TEOREMA DEL LIMITE CENTRALE

DISTRIBUZIONE DELLE STATISTICHE CAMPIONARIE

Prof. Rosario Lo Franco – Lezione 8

Riferimenti: [1] Sheldon M. Ross, *Introduzione alla statistica*, Apogeo Editore; [2] Maria Garetto, *Statistica*, Università di Torino

Campionamento

Uno degli aspetti principali della **statistica inferenziale** consiste nel trarre delle conclusioni sui parametri di una popolazione utilizzando i corrispondenti valori campionari.
La necessità di ricorrere ai metodi della statistica inferenziale deriva dalla necessità del campionamento: 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 – l'estrazione di un campione richiede meno tempo rispetto all'esame dell'intera popolazione;

2 – un campione è meno costoso;

3 – un campione è più pratico da gestire;

4 – a volte l'esame dell'intera popolazione è impossibile: ad esempio è letale estrarre tutto il sangue di un paziente per effettuare il conteggio dei globuli rossi!

5 – qualche volta è disponibile solo un piccolo campione di dati, 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.

Si usa perciò un **campione**, 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.

Distribuzione delle statistiche campionarie

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 e le corrispondenti frequenze.

Distribuzione della media campionaria
(varianza nota)

Si estrae un primo campione casuale di n elementi da una data popolazione, e si indica con $\bar{x}_1$ la sua media; se si estrae un secondo campione di n elementi dalla stessa popolazione, si ottiene un altro valore per la media $\bar{x}_2$, 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.

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

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

Distribuzione della media campionaria
(varianza nota)

$$E[\overline{X}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$$
$$= \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n}$$
$$= \frac{n\mu}{n} = \mu \qquad (6.2.2)$$

e, per la varianza,

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$
$$= \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)}{n^2} \quad \text{per l'indipendenza}$$
$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad (6.2.3)$$

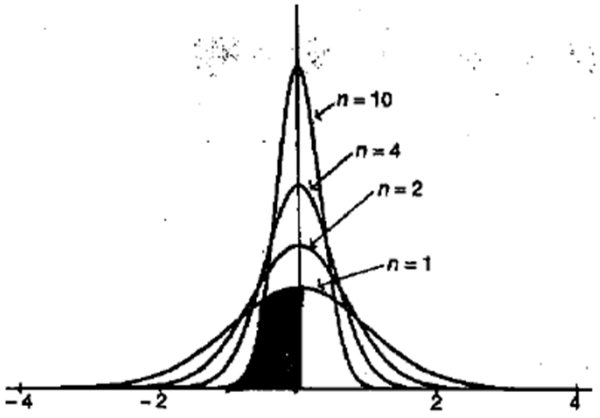


Figura 6.1 Densità delle medie campionarie di una popolazione normale standard.

Distribuzione della media campionaria
(varianza nota)

Esempio

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

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

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

Consideriamo tutti i possibili campioni di dimensione $n = 2$ estraibili da questa popolazione; con reimmissione

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

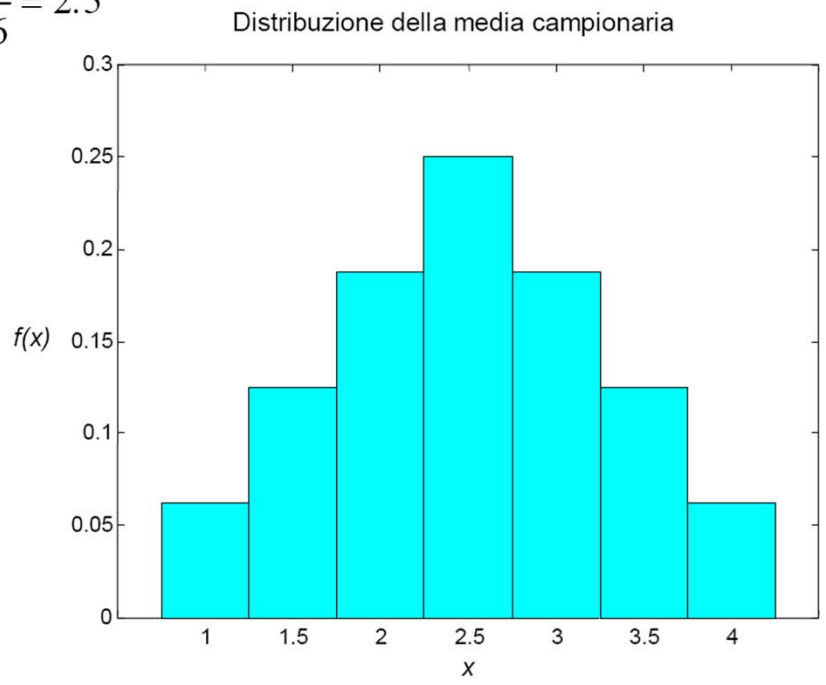
Distribuzione della media campionaria
(varianza nota)

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

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

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

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{1.25}{2} = 0.625$$



Distribuzione della media campionaria
(varianza nota)

senza reimmissione

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

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

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

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

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

Teorema 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

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

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.

Teorema del limite centrale

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 $n \geq 30$.

Teorema del limite centrale

Esempio

I pesi di 20000 cuscinetti a sfere sono distribuiti normalmente con media $\mu = 22.4$g e scarto quadratico medio $\sigma = 0.048$g. Se da questa popolazione vengono estratti 300 campioni casuali di ampiezza 36, determinare la media e lo scarto quadratico medio della distribuzione della media campionaria nel caso che il campionamento venga fatto con reimmissione o senza reimmissione.

Determinare per quanti dei campioni casuali la media

a – è compresa fra 22.39 e 22.41;

b – è superiore a 22.42;

c – è inferiore a 22.37.

$$\mu_{\overline{X}} = \mu = 22.4 \qquad \sigma_{\overline{X}} = \frac{0.048}{\sqrt{36}} = 0.008$$

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 22.4}{0.008}$$

a –

$$\overline{X} = 22.39 \quad \Rightarrow \quad Z = \frac{22.39 - 22.4}{0.008} = -1.25$$

$$\overline{X} = 22.41 \quad \Rightarrow \quad Z = \frac{22.41 - 22.4}{0.008} = 1.25$$

$$P\left(22.39 \le \overline{X} \le 22.41\right) = P(-1.25 \le Z \le 1.25) =$$

$$= 2P(Z \le 1.25) - 1 = 2 \cdot 0.8944 - 1 = 0.7888$$

Il numero di campioni atteso è $300 \cdot 0.7888 = 237$.

Teorema del limite centrale

Esempio

Per un certo segmento ampio di popolazione e per un dato anno, il numero medio di giorni di assenza dal lavoro per malattia è 5.4 con una deviazione standard di 2.8 giorni. Calcolare la probabilità che un campione casuale di 49 persone estratto da questa popolazione abbia una media di assenze

a – maggiore di 6 giorni;

b – fra 4 e 6 giorni;

c – fra 4 giorni e mezzo e 5 giorni e mezzo.

$$\mu_{\overline{X}} = \mu = 5.4 \qquad \sigma_{\overline{X}} = \frac{2.8}{\sqrt{49}} = 0.4 \qquad Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 5.4}{0.4}$$

$$\text{a} - \ \overline{X} = 6 \quad \Rightarrow \quad Z = \frac{6 - 5.4}{0.4} = 1.5$$

$$P\left(\overline{X} > 6\right) = P(Z > 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668$$

Distribuzione della media campionaria
(varianza ignota)

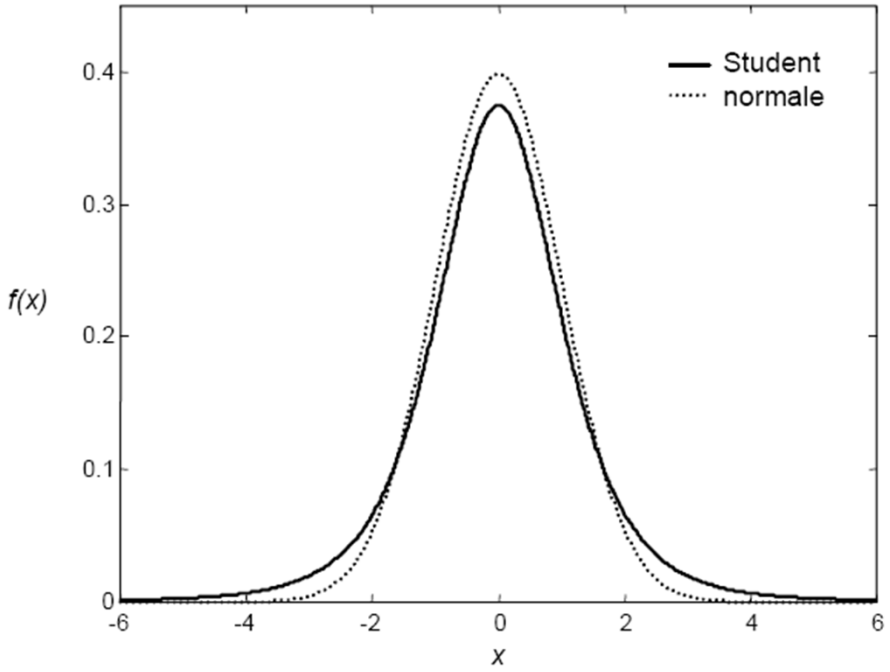
Nel caso che il numero n degli elementi del campione sia grande (**grande campione**), se σ^2 non è nota, si sostituisce a σ^2 la varianza s^2 del campione.
Se invece l'ampiezza n del campione è piccola (**piccolo campione**), si hanno dei risultati solo se il campione proviene da una popolazione normale.

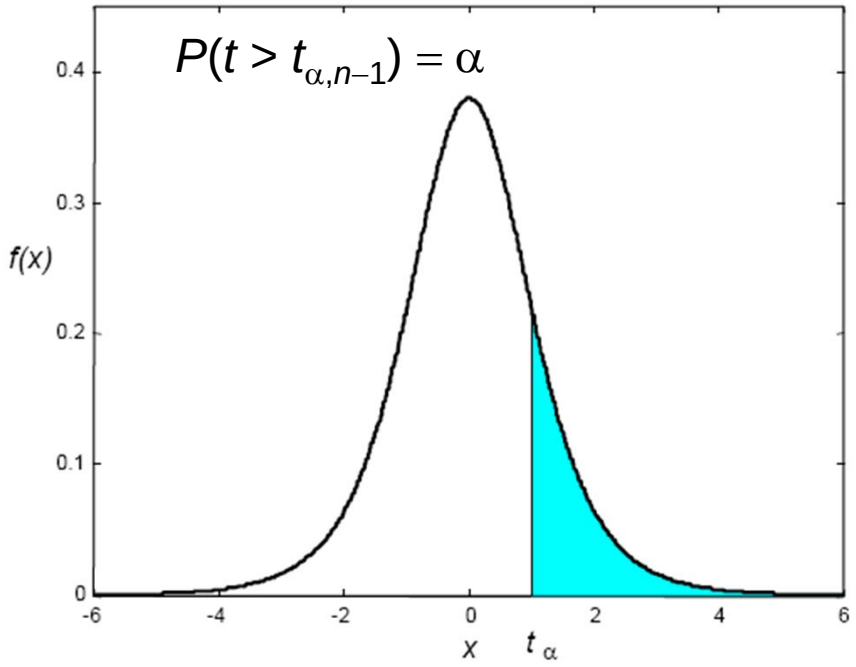
Teorema 3
Sia data una popolazione normale avente media μ e da essa si estraggano campioni casuali di ampiezza n; indicando con $\overline{X}$ la media campionaria e con S lo scarto quadratico medio campionario, la variabile

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

è una variabile aleatoria avente la **distribuzione *t* di Student**[3] con grado di libertà $\nu = n - 1$.

Distribuzione della media campionaria
(varianza ignota)





Si può dimostrare che la distribuzione t con grado di libertà ν tende alla distribuzione normale standardizzata per $\nu \to \infty$.

I valori di t_α per $\nu > 29$ sono circa uguali ai corrispondenti valori tratti dalle tavole della distribuzione normale (vedere esempi 15 e 16): infatti la distribuzione normale è una buona approssimazione della distribuzione t per valori del grado di libertà $\nu > 29$.

$$P(t > t_{\alpha,n-1}) = \alpha$$

Tabella A.3 Valori assunti da $t_{\alpha,n}$

	α				
n	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
70	1.294	1.667	1.994	2.381	2.648
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.326	2.576

Distribuzione Chi Quadrato

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

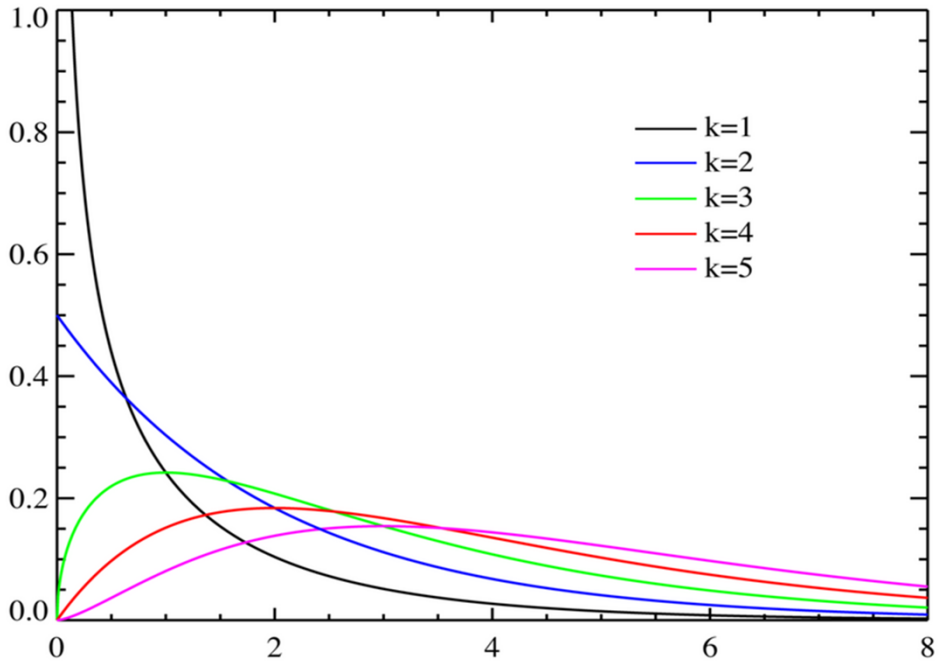
Sia data una popolazione normale avente varianza σ^2 e da essa si estraggano campioni casuali di ampiezza n; indicando con S^2 la varianza campionaria, la variabile

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

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

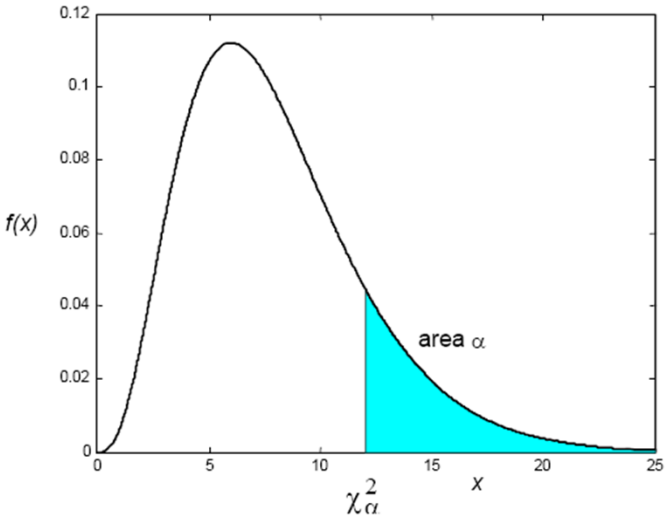
Si dimostra che la distribuzione χ^2 ha media $\mu = \nu$ e varianza $\sigma^2 = 2\nu$.

Distribuzione Chi Quadrato



$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Nota: k indica il numero di gradi di libertà. Per quanto riguarda le funzioni di ripartizione e i percentili si consultano le tabelle.
Lo vedremo solo in occasione del test chi quadrato.



$$P(\chi^2 > \chi^2_{\alpha,n-1}) = \alpha$$

Tabella A.2 Valori assunti da $\chi^2_{\alpha,n}$

	α							
n	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.00098	0.00393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672