Distribuzioni notevoli, introduzione all'inferenza statistica e statistiche campionarie

10 maggio 2017

Distribuzioni utilizzate in statistica

Definizione di distribuzione Gamma

Una v.a. X assolutamente continua è detta avere una densità Gamma di parametri $\alpha > 0$ e $\lambda > 0$ se ha densità

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbb{I}_{(0,+\infty)}(x)$$

dove $\Gamma(\cdot)$ è la funzione Gamma di Eulero e rappresenta la costante di normalizzazione che rende l'integrale di f_X su $\mathbb{R}$ uguale a uno. Scriveremo

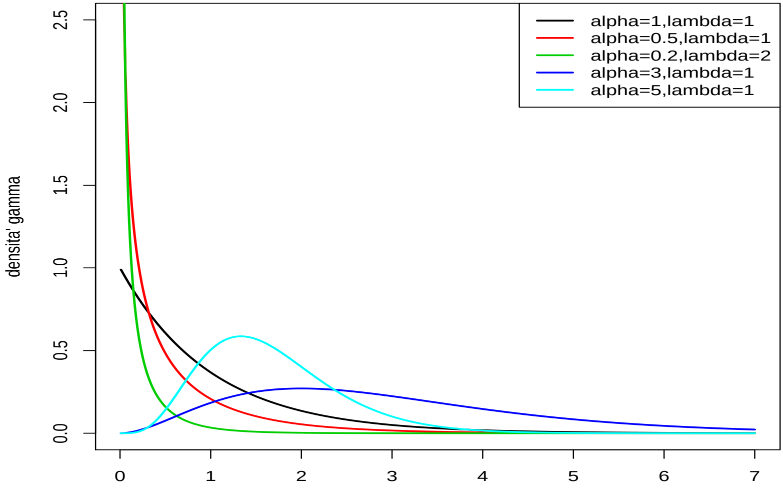
$$X \sim \Gamma(\alpha, \lambda).$$

α è detto parametro di forma

λ è detto parametro di scala.

N.B. Se $\alpha = 1$, allora $\Gamma(1, \lambda) = \mathcal{E}(\lambda)$ (esponenziale di parametro λ)

Densità Gamma



Funzione Gamma di Eulero

Definizione

Per ogni $\alpha > 0$

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x}\, dx = \int_0^{+\infty} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} dx \qquad \forall \lambda > 0$$

Proprietà

- $\Gamma(1) = \int_0^{+\infty} e^{-x}\, dx = 1;$
- Se $\alpha > 1$, integrando per parti

$$\begin{aligned}\Gamma(\alpha) &= -x^{\alpha-1} e^{-x}\Big|_0^{+\infty} + (\alpha-1)\int_0^{+\infty} x^{(\alpha-1)-1} e^{-x} dx \\ &= (\alpha-1)\Gamma(\alpha-1)\end{aligned}$$

- Utilizzando le precedenti per $n > 1$ intero:

$$\Gamma(n) = (n-1)\Gamma(n-1) = \cdots = (n-1)!$$

Funzione generatrice dei momenti (f.g.m.)

Calcoliamo la f.g.m. di $X \sim \Gamma(\alpha, \lambda)$. Sia $t < \lambda$:

$$
\begin{aligned}
m_X(t) := \mathbb{E}(e^{tX}) &= \int_0^{+\infty} e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\
&= \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} dx \\
&= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \int_0^{+\infty} \frac{(\lambda - t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} dx \\
&= \left(\frac{\lambda}{\lambda - t}\right)^\alpha.
\end{aligned}
$$

N.B. Se $X \sim \mathcal{E}(\lambda) = \Gamma(1, \lambda) \Rightarrow m_X(t) = \frac{\lambda}{\lambda - t}$ se $t < \lambda$.

Calcolo di media e varianza

Sappiamo che la f.g.m. di $X \sim \Gamma(\alpha, \lambda)$ è, per $t < \lambda$,

$$m_X(t) := \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}.$$

Calcoliamo ora la derivata prima e seconda:

$$m'_X(t) = \frac{d}{dt}\left(\frac{\lambda}{\lambda - t}\right)^{\alpha} = \frac{\alpha\lambda^{\alpha}}{(\lambda - t)^{\alpha+1}},$$

$$m''_X(t) = \frac{d^2}{(dt)^2}\left(\frac{\lambda}{\lambda - t}\right)^{\alpha} = \frac{d}{dt}\frac{\alpha\lambda^{\alpha}}{(\lambda - t)^{\alpha+1}} = \frac{\alpha(\alpha+1)\lambda^{\alpha}}{(\lambda - t)^{\alpha+2}}.$$

Quindi

$$\mathbb{E}(X) = m'_X(0) = \frac{\alpha}{\lambda} \qquad \mathbb{V}\mathrm{ar}(X) = m''_X(0) - (m'_X(0))^2 = \frac{\alpha}{\lambda^2}$$

Somma di v.a. Gamma indipendenti

Siano $X_1 \sim \Gamma(\alpha_1, \lambda)$ e $X_2 \sim \Gamma(\alpha_2, \lambda)$ v.a. indipendenti. Vogliamo determinare la distribuzione di $X_1 + X_2$.
Calcoliamo la f.g.m. di $X_1 + X_2$

$$\begin{aligned} m_{X_1+X_2}(t) &= \mathbb{E}(e^{t(X_1+X_2)}) = \mathbb{E}(e^{tX_1})\mathbb{E}(e^{tX_2}) \\ &= \Big(\frac{\lambda}{\lambda - t}\Big)^{\alpha_1}\Big(\frac{\lambda}{\lambda - t}\Big)^{\alpha_2} \\ &= \Big(\frac{\lambda}{\lambda - t}\Big)^{\alpha_1+\alpha_2} \end{aligned}$$

per l'indipendenza di X_1 e X_2. Quindi la f.g.m di $X_1 + X_2$ coincide con la f.g.m di una distribuzione $\Gamma(\alpha_1 + \alpha_2, \lambda)$ e per la corrispondenza biunivoca tra funzioni generatrici dei momenti e distribuzioni possiamo concludere

$$X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$$

Esempi

- Siano $X_1, \dots, X_n$ i.i.d., $X_i \sim \mathcal{E}(\lambda) = \Gamma(1, \lambda) \Rightarrow$ $Y = X_1 + \dots + X_n \sim \Gamma(n, \lambda)$ e quindi ha densità

$$f_Y(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} \mathbb{I}_{(0,+\infty)}(x).$$

- Sia $Z \sim \mathcal{N}(0, 1)$, allora Z^2 ha densità

$$\begin{aligned} f_{Z^2}(x) &= \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} \mathbb{I}_{(0,+\infty)}(x) \\ &= \left(\frac{1}{2}\right)^{1/2} \frac{1}{\Gamma(1/2)} x^{1/2-1} e^{-\frac{x}{2}} \mathbb{I}_{(0,+\infty)}(x) \\ &\Rightarrow Z^2 \sim \Gamma(1/2, 1/2). \end{aligned}$$

- Siano $Z_1, \dots, Z_n$ i.i.d., $Z_i \sim \mathcal{N}(0,1) \Rightarrow$ $W = Z_1^2 + \cdots + Z_n^2 \sim \Gamma(n/2, 1/2)$ e quindi ha densità

$$f_W(x) = \frac{\lambda^{n/2}}{\Gamma(n/2)} x^{n/2-1} e^{-\lambda x} \mathbb{I}_{(0,+\infty)}(x).$$

Densità chi quadrato

La densità Gamma di parametri $\alpha = k/2, \lambda = 1/2$, $k \geq 1$ intero, è detta densità chi quadrato con k gradi di libertà e si indica con $\chi^2(k)$.

$\Rightarrow Z^2 \sim \chi^2(1)$ se $Z \sim \mathcal{N}(0,1)$
$\Rightarrow Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n)$ se $Z_i \sim \mathcal{N}(0,1)$ e indipendenti.

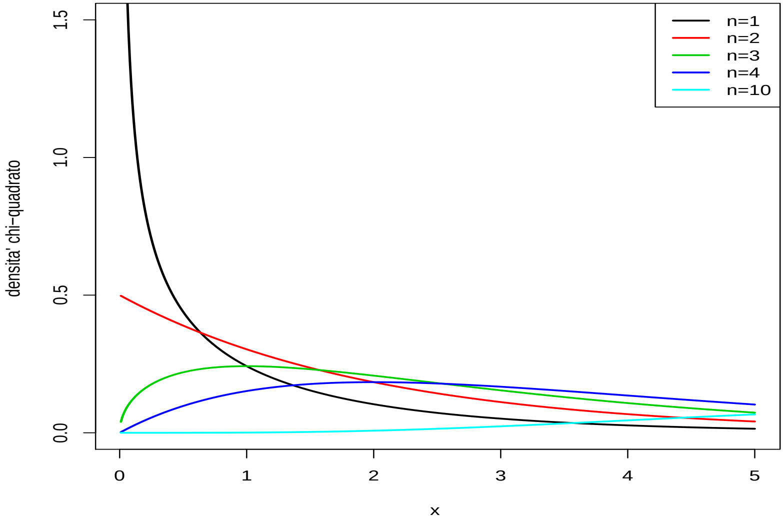
...ancora sulle distribuzioni chi quadrato

- Se $X_1 \sim \chi^2(n)$ e $X_2 \sim \chi^2(m)$ e sono indipendenti $\Rightarrow$ $X_1 + X_2 \sim \chi^2(n+m)$
- Se $X \sim \chi^2(n) \Rightarrow$ la sua f.g.m è

$$m_X(t) = \left(\frac{1/2}{1/2 - t}\right)^{n/2} = \left(\frac{1}{1-2t}\right)^{n/2} \qquad t < 1/2$$

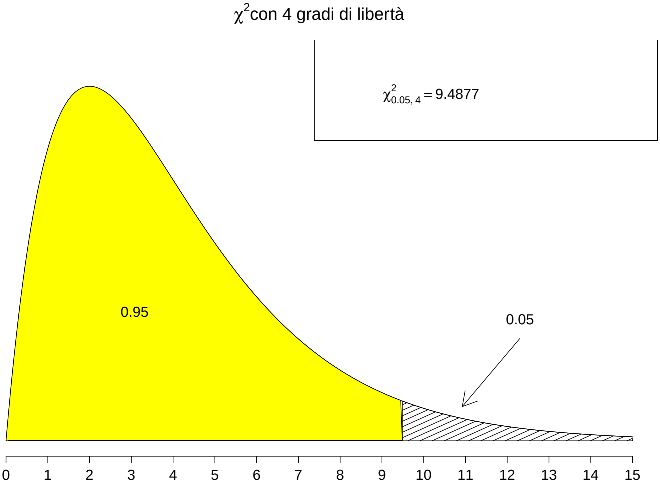
- $\chi^2(2) = \Gamma(2/2, 1/2) = \mathcal{E}(1/2)$.
- $\mathbb{E}(X) = n$ e $\mathbb{V}\mathrm{ar}(X) = 2n$.

Densità chi quadrato



- Sia $X \sim \chi^2(n)$. Quantile di coda destra di ordine $\alpha \in (0,1)$: è l'unico numero che indichiamo con $\chi^2_{\alpha,n}$ tale che

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



...ancora sulle distribuzioni utilizzate in statistica

Definizione di distribuzione t di Student

Se Z e χ_n^2 sono v.a. indipendenti con $Z \sim \mathcal{N}(0,1)$ e $\chi_n^2 \sim \chi^2(n)$, allora la distribuzione della v.a.

$$T_n := \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

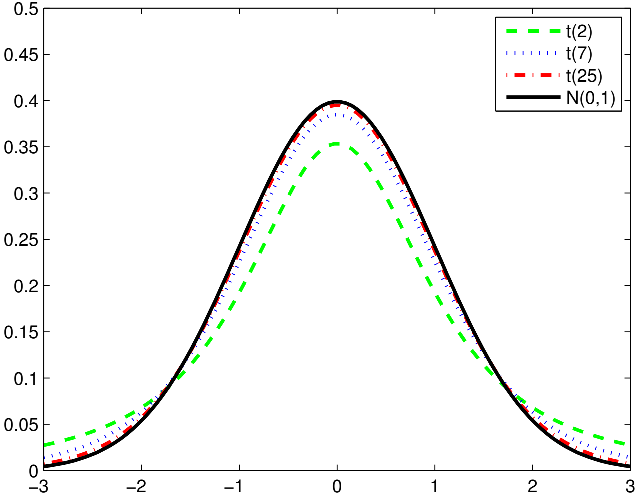
è detta t (di Student) con n gradi di libertà e indicheremo

$$T_n \sim t(n).$$

Proprietà di $T_n \sim t(n)$

- Le v.a. T_n e $-T_n$ hanno la stessa distribuzione.
  Infatti $Z \sim \mathcal{N}(0,1)$ e quindi anche $-Z \sim \mathcal{N}(0,1)$ ha la stessa distribuzione. Allora
  $$-T_n = \frac{-Z}{\sqrt{\frac{\chi_n^2}{n}}} \sim T_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}.$$
- Si può dimostrare che T_n è una v.a. assolutamente continua e quindi, la sua densità f_{T_n} è una funzione simmetrica rispetto all'asse delle ordinate.
- Per $n \to +\infty$
  $$F_{T_n}(x) := P(T_n \leq x) \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du.$$
  Quindi, per "n grande" ($n \simeq 100$), si possono usare le tavole della f.d.r. della gaussiana standard.

Grafici di densità t-Student e gaussiana standard



- Se $n > 2$ si ha $\mathbb{E}(T_n) = 0$ e $\mathbb{V}\mathrm{ar}(T_n) = \frac{n}{n-2}$.
- Quantile di coda destra di ordine $\alpha \in (0, 1)$: è l'unico numero che indichiamo con $t_{\alpha,n}$ tale che

$$P(T_n \geq t_{\alpha,n}) = P(T_n > t_{\alpha,n}) = \alpha$$

N.B. Per la simmetria della densità t-Student si ha

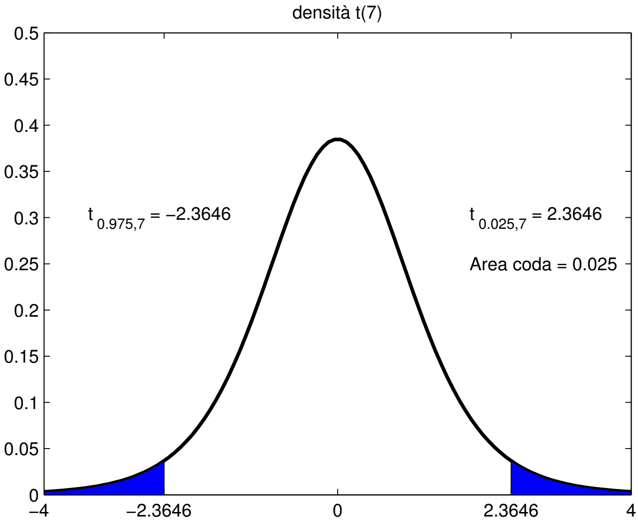
$$-t_{\alpha,n} = t_{1-\alpha,n}.$$

Infatti

$$\begin{aligned} P(T_n \geq -t_{\alpha,n}) &= 1 - P(T_n \leq -t_{\alpha,n}) \\ &= 1 - P(T_n \geq t_{\alpha,n}) = 1 - \alpha \end{aligned}$$

Quindi i quantili sono tabulati solo per $\alpha \leq 0.5$.

Quantile di ordine $\alpha = 0.025$ e $1 - \alpha = 0.975$ di una $t(7)$



...ancora sulle distribuzioni utilizzate in statistica

Definizione di distribuzione F di Fisher

Se χ_n^2 e $\bar{\chi}_m^2$ sono v.a.indipendenti con $\chi_n^2 \sim \chi^2(n)$ e $\bar{\chi}_m^2 \sim \chi^2(m)$, allora la distribuzione della v.a.

$$F_{n,m} := \frac{\chi_n^2/n}{\bar{\chi}_m^2/m}$$

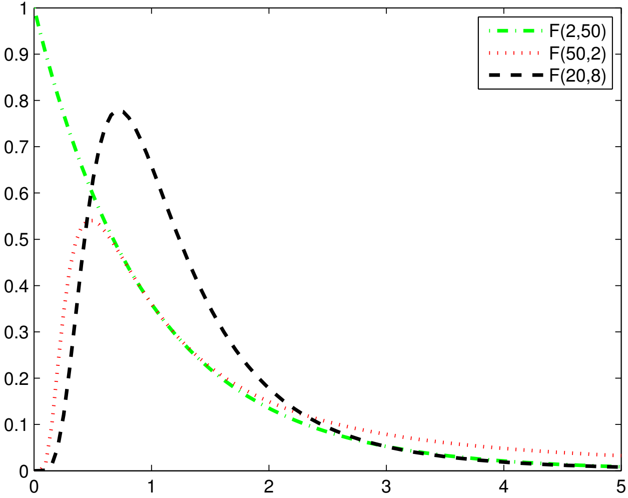
è detta F (di Fisher) con n e m gradi di libertà. Scriveremo

$$F_{n,m} \sim \mathbb{F}(n, m)$$

.

N.B. È importante l'ordine con cui compaiono n e m. Il primo intero n si riferisce ai gradi di libertà del numeratore, il secondo m a quelli del denominatore.

Grafici di densità F di Fisher



Proprietà di $F_{n,m}$ v.a. di Fisher

- $F_{n,m}$ è una v.a. positiva in quanto quoziente di due v.a. χ^2.
- Quantile di coda destra di ordine $\alpha \in (0,1)$: è l'unico numero che indichiamo con $f_{\alpha,n,m}$ tale che

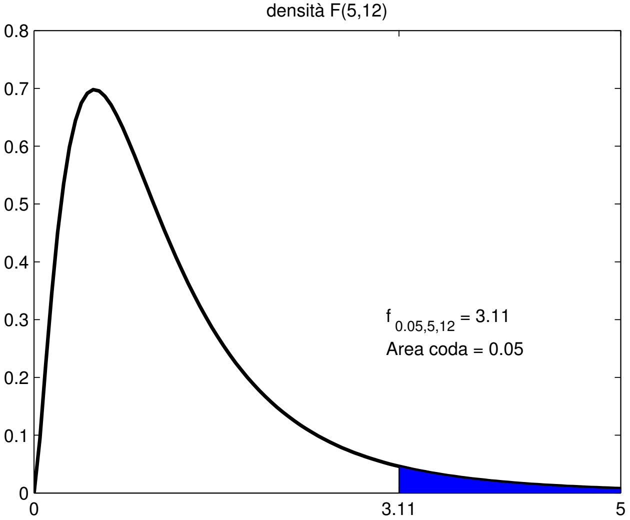
$$P(F_{n,m} \geq f_{\alpha,n,m}) = P(F_{n,m} > f_{\alpha,n,m}) = \alpha.$$

N.B. I quantili $f_{\alpha,n,m}$ sono tabulati per diversi valori di m e n e per valori di $\alpha \leq 0.5$. Per gli ordini > 0.5 basta osservare che:

$$\begin{aligned} 1-\alpha &= 1 - P\left(\frac{\chi^2_n/n}{\bar{\chi}^2_m/m} \geq f_{\alpha,n,m}\right) \\ &= 1 - P\left(\frac{\bar{\chi}^2_m/m}{\chi^2_n/n} < \frac{1}{f_{\alpha,n,m}}\right) = P\left(\frac{\bar{\chi}^2_m/m}{\chi^2_n/n} \geq \frac{1}{f_{\alpha,n,m}}\right) \end{aligned}$$

e quindi $f_{1-\alpha,m,n} = \frac{1}{f_{\alpha,n,m}}$.

Quantile di ordine $\alpha = 0.05$ di una $\mathbb{F}(5, 12)$ (distribuzione F di Fisher con 5,12 gradi di libertà)



Il problema dell'inferenza statistica

Abbiamo già detto che la statistica è la scienza che si occupa di trarre conclusioni dai dati sperimentali.

La situazione tipica è quella in cui si studia un insieme molto grande, detto POPOLAZIONE, di oggetti a cui sono associate quantità misurabili.

L'approccio statistico consiste nel selezionare un sottoinsieme ridotto di oggetti detto CAMPIONE e "analizzarlo" per trarre conclusioni valide per la popolazione nel suo insieme (INFERENZA).

Per basare sui dati del campione delle inferenze che riguardino l'intera popolazione è necessario assumere qualche relazione tra il campione e l'intera popolazione.

Un'ipotesi fondamentale - e in molti casi del tutto ragionevole - è che vi sia una distribuzione di probabilità tipica della popolazione, nel senso che da essa si estraggono in modo casuale degli oggetti, e le quantità numeriche loro associate possono essere pensate come valori assunti da v.a. indipendenti e tutte con la stessa distribuzione.

Cioè se osserviamo i valori $x_1, \dots, x_n$ pensiamo che siano i valori assunti da n v.a. i.i.d. $X_1, \dots, X_n$.

Esempio: La popolazione è costituita dagli studenti maschi del Politecnico, X è la v.a. che rappresenta la votazione media degli esami conseguiti $\Rightarrow x_1, \dots, x_n$ la votazione media conseguita da n studenti.

Definizioni

Sia F una funzione di ripartizione su R.

Definizione 1:

Le v.a. $X_1, \ldots, X_n$ sono dette campione casuale di dimensione n estratto da F se sono indipendenti e tutte con la stessa funzione di ripartizione F (i.i.d.).

Definizione 2:

Se osserviamo i valori $X_1 = x_1, \ldots, X_n = x_n$ allora $(x_1, \ldots, x_n)$ è detta realizzazione del campione. (DATI).

Definizione 3:

Problema di inferenza parametrica: F è nota a meno di uno o più parametri incogniti.

Esempio: Si sa che F è una distribuzione gaussiana, ma non si conoscono la media e la varianza. Oppure F è una distribuzione di Poisson di parametro λ incognito.

Definizione 4:

Problema di inferenza non parametrica: non conosciamo la forma analitica della funzione di ripartizione.

Esempio: Si sa solamente che F è una distribuzione assolutamente continua o discreta.

Statistiche

Sia $X_1, \ldots, X_n$ un campione aleatorio estratto da F.

Definizione: statistica basata sul campione

Una statistica basata sul campione è una funzione nota del campione, i.e.

$$D_n = d_n(X_1, \ldots, X_n)$$

dove d_n è una funzione nota delle n v.a. $X_1, \ldots, X_n$.

$\Rightarrow$ una statistica D_n è una v.a..

Esempio: Media campionaria

Sia $X_1, \ldots, X_n$ un campione aleatorio estratto da F.
Indichiamo con μ e σ^2 la loro media e la loro varianza, rispettivamente.

μ e σ^2 sono dette MEDIA E VARIANZA DELLA POPOLAZIONE

Definizione: media campionaria

$$\overline{X}_n := \frac{X_1 + \cdots + X_n}{n}$$

per ogni $n = 1, 2, \ldots$.

N.B. Per ogni n è una funzione nota del campione:

$$\overline{X}_n = d_n(X_1, \ldots, X_n)$$

con $d_n(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n}$.

Proprietà della media campionaria

1. $\mathbb{E}(\overline{X}_n) = \mu$. Infatti:

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)}{n}$$
$$= \frac{n\mu}{n} = \mu.$$

2. $\mathbb{V}\mathrm{ar}(\overline{X}_n) = \frac{\sigma^2}{n}$. Infatti:

$$\mathbb{V}\mathrm{ar}(\overline{X}_n) = \mathbb{V}\mathrm{ar}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\mathbb{V}\mathrm{ar}(X_1) + \cdots + \mathbb{V}\mathrm{ar}(X_n)}{n^2}$$
$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

N.B. La media campionaria è una statistica la cui media coincide con la media μ della popolazione.

La varianza della media campionaria tende a zero per $n \to +\infty$.

3. La Legge debole dei grandi numeri ci dice che: per ogni $\epsilon > 0$

$$P(|\overline{X}_n - \mu| > \epsilon) \to 0$$

per $n \to +\infty$.

4. Per il Teorema centrale del limite:

$$P\left(\frac{(\overline{X}_n - \mu)\sqrt{n}}{\sigma} \leq x\right) \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{\mathrm{u}^2}{2}} \mathrm{du},$$

cioè per "n grande" $\overline{X}_n \simeq \mathcal{N}(\mu, \sigma^2/n)$, qualunque sia la distribuzione F comune alle X_i (purché ammetta media e varianza finite).

Esempio: Varianza campionaria

Sia (come sopra) $X_1, \ldots, X_n$ un campione aleatorio estratto da F.

Definizione: varianza campionaria

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

per ogni $n = 2, 3, \ldots$.

Definizione: deviazione standard campionaria

$$S_n := \sqrt{S_n^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2}$$

per ogni $n = 2, 3, \ldots$.

N.B. Per ogni n sono funzioni note del campione e quindi sono statistiche.

Proprietà della varianza campionaria

1. $\mathbb{E}(S_n^2) = \sigma^2$. Infatti, osserviamo che:

$$(n-1)S_n^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n}(X_i^2 + \overline{X}_n^2 - 2X_i\overline{X}_n)$$
$$= \sum_{i=1}^{n} X_i^2 + n\overline{X}_n^2 - 2n\overline{X}_n^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2.$$

Quindi

$$\mathbb{E}(S_n^2) = \frac{1}{n-1}\left[\mathbb{E}\left(\sum_{i=1}^{n} X_i^2\right) - \mathbb{E}(n\overline{X}_n^2)\right] = \frac{1}{n-1}\left[n\mathbb{E}(X_1^2) - n\mathbb{E}(\overline{X}_n^2)\right]$$
$$= \frac{n}{n-1}\left[\mathbb{V}\mathrm{ar}(X_1) + (\mathbb{E}(X_1))^2 - \mathbb{V}\mathrm{ar}(\overline{X}_n) - (\mathbb{E}(\overline{X}_n))^2\right]$$
$$= \frac{n}{n-1}\left[\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\right] = \sigma^2.$$

Analogamente si dimostra che

2. $\mathbb{V}\mathrm{ar}(S_n^2) = \frac{1}{n}\left[\mu_4 - \frac{n-3}{n-1}\sigma^4\right]$

dove $\mu_4 = \mathbb{E}[(X_1 - \mu)^4]$ e $\sigma^4 = (\sigma^2)^2$

N.B. La varianza campionaria è una statistica la cui media coincide con la varianza σ^2 della popolazione.

La varianza della varianza campionaria tende a zero per $n \to +\infty$.

Distribuzione congiunta delle statistiche $\overline{X}_n$ e S_n^2 nel caso di popolazioni gaussiane

Sia $X_1, \ldots, X_n$ un campione aleatorio estratto da una popolazione gaussiana, i.e.

$$X_1, \ldots, X_n \quad \text{i.i.d.} \quad X_i \sim \mathcal{N}(\mu, \sigma^2)$$

Problema: determinare la distribuzione congiunta di

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{e} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Sappiamo che:

- $\overline{X}_n$ è combinazione lineare di v.a. gaussiane indipendenti e quindi è una v.a. gaussiana. La sua media è $\mathbb{E}(\overline{X}_n) = \mu$ e $\mathbb{V}\mathrm{ar}(\overline{X}_n) = \sigma^2/n$.
- $$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n) \quad \Longleftrightarrow \quad \frac{(\overline{X}_n - \mu)\sqrt{n}}{\sigma} \sim \mathcal{N}(0, 1).$$

Proposizione

Se $X_1, \ldots, X_n$ è un campione di dimensione n estratto da una popolazione gaussiana di media μ e varianza σ^2, allora

$$\overline{X}_n \qquad \text{e} \qquad S_n^2 \qquad \text{sono indipendenti}$$

e tali che

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad (n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

Corollario 1

Se $X_1, \ldots, X_n$ è un campione di dimensione n estratto da una popolazione gaussiana di media μ e varianza σ^2, allora

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

dove $S_n = \sqrt{S_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2}$ è la deviazione standard campionaria e $t(n-1)$ indica la distribuzione t-Student con $n-1$ gradi di libertà.

Dimostrazione.

Ricordiamo che se $Z \sim \mathcal{N}(0,1)$ e $\chi_k^2 \sim \chi^2(k)$, e Z e χ_k^2 sono indipendenti, allora

$$\frac{Z}{\sqrt{\frac{\chi_k^2}{k}}} \sim t(k).$$

Poiché dalla precedente Proposizione sappiamo che.

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad (n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

e sono indipendenti, segue che

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}} = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1).$$

$\square$

Corollario 2

Se $X_1, \ldots, X_n$ è un campione di dimensione n estratto da una popolazione gaussiana di media μ_X e varianza σ_X^2. Se $Y_1, \ldots, Y_m$ è un campione, indipendente dal precedente, di dimensione m estratto da una popolazione gaussiana di media μ_Y e varianza σ_Y^2. Indichiamo con S_X^2 e S_Y^2 le rispettive varianze campionarie, cioè

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \qquad \text{e} \qquad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2$$

Allora, se $\sigma_X^2 = \sigma_Y^2$

$$\frac{S_X^2}{S_Y^2} \sim \mathbb{F}(n-1, m-1)$$

dove $\mathbb{F}(n-1, m-1)$ indica la distribuzione di Fisher con $n-1,\ m-1$ gradi di libertà.

Dimostrazione.

Ricordiamo che se $\chi_k^2 \sim \chi^2(k)$ e $\bar{\chi}_r^2 \sim \chi^2(r)$, e χ_k^2 e $\bar{\chi}_r^2$ sono indipendenti, allora

$$\frac{\chi_k^2}{k} \Big/ \frac{\bar{\chi}_r^2}{r} \sim \mathbb{F}(k, r).$$

Poichè dalla precedente Proposizione sappiamo che.

$$(n-1)\frac{S_X^2}{\sigma_X^2} \sim \chi^2(n-1) \qquad (m-1)\frac{S_Y^2}{\sigma_Y^2} \sim \chi^2(m-1)$$

e sono indipendenti, se $\sigma_X^2 = \sigma_Y^2$ segue che

$$\frac{(n-1)S_X^2}{(n-1)\sigma_X^2} \Big/ \frac{(m-1)S_Y^2}{(m-1)\sigma_Y^2} = \frac{S_X^2}{S_Y^2} \sim \mathbb{F}(n-1, m-1).$$

□