Facoltà di Ingegneria
Università degli Studi di Firenze
Dipartimento di Elettronica e Telecomunicazioni

Misure Elettriche

- Inferenza statistica -

2. Teoria dei Campioni

Ing. Andrea Zanobini
Dipartimento di Elettronica e Telecomunicazioni

Popolazioni

Per **popolazione** si intende oggi un insieme o collezione di oggetti, numeri, misure o osservazioni. La popolazione può essere **finita** o **infinita**; ad esempio la popolazione costituita da tutti i bulloni prodotti in una fabbrica in un dato giorno è finita; la popolazione costituita da tutte le possibili uscite T o C in successivi lanci di una moneta è infinita.

Se la popolazione è infinita, è impossibile osservarne tutti i valori, ma anche quando è finita, questo può essere non pratico o antieconomico.
Le ragioni per cui la ricerca viene effettuata per campione, piuttosto che attraverso una rilevazione totale, sono principalmente le seguenti:

1 – risorse limitate: ad esempio nelle rilevazioni pre-elettorali non sono disponibili i fondi per osservare un'intera popolazione;

2 – pochi dati disponibili: qualche volta è disponibile solo un piccolo campione, e non per motivi economici. Si pensi ad esempio ad un antropologo che vuole provare una certa teoria riguardante una popolazione oggi quasi estinta ed ha a disposizione solo gli ultimi sopravvissuti, 1000 persone che vivono in una certa isola: la dimensione del campione è fissata dalla natura e non dalle risorse finanziarie;

3 – impossibilità a compiere certi test: il campionamento può essere l'unica soluzione in quei casi in cui il test distrugge l'oggetto in esame. Ad esempio, data una produzione di lampadine, si vuole conoscere la loro vita media: è illogico insistere nell'osservare l'intera popolazione di lampadine finché siano tutte bruciate.

Campioni

Si usa perciò un **campione**, una parte della popolazione, e si traggono da esso, ossia si inferiscono, risultati riguardanti l'intera popolazione. La **teoria dei campioni** è lo studio delle relazioni esistenti tra una popolazione ed i campioni estratti da essa.

Tale teoria è utile per ottenere la **stima dei parametri** ignoti di una popolazione, media, varianza, scarto quadratico medio, quando si conoscono i valori corrispondenti del campione.

E' utile anche per stabilire se le differenze osservate tra due campioni possono essere dovute al caso o se sono significative: le risposte a questo tipo di quesito implicano l'uso dei **test di ipotesi**.

Affinché le conclusioni della teoria dei campioni siano valide, i campioni devono essere scelti in modo da essere rappresentativi della popolazione.

Campionamento casuale

Il miglior modo per assicurarsi un campione non distorto consiste nel fornire a ciascun membro della popolazione un'eguale possibilità di essere incluso nel campione: questa è la definizione di **campione casuale**.

Il **campionamento** è detto **con reimmissione** se ogni elemento può essere scelto più di una volta: ad esempio nell'estrazione con reimmissione da un'urna l'elemento viene estratto la prima volta, poi rimesso nell'urna e può essere estratto una seconda volta, e così via.

Se invece ciascun elemento può essere scelto una sola volta, cioè non viene reimmesso dopo la prima estrazione, si parla di **campionamento senza reimmissione**.

Lo scopo, come già detto, è quello di generalizzare le informazioni dal campione alla popolazione, soprattutto stimare i **parametri**, come la media μ o la varianza σ^2 della popolazione.

Per far questo ci serviamo delle quantità calcolate dal campione, media $\bar{x}$ e varianza s^2 ; queste quantità vengono chiamate anche **statistiche**.

Distribuzioni di Campionamento

Consideriamo tutti i possibili campioni casuali di ampiezza n che possono essere estratti da una data popolazione, con o senza reimmissione. Per ciascun campione si può calcolare una data statistica, come la media, la varianza o lo scarto quadratico medio, che potrà variare da campione a campione. In tal modo otteniamo una distribuzione della statistica, detta distribuzione di campionamento della statistica stessa.

Definizione
Si definisce **distribuzione di campionamento** di una data statistica la distribuzione di tutti i possibili valori che possono essere assunti dalla statistica stessa, calcolati da campioni casuali della stessa dimensione estratti dalla stessa popolazione.

1 – da una popolazione finita di dimensione N si estraggono tutti i possibili campioni casuali di ampiezza n;

2 – si calcola la statistica di interesse per ogni campione;

3– si costruisce una tabella contenente i vari valori distinti assunti dalla statistica

Le caratteristiche importanti di una distribuzione di campionamento, a cui siamo interessati, sono la sua media, la sua varianza e la sua forma.

Distribuzione della media campionaria (varianza σ^2 nota)

Si estrae un primo campione casuale di n elementi da una data popolazione, e si indica con $\overline{x}_1$ la sua media; se si estrae un secondo campione di n elementi dalla stessa popolazione, si ottiene un altro valore per la media $\overline{x}_2$, di solito diverso dal precedente; se si estraggono successivamente altri campioni, i valori delle medie saranno in generale diversi fra loro.

I valori delle medie possono essere visti come i valori assunti da una variabile aleatoria $\overline{X}$, detta **media campionaria**, su tutti i possibili campioni di ampiezza n che possono essere estratti dalla popolazione. La differenza fra i valori delle medie è dovuta al caso, e questo fatto suggerisce di studiare la distribuzione di tali valori.

Esempio 1

Si consideri una popolazione finita, costituita da $N = 4$ elementi, e avente la seguente distribuzione uniforme discreta

x_i	1	2	3	4
$f(x_i)$	0.25	0.25	0.25	0.25

Tabella 1

Esempio 1
Si consideri una popolazione finita, costituita da $N = 4$ elementi, e avente la seguente distribuzione uniforme discreta

x_i	1	2	3	4
$f(x_i)$	0.25	0.25	0.25	0.25

Tabella 1

La media μ e la varianza σ^2 di questa popolazione sono

$$\mu = \frac{1+2+3+4}{4} = 2.5$$

$$\sigma^2 = 1\cdot\frac{1}{4} + 4\cdot\frac{1}{4} + 9\cdot\frac{1}{4} + 16\cdot\frac{1}{4} - (2.5)^2 = 1.25$$

Con reimbussolamento

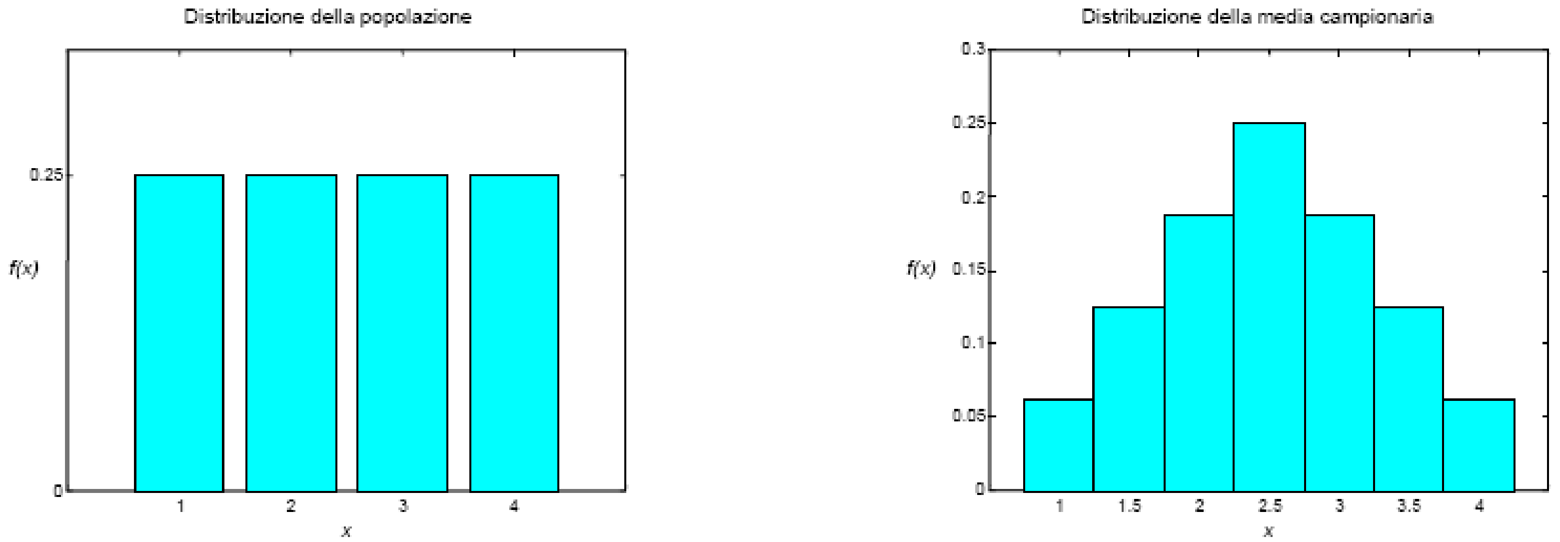
Consideriamo tutti i possibili campioni di dimensione $n = 2$ estraibili da questa popolazione; quando il campionamento avviene con reimmissione, i campioni di ampiezza 2 sono in numero di $4^2 = 16$; tali campioni sono elencati nella tabella 2, insieme con le corrispondenti medie.

Campioni	*Medie*	*Campioni*	*Medie*
(1,1)	1	(3,1)	2
(1,2)	1.5	(3,2)	2.5
(1,3)	2	(3,3)	3
(1,4)	2.5	(3,4)	3.5
(2,1)	1.5	(4,1)	2.5
(2,2)	2	(4,2)	3
(2,3)	2.5	(4,3)	3.5
(2,4)	3	(4,4)	4

Nella tabella 3 è riportata la distribuzione della media campionaria, ottenuta elencando i diversi valori della media campionaria nella prima riga e le rispettive frequenze nella seconda riga.

$\overline{x}_i$	1	1.5	2	2.5	3	3,5	4
$f(\overline{x}_i)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Tabella 3



Calcoliamo la media della distribuzione della media campionaria

$$\mu_{\overline{X}} = 1 \cdot \frac{1}{16} + 1.5 \cdot \frac{2}{16} + 2 \cdot \frac{3}{16} + 2.5 \cdot \frac{4}{16} + 3 \cdot \frac{3}{16} + 3.5 \cdot \frac{2}{16} + 4 \cdot \frac{1}{16} = 2.5$$

Questa media è uguale alla media della popolazione.

Calcoliamo infine la varianza della distribuzione della media campionaria

$$\sigma_{\overline{X}}^2 = 1 \cdot \frac{1}{16} + (1.5)^2 \cdot \frac{2}{16} + (2)^2 \cdot \frac{3}{16} + (2.5)^2 \cdot \frac{4}{16} + (3)^2 \cdot \frac{3}{16} + (3.5)^2 \cdot \frac{2}{16} + (4)^2 \cdot \frac{1}{16} - (2.5)^2 = 0.625$$

Questa varianza non è uguale alla varianza della popolazione, tuttavia si osserva che vale la relazione

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n} = \frac{1.25}{2} = 0.625$$

Senza reimbussolamento

Campioni	*Medie*
(1,2)	1.5
(1,3)	2
(1,4)	2.5
(2,3)	2.5
(2,4)	3
(3,4)	3.5

Tabella 4

$\overline{x}_i$	1.5	2	2.5	3	3,5
$f(\overline{x}_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tabella 5

Campioni	*Medie*
(1,2)	1.5
(1,3)	2
(1,4)	2.5
(2,3)	2.5
(2,4)	3
(3,4)	3.5

Tabella 4

$\overline{x}_i$	1.5	2	2.5	3	3,5
$f(\overline{x}_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tabella 5

$$\mu_{\overline{X}} = 1.5 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 2.5 \cdot \frac{2}{6} + 3 \cdot \frac{1}{6} + 3.5 \cdot \frac{1}{6} = 2.5$$

$$\sigma^2_{\overline{X}} = (1.5)^2 \cdot \frac{1}{6} + (2)^2 \cdot \frac{1}{6} + (2.5)^2 \cdot \frac{2}{6} + (3)^2 \cdot \frac{1}{6} + (3.5)^2 \cdot \frac{1}{6} - (2.5)^2 = \frac{5}{12}$$

Osserviamo che la media della distribuzione della media campionaria è ancora uguale alla media della popolazione, mentre per la varianza si può verificare che vale la relazione

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{1.25}{2} \cdot \frac{4-2}{4-1} = \frac{1.25}{3} = \frac{5}{12}$$

Teorema 1

Se si estraggono campioni casuali di ampiezza n da una popolazione avente media μ e varianza σ^2, allora la distribuzione della media campionaria $\overline{X}$ ha media

$$\mu_{\overline{X}} = \mu \, . \tag{6.1}$$

Per campioni estratti da popolazioni infinite, o se il campionamento è fatto con reimmissione, la varianza della distribuzione della media campionaria è

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \, . \tag{6.2}$$

Per campioni estratti senza reimmissione da una popolazione finita di ampiezza N la varianza della distribuzione della media campionaria è

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \, . \tag{6.3}$$

Lo scarto quadratico medio $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ è detto **errore standard della media**

proporzione alla radice quadrata di n: per esempio è necessario quadruplicare l'ampiezza del campione per dimezzare l'errore standard della distribuzione della media campionaria.

Il fattore $\dfrac{N-n}{N-1}$, detto **fattore correttivo per la popolazione finita**, ha un valore prossimo a 1

Teorema del limite centrale

Teorema 2 – Teorema del limite centrale
Sia data una popolazione avente media μ e varianza σ^2, e da essa si estraggano campioni casuali di ampiezza n; indicando con $\overline{X}$ la media campionaria, la variabile

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{6.4}$$

è una variabile aleatoria la cui distribuzione tende alla distribuzione normale standardizzata per $n \to \infty$.

Qualunque sia la distribuzione della popolazione, si può quindi affermare che la distribuzione della media campionaria $\overline{X}$ è approssimativamente normale con media μ e varianza $\frac{\sigma^2}{n}$, per n sufficientemente grande.

In pratica nella maggior parte dei casi la distribuzione normale è una buona approssimazione della distribuzione della media campionaria per $n \geq 30$, qualunque sia la distribuzione della popolazione.

...riassumendo

Schema riassuntivo – Proprietà della distribuzione della media campionaria

1. Campionamento da una popolazione distribuita normalmente con media μ e varianza σ^2:

a – $\mu_{\overline{X}} = \mu$

b – $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$

c – la distribuzione della media campionaria $\overline{X}$ è normale.

2. Campionamento da una popolazione non distribuita normalmente con media μ e varianza σ^2:

a – $\mu_{\overline{X}} = \mu$

b – $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$ se $\dfrac{n}{N} \leq 0.05$

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

c – la distribuzione della media campionaria è approssimativamente normale, per $n \geq 30$.

Esempio 2
La variabile aleatoria continua X ha media $\mu = 5$ e varianza $\sigma^2 = 25$. Si estrae un campione di 100 elementi da questa popolazione; determinare la probabilità che la media del campione sia maggiore di 5.4.

In base al teorema 1, la media campionaria $\overline{X}$ ha il valor medio e la varianza seguenti

$$\mu_{\overline{X}} = \mu = 5 \qquad\qquad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{25}{100} = \frac{1}{4}.$$

Applicando il teorema del limite centrale, si può affermare che la variabile $\overline{X}$ ha approssimativamente la distribuzione normale.

Per calcolare la probabilità che la media del campione sia maggiore di 5.4, occorre standardizzare la media campionaria con la formula

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 5}{\sqrt{\frac{1}{4}}}$$

$$\overline{X} = 5.4 \quad \Rightarrow \quad Z = \frac{5.4 - 5}{0.5} = 0.8$$

$$P(\overline{X} > 5.4) = P(Z > 0.8) = 1 - P(Z < 0.8) = 1 - 0.7881 = 0.2119$$

Fare da soli gli esempi 3 e 4 !!

Distribuzione della media campionaria (varianza σ^2 incognita)

Nel caso che il numero n degli elementi del campione sia grande (**grande campione**), se σ^2 non è nota, si sostituisce a σ^2 la varianza s^2 del campione.
Se invece l'ampiezza n del campione è piccola (**piccolo campione**), si hanno dei risultati solo se il campione proviene da una popolazione normale.

Teorema 3
Sia data una popolazione normale avente media μ e da essa si estraggano campioni casuali di ampiezza n; indicando con $\overline{X}$ la media campionaria e con S lo scarto quadratico medio campionario, la variabile

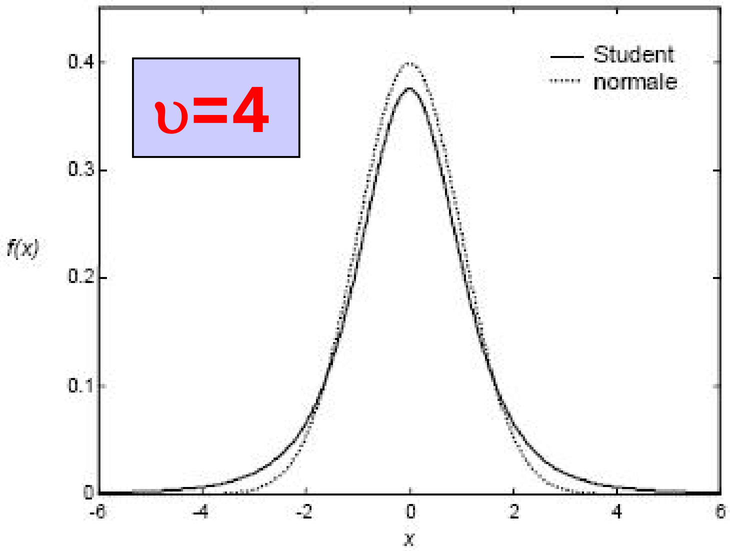
$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \qquad (6.5)$$

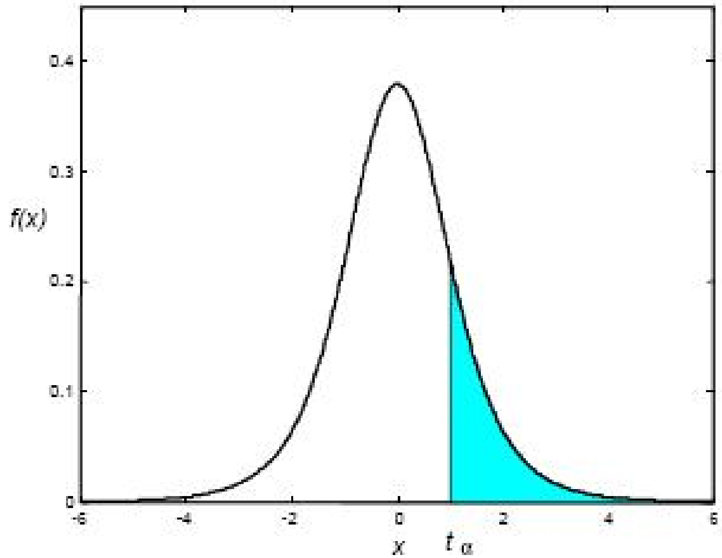
è una variabile aleatoria avente la **distribuzione t di Student**[1] con grado di libertà $\nu = n - 1$.

Questo teorema da un lato è più generale del teorema del limite centrale, nel senso che non richiede la conoscenza di σ, ma d'altra parte richiede l'ipotesi più restrittiva di una popolazione normale.

Distribuzione t di Student

La distribuzione t di Student non è un'unica distribuzione, ma una famiglia di distribuzioni dipendenti dal parametro ν, detto **grado di libertà**.





Come la distribuzione normale, la distribuzione t ha media $\mu = 0$; la sua varianza dipende dal grado di libertà ν; la varianza è maggiore di 1, e tende a 1 al crescere del grado di libertà.

I valori di t_α per $\nu > 29$ sono circa uguali ai corrispondenti valori tratti dalle tavole della distribuzione normale (vedere esempi 9 e 10): infatti la distribuzione normale è una buona approssimazione della distribuzione t per valori del grado di libertà $\nu > 29$.

Esempio 5

Data la distribuzione t con grado di libertà $\nu = 9$, trovare il valore di t_α tale che l'area a destra di t_α vale $\alpha = 0.05$ (figura 5).

Dalle tavole si deduce che $t_\alpha = 1.833$

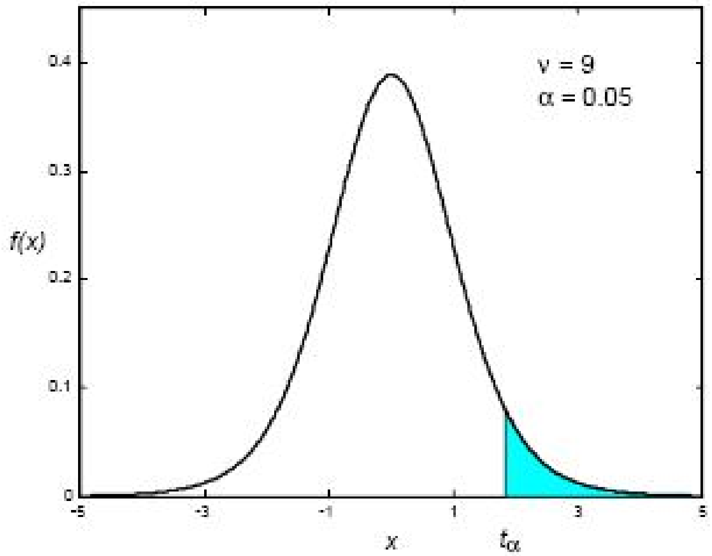


Figura 5

Fare da soli l'esempio 6 !!

Esempio 7

Data la distribuzione t con grado di libertà $\nu = 10$, trovare il valore di t_α tale che l'area compresa fra $-t_\alpha$ e t_α vale $\alpha = 0.90$ (figura 7)

Area compresa fra $-t_\alpha$ e $t_\alpha = \alpha = 0.90 \Rightarrow$ Area totale delle due code $= 1 - \alpha = 0.1 \Rightarrow$

Area di una coda $= \dfrac{1-\alpha}{2} = 0.05$

Dalle tavole si deduce $t_\alpha = t_{0.05} = 1.812$

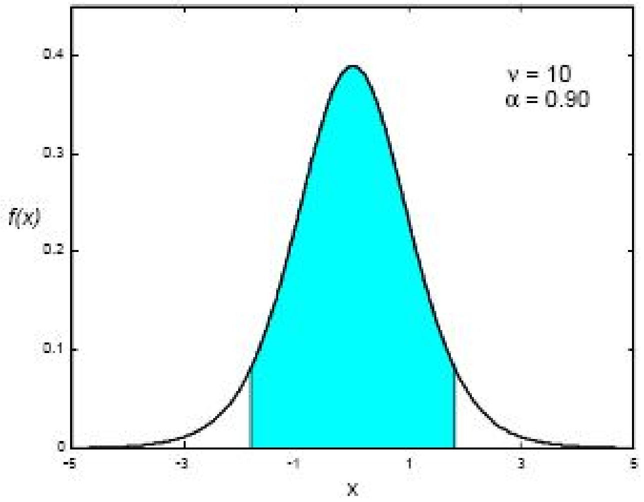


Figura 7

Fare da soli l'esempio 8 !!

Esempio 9

Data la distribuzione t con grado di libertà $\nu > 29$, trovare il valore di t_α tale che l'area a destra di t_α vale $\alpha = 0.025$.

Verificare che si ottiene lo stesso valore con la tavola della distribuzione normale.

Dalla tavola della distribuzione t si ottiene

$$t_\alpha = 1.960$$

Dalla tavola dei quantili della distribuzione normale standardizzata si ottiene lo stesso valore

$$z_\alpha = 1.960$$

Esempio 10

Data la distribuzione t, trovare i valori di t_α tale che l'area a destra di t_α vale $\alpha = 0.05$ per i gradi di libertà $\nu = 16$, $\nu = 27$, $\nu = 200$.

Dalle tavole si trova

a –	$\nu = 16$	$t_\alpha = 1.746$
b –	$\nu = 27$	$t_\alpha = 1.703$
c –	$\nu = 200$	$t_\alpha = 1.645$

Quest'ultimo valore è uguale al valore che si trova dalla tavola dei quantili della distribuzione normale standardizzata

$$z_\alpha = 1.645.$$

Distribuzione della varianza campionaria

Studiamo la **distribuzione di campionamento della varianza campionaria** per campioni provenienti da una popolazione normale; otteniamo questa distribuzione estraendo tutti i possibili campioni casuali di ampiezza n da una popolazione avente distribuzione normale e determinando per ciascuno di essi la varianza campionaria s^2. Poiché s^2 non può essere negativa, ci si attende che la distribuzione della varianza campionaria non sia simmetrica, cioè non sia di tipo normale.

Teorema 4
Sia data una popolazione normale avente varianza σ^2 e da essa si estraggano campioni casuali di ampiezza n; indicando con S^2 la varianza campionaria, la variabile

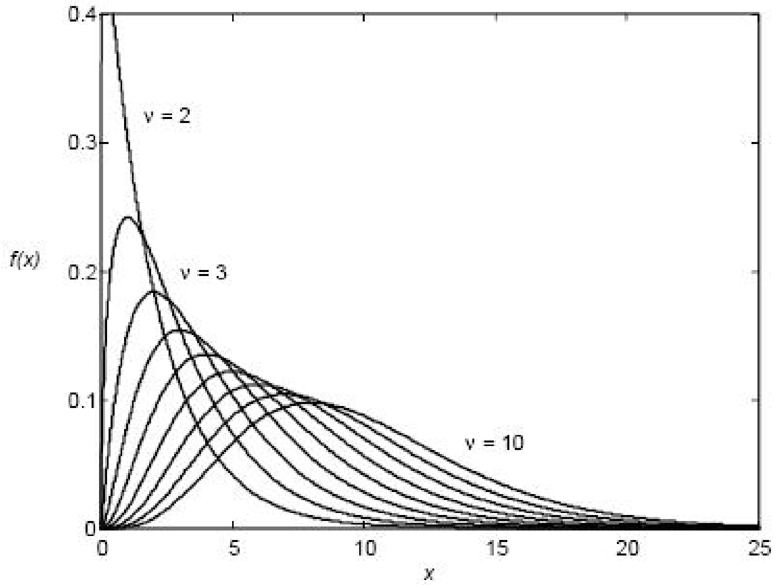
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \tag{6.6}$$

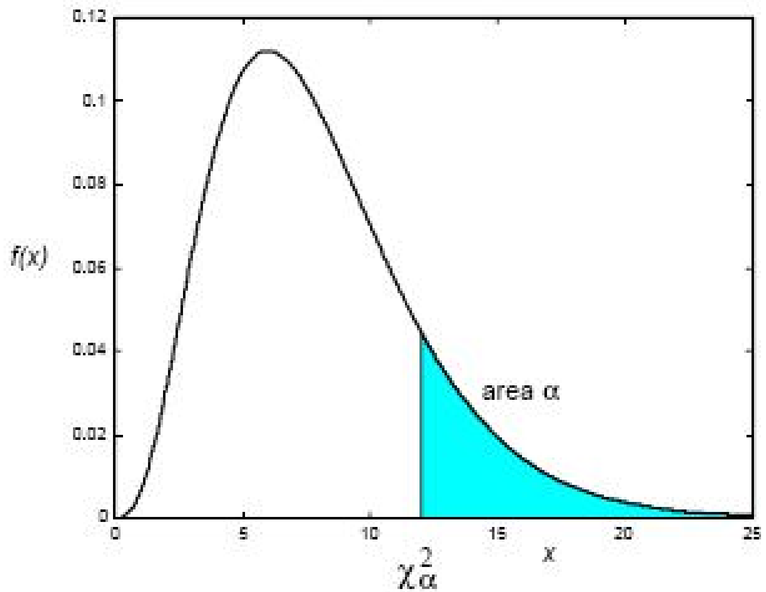
è una variabile aleatoria avente la **distribuzione** χ^2 (**chi quadro**) di parametro $\nu = n-1$.

Si dimostra che la distribuzione χ^2 ha media $\mu = \nu$ e varianza $\sigma^2 = 2\nu$.

Distribuzione chi quadro

Il parametro ν è detto **grado di libertà**. Anche la distribuzione chi quadro non è un'unica distribuzione, ma una famiglia di distribuzioni dipendenti dal grado di libertà ν.



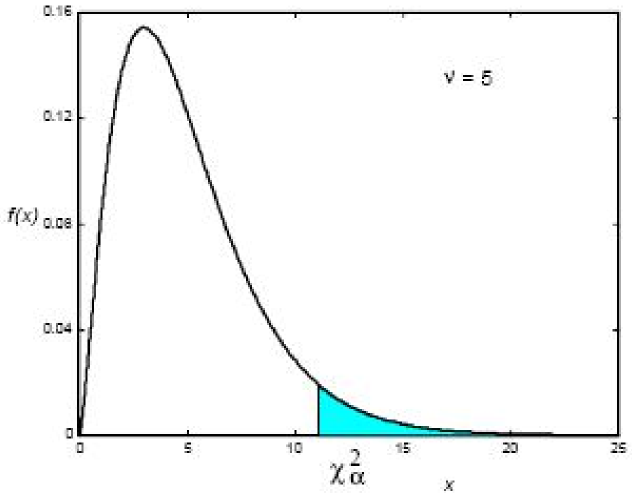


La distribuzione chi quadro è definita solo per valori positivi di x e in generale è asimmetrica; l'asimmetria diminuisce per valori elevati di ν.

Esempio 11

Data la distribuzione χ^2 con grado di libertà $\nu = 5$, trovare il valore di χ^2_α tale che l'area a destra di χ^2_α vale $\alpha = 0.05$ (figura 11).

Dalle tavole, per $\nu = 5$ e $\alpha = 0.05$ si deduce $\chi^2_\alpha = 11.070$.



Fare da soli l'esempio 12 !!

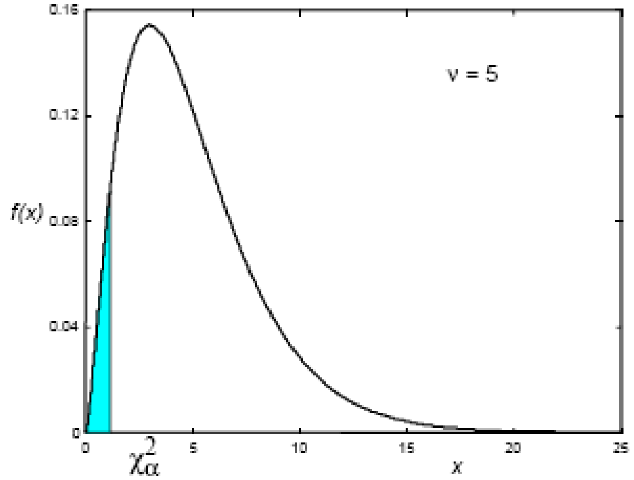
Esempio 13

Data la distribuzione χ^2 con grado di libertà $\nu = 5$, trovare il valore di χ^2_α tale che l'area a sinistra di χ^2_α vale $\alpha = 0.05$ (figura 12, pag. seguente).

Area a sinistra di $\chi^2_\alpha = \alpha = 0.05 \quad \Rightarrow$ Area a destra di $\chi^2_\alpha = 1 - \alpha = 0.95$

Dalle tavole si deduce

$$\chi^2_\alpha = 1.145 \; .$$



Fare da soli l'esempio 14 !!

Distribuzione del rapporto di due varianze di due campioni indipendenti

Per studiare il rapporto di due varianze si utilizza la distribuzione di campionamento della variabile

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

Teorema 5

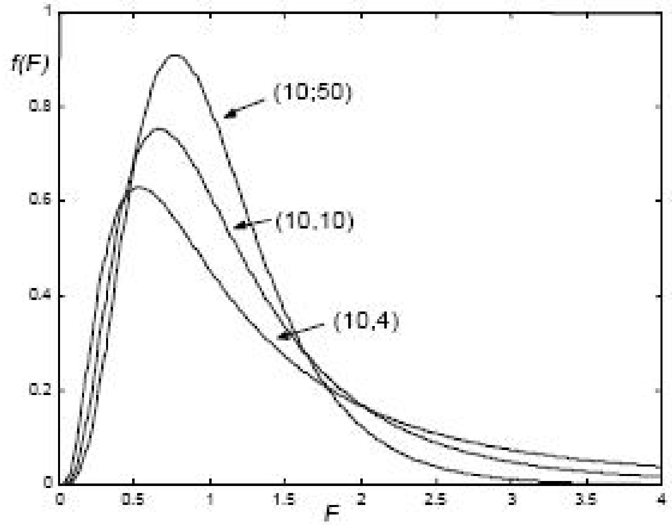
Siano date due popolazioni normali aventi varianze σ_1^2 e σ_2^2, e si estraggano da esse campioni casuali indipendenti di ampiezza rispettivamente n_1 e n_2; indicando con S_1^2 e S_2^2 le varianze campionarie, la variabile

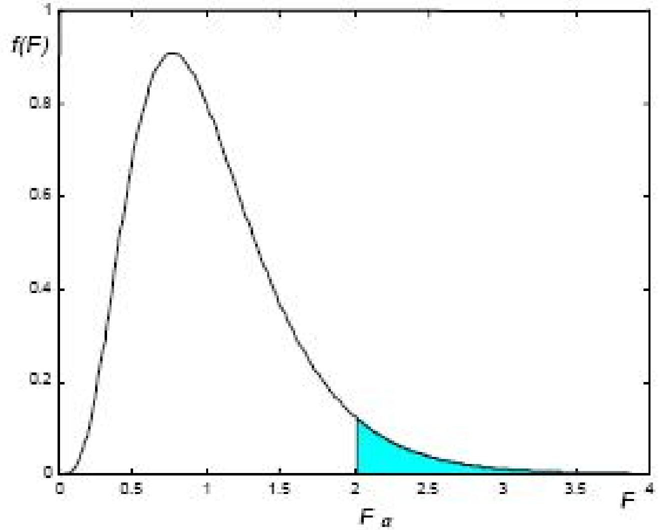
$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{6.7}$$

è una variabile aleatoria avente la **distribuzione *F***, detta anche **distribuzione di Fisher**, di parametri $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$.

Distribuzione F di Fisher

La distribuzione F dipende dai due parametri ν_1 e ν_2, detti **gradi di libertà del numeratore e del denominatore**.





$$F_{1-\alpha}(\nu_1,\nu_2)=\frac{1}{F_\alpha(\nu_2,\nu_1)} \tag{6.8}$$

Esempio 15

Data la distribuzione F con gradi di libertà $\nu_1 = 15$, $\nu_2 = 25$, trovare il valore F_α tale che l'area a destra di F_α vale

a – $\alpha = 0.10$;
b – $\alpha = 0.05$;
c – $\alpha = 0.01$.

Dalle tavole, per i gradi di libertà $\nu_1 = 15$, $\nu_2 = 25$, si deduce

a – $F_{0.10}(15{,}25) = 1.77$

b – $F_{0.05}(15{,}25) = 2.09$

c – $F_{0.01}(15{,}25) = 2.85$

Fare da soli l'esempio 16 !!

Esempio 17

Data la distribuzione F con gradi di libertà $\nu_1 = 10$, $\nu_2 = 15$, trovare i valori $F_{1-\frac{\alpha}{2}}$ e $F_{\frac{\alpha}{2}}$ tali che l'area compresa fra essi vale $\alpha = 0.90$.

Dato che la distribuzione F non è simmetrica, di solito si scelgono le due code in modo che abbiano uguale area; in questo esempio entrambe hanno area uguale a 0.05.

Dalle tavole, per $\frac{\alpha}{2} = 0.05$, $\nu_1 = 15$, $\nu_2 = 10$ si desume che

$$F_{0.05}(15{,}10) = 2.85$$

Con la (6.8), per $1 - \frac{\alpha}{2} = 0.95$ si ha

$$F_{0.95}(15{,}10) = \frac{1}{F_{0.05}(10{,}15)} = \frac{1}{2.54} = 0.394\,.$$

Si ricordi che i teoremi 4 e 5 richiedono l'ipotesi che i campioni vengano estratti da una popolazione normale. Contrariamente a quanto accade con la distribuzione t (teorema 3), scostamenti anche modesti dalla distribuzione normale possono avere conseguenze serie sulle distribuzioni campionarie.