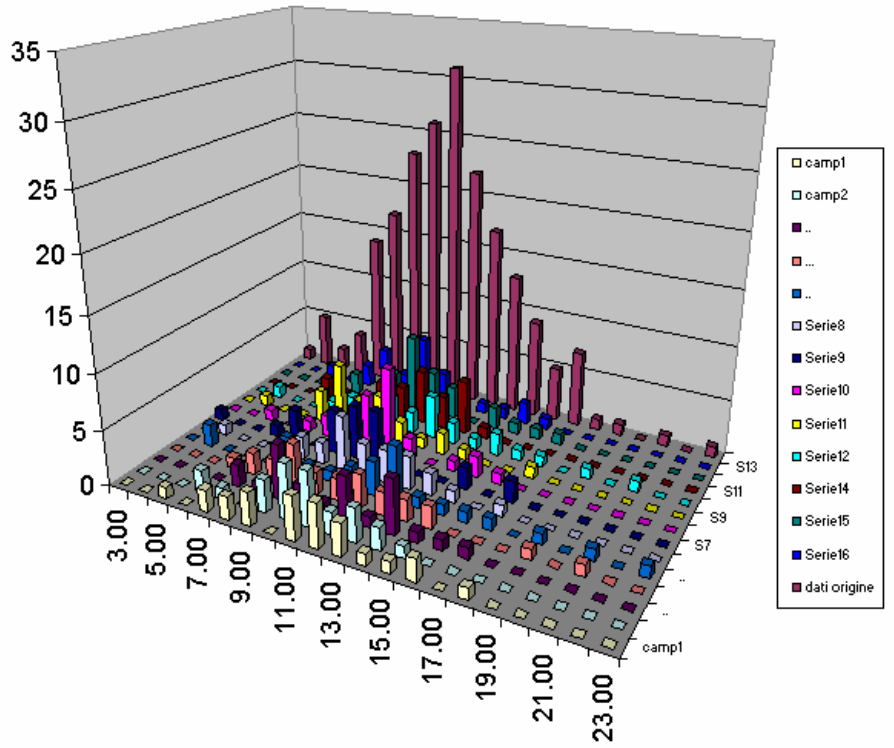
$\sigma_{\bar{x}}$ indica il grado di incertezza da associare alla stima della media ottenuta utilizzando un campione dell'intera popolazione

Interpretazione: se effettuo diversi campionamenti (al limite tutti i possibili campionamenti) da una data popolazione le medie ottenute per i vari campionamenti si distribuiscono attorno al valore μ. La larghezza della distribuzione dei valori medi sarà tanto più stretta intorno al valore vero quanti più elementi scelgo per ogni campionamento (m).

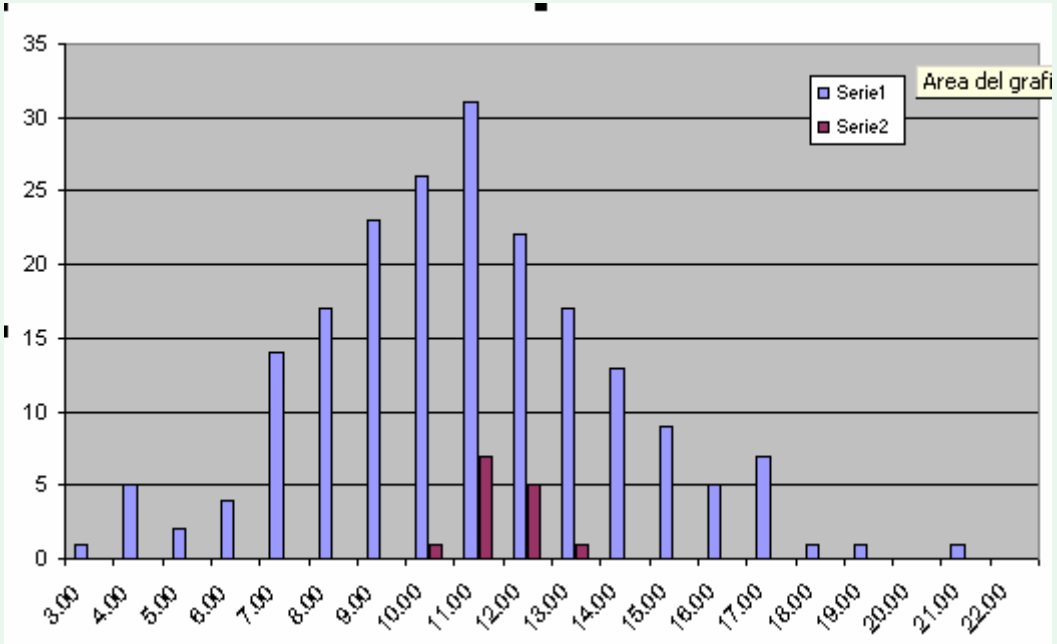
ATTENZIONE: l'errore standard sulla media è funzione della deviazione standard della distribuzione ma non è la deviazione standard della distribuzione.

La distribuzione reale confrontata con un ipotetico campione





La distribuzione reale confrontata con più campioni



La distribuzione reale confrontata la distribuzione delle medie

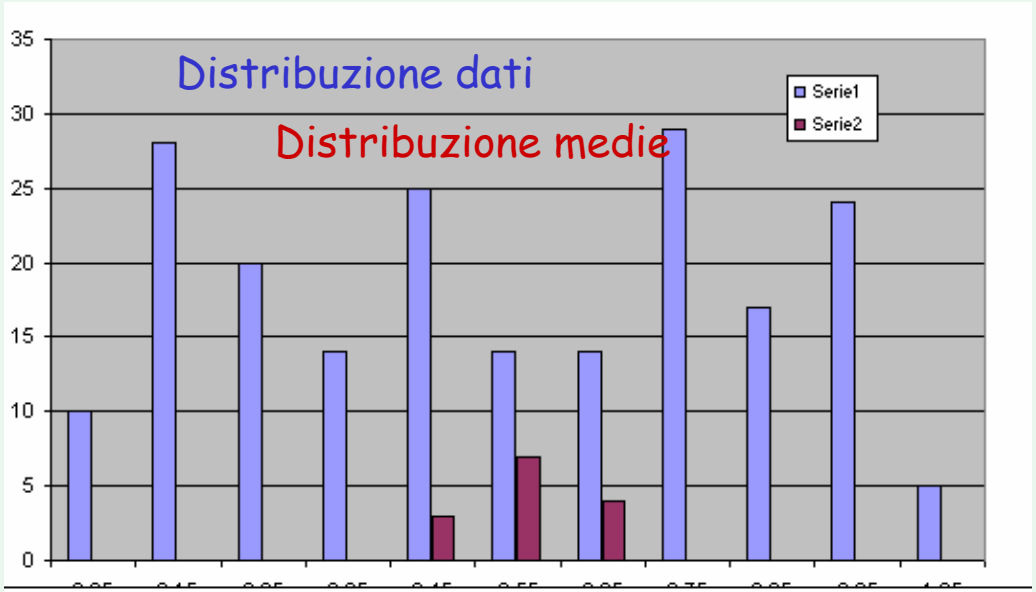
Teorema del limite centrale per una distribuzione uniforme dei dati.

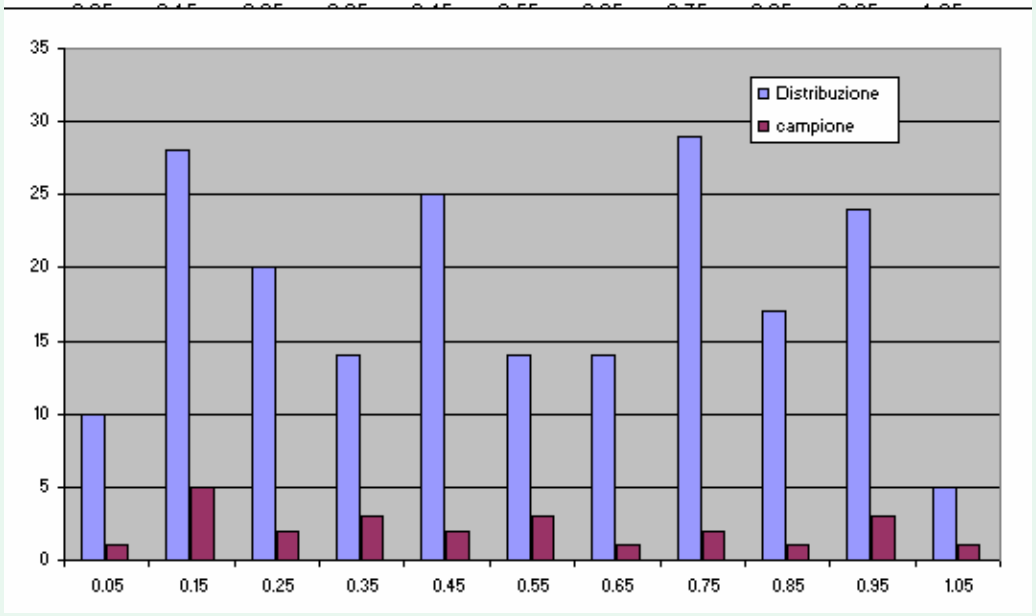
Il teorema del limite centrale mi dice che:

- la distribuzione delle media campionarie ha una forma gaussiana qualunque sia la distribuzione della popolazione,

- il valore medio della distribuzione delle medie campionarie è il valore medio della popolazione

la varianza della distribuzione delle medie campionarie è la varianza della popolazione diviso m (numero di campionamenti)





Accuratezza delle stime

Il valor medio ottenuto da **un** solo campione di **m** elementi è una stima del valore aspettato della popolazione.

L'errore standard della media rappresenta una stima dell'errore fatto nella stima del valore atteso. Se non conosco la deviazione standard della popolazione utilizzo la stima della deviazione standard (s) per valutare l'errore sulla media

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}} \sim \frac{s}{\sqrt{m}}$$

Risultato di un'osservazione:

$$\bar{x} \pm \sigma_{\bar{x}}$$

$$\bar{x} - \sigma_{\bar{x}} \leq \text{valore vero} \leq \bar{x} + \sigma_{\bar{x}}$$

Nota: attenziona al significato di queste formule

Accuratezza delle stime

Per migliorare la stima del valore atteso si puó ripetere l'esperimento utilizzando K campioni indipendenti

In questo caso la migliore stima del valore atteso è la media delle medie campionarie:

$$\bar{x} = \frac{1}{K}\sum_{i=1}^{K}\bar{x}_i$$

Utilizzando K campioni indipendenti l'errore standard della media si calcola (radice quadrata) dalla varianza della distribuzione delle medie campionarie:

varianza:

$$\sigma^2_{\bar{x}} = \frac{1}{K}\sum_{i=1}^{K}(\bar{x}_i - \bar{x})^2$$

Stima della varianza:

$$s^2_{\bar{x}} = \frac{1}{K-1}\sum_{i=1}^{K}(\bar{x}_i - \bar{x})^2$$

Standardizzazione e normalizzazione

La distribuzione delle medie campionarie $\bar{x}$ su campioni di m elementi segue una distribuzione **normale** indipendentemente dalla distribuzione della popolazione d'origine

$$\frac{1}{\sqrt{2\pi\sigma_{\bar{x}}^2}} e^{-\frac{(\bar{x}_i-\mu)^2}{2\sigma_{\bar{x}}^2}}$$

La variabile: $t = \dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}}$

è una variabile aleatoria che, per m molto grande, ha una distribuzione Normale Standard (ha media nulla e varianza unitaria):

$$g(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Intervalli di Confidenza

ovvero quale è la probabilità di sbagliare la stima?

Pb.: un'osservazione su un campione di m elementi fornisce come risultato il valor medio $\bar{x}$ di una variabile aleatoria.

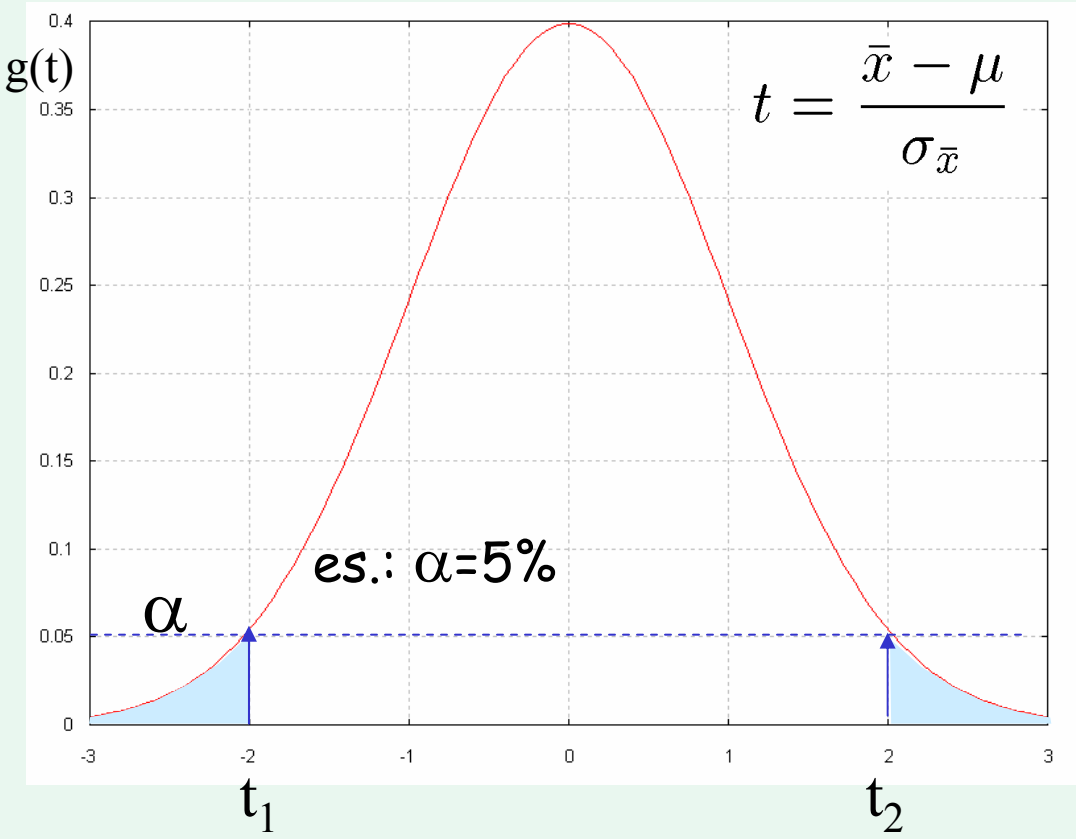
1) costruisco una variabile aleatoria con distribuzione nota, es.:

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad g(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

2) sulla base della g(t) determino i valori di t che hanno una bassa probabilità di essere osservati, cioè:

- fisso un livello di confidenza α.
- determino un intervallo di valori $t_{\alpha 1}$–$t_{\alpha 2}$ (intervallo di confidenza) tale che la probabilità di osservare t all'esterno dell'intervallo dato sia minore di α

$$\alpha = P(t < \alpha) = P(t < t_1) + P(t > t_2) = \int_{-\infty}^{t_1} g(t)dt + \int_{t_2}^{\infty} g(t)dt$$



Se $t_1 < t < t_2$ la probabilità di osservare il valore di t, calocolato in base ai dati, è (1-α)

Se $t < t_1$ o $t > t_2$ la probabilità di osservare il valore di t, calocolato in base ai dati, è α

$$P(t_1 < t < t_2)) = \int_{t_1}^{t_2} g(t)dt = 1 - \alpha$$

probabilità che t appartenga all'intervallo t_1 - t_2

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$P(t_1 < t < t_2) = P(\bar{x} - \sigma_{\bar{x}} t_1 < \mu < \bar{x} + \sigma_{\bar{x}} t_2)$$

dato il valore medio $\bar{x}$, osservato su un campione di **m** elementi, il valore aspettato della popolazione (μ) è contenuto nell'intervallo:

$$[\bar{x} - \sigma_{\bar{x}} t_1 ; \bar{x} + \sigma_{\bar{x}} t_1]$$

con probabilità 1-α.

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$P(t_1 < t < t_2) = P(\mu - \sigma_{\bar{x}} t_1 < \bar{x} < \mu + \sigma_{\bar{x}} t_2)$$

dato il valore medio $\bar{x}$, osservato su un campione di **m** elementi, 1-α è la probabilità che questo sia compreso nell'intervallo

$$[\mu - \sigma_{\bar{x}} t_1 ; \mu + \sigma_{\bar{x}} t_2]$$

che puó essere detto anche: α è la probabilità di fare un errore maggiore di $\sigma_{\bar{x}} t$ utilizzando la media come stima del valore atteso (quantificare il rischio)

funzione EXCEL: **CONFIDENZA(α, dev.st, m)**

Intervalli di confidenza: varianza nota

Se le osservazioni sono distribuite con:

valor medio $\bar{x}$

dev.st. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}}$

funzione EXCEL: **CONFIDENZA(α, dev.st, m)**

La resistenza elettrica di un cavo viene misurata con uno strumento che ha un'incertezza σ=0.5 Ω. Vengono effettuate 5 misure, ne risulta un valor medio $\bar{R}$=4.52 Ω

CONFIDENZA(0.05, 0.5, 5)=0.438

La resistenza vera del cavo è:

R = 4.52 ± 0.44 Ω oppure: R = [4.08, 4.96]

Nota: α rappresenta il rischio di sbagliare, cioè la probabilità che il valore vero della resistenza sia esterno all'intervallo dato

Intervalli di confidenza: varianza campionaria

Molto piú spesso non conosco la varianza della distribuzione. La migliore stima della varianza in un campione di m elementi è:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{m}} \qquad \text{da cui:} \qquad t = \frac{\bar{x} - \mu}{\sqrt{s^2/m}}$$

$$P(\mu - t_1 \frac{s}{\sqrt{m}} < \bar{x} < \mu + t_2 \frac{s}{\sqrt{m}})$$

probabilità che il valore osservato sia nell'intervallo $[\mu - \sigma_{\bar{x}} t_1 ; \mu + \sigma_{\bar{x}} t_2]$ intorno al valore vero.

$$\left[\bar{x} - t_1 \frac{s}{\sqrt{m}}; \bar{x} + t_1 \frac{s}{\sqrt{m}}\right]$$

funzione EXCEL: INV.T(α,m-1) = t_a

Nota: la variabile t cosi definita ha una distribuzione nota (t-Student) con ν = m-1 gradi di libertà. La t-Student approssima una distribuzione Gaussiana per ν che tende a infinito

Una misura dell'altezza di un gruppo di 20 studenti fornisce il valore medio: H = 1.68 m con la deviazione standard stimata s = 9 cm.

Determinare gli intervalli di confidenza con un incertezza minore di 1%, 0.5% e 0.05%

Dati	
numero di osservazioni: m	20
valor medio	1.68
dev. standard: s	0.09

Risultati						
H	1.62	1.74	1.61	1.75	1.59	1.77
α %	1		0.5		0.05	
t_a	2.86		3.17		4.19	
t*s/m^0.5	0.058		0.064		0.084	
confidenza	0.052		0.056		0.070	

Risultati				
H=	1.68+/-0.06	1.68+/-0.07	1.68+/-0.08	1.68+/-0.09
α %	1	0.5	0.1	0.05
t_a	2.86	3.17	3.88	4.19
t*s/m^0.5	0.058	0.064	0.078	0.084
confidenza	0.052	0.056	0.066	0.070

inv.T(α; m-1)

$$t_\alpha \frac{s}{\sqrt{m}}$$

CONCATENA(**TESTO**(valore,"0.00"),"testo",...)

Nota: Se conosco la varianza utilizzo la funzione "confidenza", se devo stimare la varianza utilizzo la funzione "inv.T"

Test di ipotesi, test statistici

1) ipotesi da verificare
ipotesi nulla: H_0

Es.: il valore misurato è compatibile con il valore vero?

2) costruisco una variabile aleatoria con distribuzione nota, es.:

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

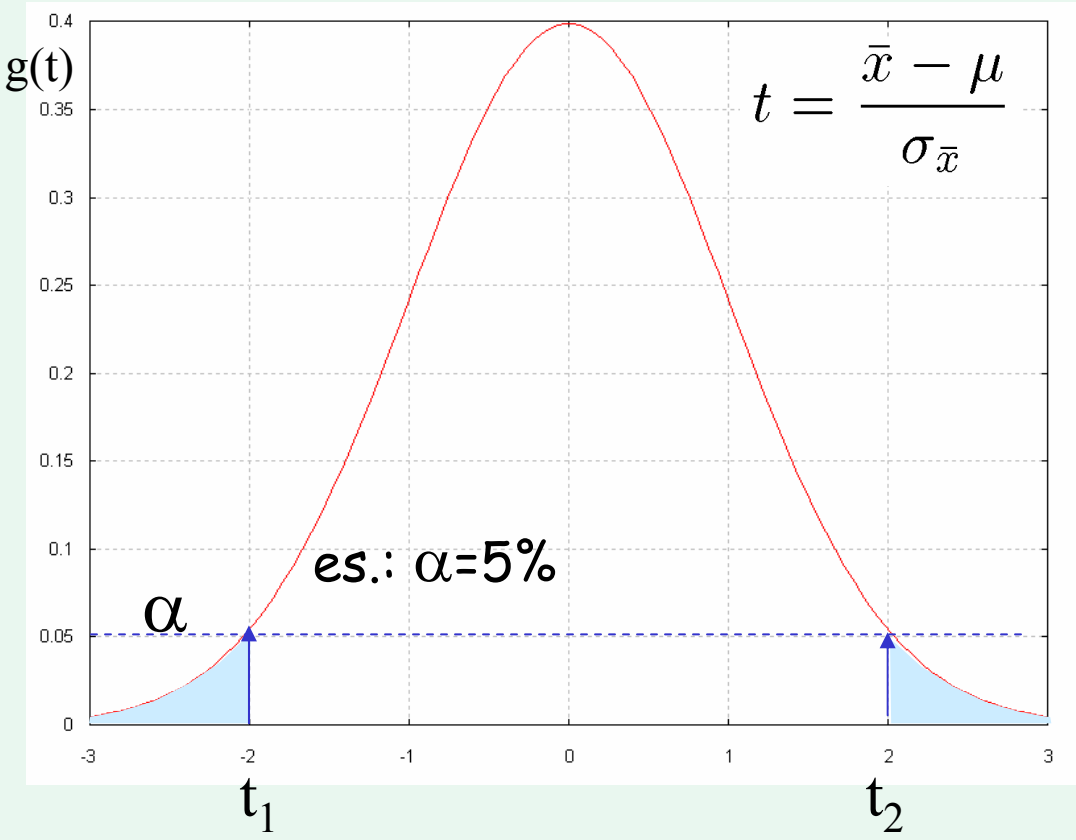
$$g(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

3) sulla base della g(t) determino i valori di t che hanno una bassa probabilità di essere osservati, fissando il livello di confidenza α. Se, in base alla distribuzione scelta, il valore osservato fornisce un valore di t con bassa probabilità di essere osservato, l'ipotesi deve essere rifiutata.

Quale è il rischio di scartare un dato compatibile?

$$P(t < \alpha) = P(t < t_1) + P(t < t_1) = \int_{-\infty}^{t_1} g(t)dt + \int_{t_2}^{\infty} g(t)dt = \alpha$$

$$P(t > \alpha) = P(t_1 < t < t_1) = \int_{t_1}^{t_2} g(t)dt = 1 - \alpha$$



Se $t_1 < t < t_2$ il risultato (cui è associato il valore t) è compatibile l'ipotesi fatta con una probabilità del P = (1-α)

Se $t < t_1$ o $t > t_2$ il risultato (cui è associato il valore t) non è compatibile l'ipotesi fatta con una probabilità del P = (1-α)

α rappresenta la probabilità di sbagliare e scartare un'ipotesi corretta.

F-test

Si hanno K esperimenti, ognuno effettuato su un campione di m_k osservazioni, $m=m_1+m_2+....m_K$ numero di osservazioni

Ipotesi: (H_o) i diversi campioni derivano tutti dalla stessa popolazione (stessa media e stessa varianza)

	Exp_1	Exp_2	. . .	Exp_K
	m_1	m_2		m_K
	$x_{1,1}$	$x_{2,1}$		$x_{K,1}$
	$x_{1,2}$	$x_{2,2}$		$x_{K,2}$
	$x_{1,3}$	$x_{2,3}$		$x_{K,3}$
	...	...		
	x_{1,m_1}	x_{2,m_2}		x_{K,m_2}
medie:	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_K$
dev.st:	s_1	s_2		s_K

Nota: La media delle medie campionarie è, in generale, diversa dalla media sull'intera popolazione

Media delle medie

$$\overline{X} = \frac{1}{K}\sum_{k=1}^{K} \bar{X}_k$$

Media sull'intera popolazione: media di medie pesate per la numerosità di ciascun campione

$$\bar{X} = \frac{\sum_{i=1}^{m_1} x_i + \sum_{i=1}^{m_2} x_i + \ldots + \sum_{i=1}^{m_K}}{m_1 + m_2 + \ldots + m_k} = \frac{\sum_{k=1}^{K} m_k \bar{X}_k}{\sum_{k=1}^{K} m_k} = \frac{\sum_{k=1}^{K} m_k \bar{X}_k}{m}$$

$$\bar{X} = \frac{1}{m}\sum_{k=1}^{K}\sum_{i=1}^{m_k} x_{k,i}$$

Media sull'intera popolazione: è una stima migliore della media.

Stima della varianza **A**

Se i gruppi appartengono alla stessa popolazione possiamo stimare la varianza in due modi

A: media pesata delle stime della varianza ottenute nei vari campionamenti (in):

$$s^2_{in} = \frac{\sum_{k=1}^{K}(m_k - 1)s^2_k}{\sum_{k=1}^{K}(m_k - 1)} = \frac{\sum_{k=1}^{K}(m_k - 1)s^2_k}{m - K}$$

media delle stime della varianza **in** ogni gruppo

$$s^2_{in} = \frac{1}{K}\sum_{k=1}^{K} s^2_k$$

se le numerosità dei campioni sono eguali: $m_i = m_o$

Stima della varianza B

B: assumiamo, nell'ipotesi che i campioni provengano tutti dalla stessa popolazione, che i valori medi ottenuti nei vari campionamenti siano tutti una stima dello stesso valore medio X della popolazione.

Possiamo stimare la la varianza della popolazione come media delle varianze stimata per ciascun campione:

$$s^2_{tra} = \frac{1}{K-1}\sum_{k=1}^{K} m_k \left(\bar{X}_k - \bar{X}\right)^2$$

Nota: per campioni della stessa numerosità:

$$\frac{s^2_{tra}}{m_o} = \frac{1}{K-1}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})^2 = s^2_{\bar{X}}$$

dove: $s_{\bar{X}}$ è l'errore standard sulla media

quindi:

$$s^2_{tra} = m_o s^2_{\bar{X}}$$

La variabile F

Stime della varianza della popolazione

varianza delle medie campionarie

$$s^2_{tra} = \frac{1}{K-1}\sum_{k=1}^{K} m_k \left(\bar{X}_k - \bar{X}\right)^2$$

media delle varianze campionarie

$$s^2_{intra} = \frac{\sum_{k=1}^{K}(m_k-1)s_k^2}{\sum_{k=1}^{K}(m_k-1)} = \frac{\sum_{k=1}^{K}(m_k-1)s_k^2}{m-K}$$

$$\mathrm{F} = \frac{\text{varianza della popolazione stimata dalle medie campionarie (tra)}}{\text{varianza della popolazione come media delle varianze campionarie (in)}}$$

$$F(\nu_1,\nu_2) = \frac{S^2_{tra}}{S^2_{intra}} \qquad \nu_1 = K-1\,, \quad \nu_2 = m-K$$

La variabile F

$$F = \frac{\text{varianza della popolazione stimata dalle medie campionarie (tra)}}{\text{varianza della popolazione come media delle varianze campionarie (in)}}$$

$$F(\nu_1, \nu_2) = \frac{s^2_{tra}}{s^2_{in}}$$

$s_{in} < s_{tra}$

S_{in} e S_{tra} sono due stime della stessa grandezza quindi mi aspetto che il loro rapporto sia F~1

Valori di F~1 indicano che le deviazioni standard dei campioni sono simili, quindi i campioni provengono dalla stessa popolazione.

Valori di F>>1 indicano che le deviazioni standard dei campioni sono diverse, quindi i campioni provengono da popolazioni divere.

Quanto sono significative queste differenze?

Gradi di libertà

$F(\nu_n, \nu_d)$

K = numero di campioni

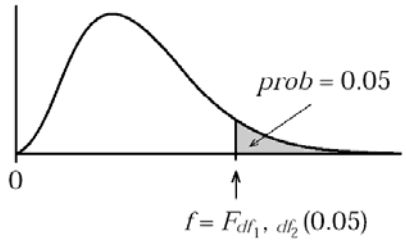
ν_n: gradi di lib. numeratore: K-1

ν_d : gradi di lib. denominatore

$\nu_d = m_1 + m_2 + \ldots m_K - K = m - K$

Appendix 2 ***F*-distribution, 5% Table**

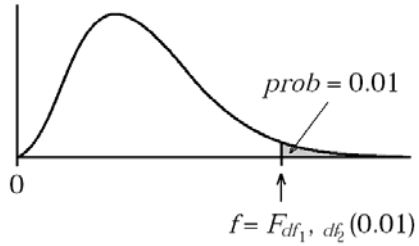
For fixed df_1, df_2 the tabulated value is the number $f = F_{df_1,df_2}(0.05)$ such that for $F \sim F(df_1, df_2)$, $\text{pr}(F \geq f) = 0.05$.



ν_d	ν_n 1	2	3	4	5	6	7	8	9	10	12	15	20	30	60
1	161	199	215	224	230	233	236	238	240	241	243	245	248	250	252
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.01
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.62

Appendix 3 ***F*-distribution, 1% Table**

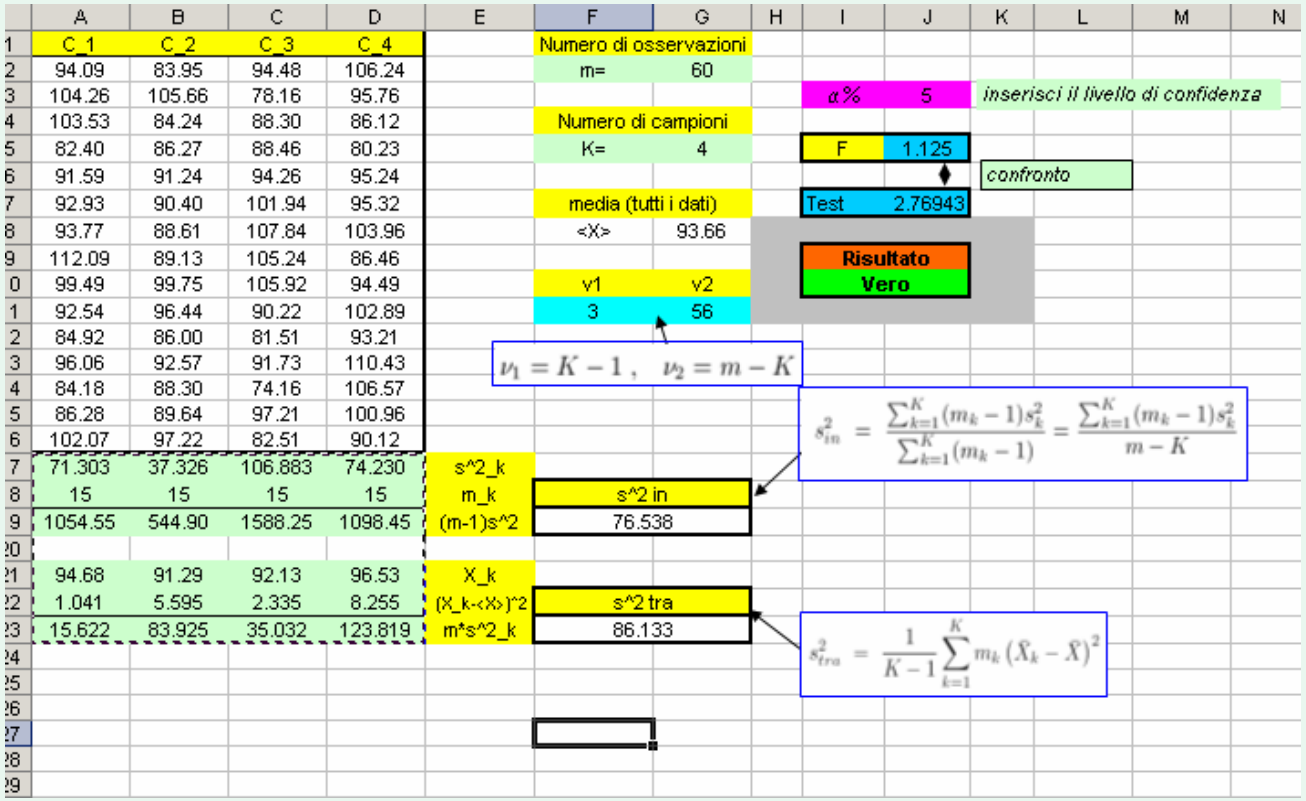
For fixed df_1, df_2 the tabulated value is the number $f = F_{df_1,df_2}(0.01)$ such that for $F \sim F(df_1, df_2)$, $\text{pr}(F \geq f) = 0.01$.



ν_d	ν_n 1	2	3	4	5	6	7	8	9	10	12	15	20	30
1	4052	4999	5403	5624	5763	5858	5928	5981	6022	6055	6106	6157	6208	6260
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.5
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.38
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25

Funzione EXCEL: INV.F (α, ν_1, ν_2)

Il livello di confidenza α indica il rischio (probabilità) di sbagliare affermando che i campioni sono diversi quando in effetti derivano dalla stessa popolazione.



Confronto fra due popolazioni
t-test

Problema: si vogliono confrontare se due popolazioni normali, X_1 e X_2. Supponiamo per ora che queste abbiano la stessa varianza (omeoschedasticità).

Ipotesi: (H_o) le due popolazioni hanno la stessa media

$$t = \frac{\text{differenza delle medie campionarie}}{\text{errore standard sulla differenza delle medie campionarie}}$$

$$t_\nu = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}$$

$$t_\nu = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}}$$

$$\sigma^2_{\bar{x}_i} = \frac{s_i^2}{m_i}$$

$$\nu = m_1 + m_2 - 2$$

La variabile aleatoria t_ν segue una distribuzione nota (t-student) con ν gradi di libertà

$$t_\nu = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}} \qquad \sigma^2_{\bar{x}_i} = \frac{s_i^2}{m_i}$$

Se m_1 è diverso da m_2

$$t_\nu = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2/m_1 + s^2/m_2}} \qquad s^2 = \frac{(m_1 - 1)s_1^2 + (m_2 - 1)s_2^2}{m_1 + m_2 - 2}$$

Se $m_1 = m_2 = m$

$$t_\nu = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{2s^2/m}} \qquad s^2 = \frac{(s_1^2 + s_2^2)}{2}$$

$$\nu = 2(m - 1)$$

Confronto fra due popolazioni: t-test

- La distribuzione della variabile *t*, ha una forma nota come "*student's t distribution*" (tende alla distribuzione normale di Gauss per N→∞)
- La forma della distribuzione dipende da un solo parametro, legato alla numerosità del campione: il numero di gradi di libertà

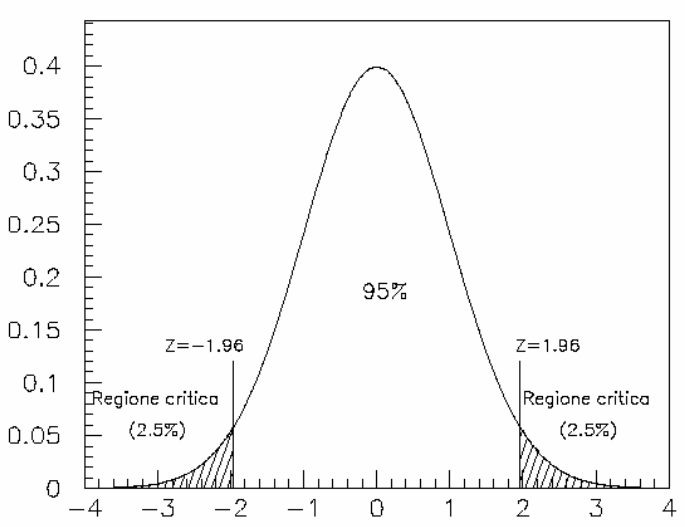
$$\nu = m_1 + m_2 - 2$$

- Valori "piccoli" di *t* indicano che la differenza fra le medie dei due campioni non è significativa (i campioni sono consistenti), valori "grandi" indicano una differenza significativa
- Per formalizzare il concetto, si considera la probabilità che il valore di *t* sia maggiore (in valore assoluto) di un dato limite.

  I valori di cui $|t|$ è maggiore con una data probabilità sono detti "valori critici" e si trovano tabulati (→)

t-test

Fissato il numero di gradi di libertà, la tabella indica i valori di t tali per cui la probabilità di ottenere un valore maggiore (in modulo) di quello indicato sia pari ad α



α: Two Tails:	0.500	0.200	0.100	0.050	0.020	0.010
df:						
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
					2.567	2.898
					2.552	2.878
					2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845

la probabilità che t_{10} >2.228 è <5%

funzione EXCEL: **TEST.T(X_A, X_B , coda, tipo)**

X_A X_A insiemi (matrici) dei dati corrispondenti ai due campionamenti

coda: **1** - test a una coda
2 – test a due code

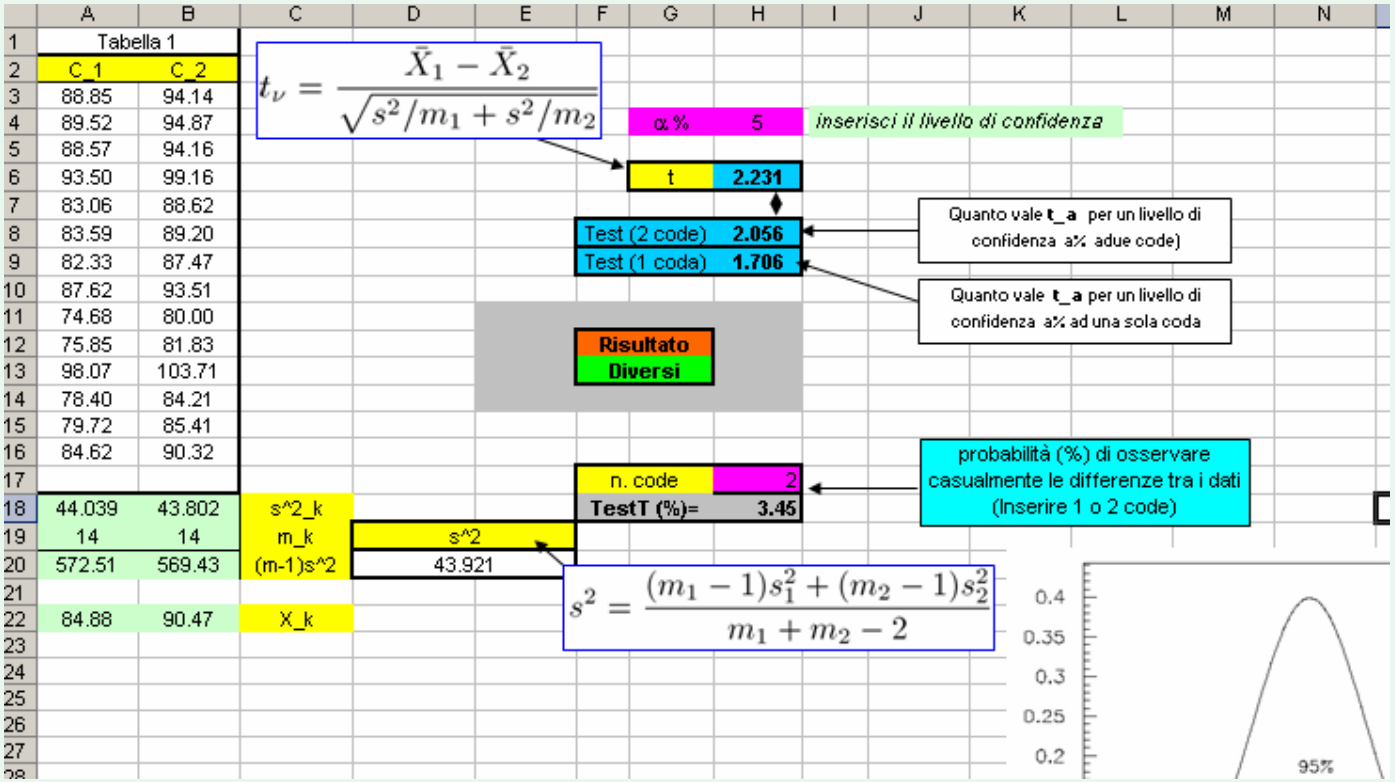
tipo: **1** test accoppiato (stesso numero di valori)
2 test omeoschedastico (stessa varianza)
3 test eteroschedastico (varianza diversa)

Il risultato rappresenta l'indice di confidenza del test. Ad esempio, un valore

TEST.T(....)=0.02

indica che, in base ai dati, la probabilità di sbagliare dicendo che le due medie sono diverse è il 2%.

Non posso dire che al 98% sono eguali! E' sbagliato dire che le medie sono diverse con il 2% di probabilità



In Excel la funzione Test.T restituisce la probabilità di osservare casualmente la differenza riscontrata.

Si distingue il caso in cui non si conosce il segno della differenza (test a due code) da quello in cui si conosce il segno della differenza (una coda)