3. Campioni statistici

3.1 Introduzione

Scopo della statistica inferenziale è lo stabilire metodi rigorosi per ottenere – con un calcolabile *grado di certezza* proprietà generali di una popolazione a partire da una raccolta di dati sulla popolazione stessa.

Possiamo sintetizzare il modello matematico che applichiamo come segue

- Se rileviamo un carattere su una popolazione di n individui, consideriamo ciascun dato rilevato come il valore assunto da $X_1, X_2, \ldots, X_n$ variabili aleatorie aventi tutte la stessa distribuzione $\mathbb{P}_X$ e che (molto spesso) si possono supporre indipendenti.
- La distribuzione $\mathbb{P}_X$ è (parzialmente) incognita; si cercano informazioni su $\mathbb{P}_X$ a partire dai dati rilevati. Le informazioni ricavate sulla distribuzione $\mathbb{P}_X$ sono di natura probabilistica. Per esempio, non riusciremo ad ottenere informazioni del tipo *il valore atteso della distribuzione* $\mathbb{P}_X$ *è* 50 ma informazioni del tipo *il valore atteso della distribuzione* $\mathbb{P}_X$ *è compresa tra 49.8 e 50.2 con probabilità del 90%.*

Comunemente si suppone di conoscere il *tipo* della distribuzione $\mathbb{P}_X$, ovvero si suppone di sapere se è gaussiana, esponenziale o binomiale o altro, ma di non conoscere i parametri che la caratterizzano.

Definizione 3.1.1 (Campione statistico)**.** Una famiglia di variabili aleatorie

$$X_1, \ldots, X_n$$

si dice un *campione statistico di numerosità* n se le v.a. $X_1, \ldots, X_n$ sono indipendenti ed identicamente distribuite.
Se $f\colon \mathbb{R} \to [0, +\infty]$ è la comune densità delle v.a. $X_1, \ldots, X_n$, allora la v.a. vettoriale $X := (X_1, \ldots, X_n)$ ha densità congiunta

$$g_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n).$$

La comune distribuzione delle X_i si dice *distribuzione campionaria di* $X_1, \ldots, X_n$.

Osservazione 3.1.1. Poiché le v.a. $X_1, \ldots, X_n$ seguono la stessa distribuzione, esse hanno anche lo stesso valore atteso e la stessa varianza (se queste quantità esistono).

Definizione 3.1.2 (Statistica, stimatori)**.** Sia $X_1, \ldots, X_n$ un campione statistico. Sia $f\colon \mathbb{R}^n \to \mathbb{R}$ una funzione misurabile secondo Borel. Allora la v.a. $Y := f(X_1, \ldots, X_n)$ si dice una statistica del campione $X_1, \ldots, X_n$.

Supponiamo ora di avere una successione $(X_i)_{i\in\mathbb{N}}$ di v.a. i.i.d. e che per ogni $n \in \mathbb{N}$ sia definita una statistica $Y_n = f_n(X_1, \ldots, X_n)$. Se Y_n converge in probabilità ad una v.a. costante con valore λ, dico che Y_n è uno *stimatore consistente in senso debole* di λ. Se Y_n converge quasi certamente ad una v.a. costante con valore λ, dico che Y_n è uno *stimatore consistente in senso forte* di λ.

Se Y_n è uno stimatore consistente di λ e se $\mathbb{E}[Y_n] = \lambda$ per ogni n, dico che Y_n è uno *stimatore corretto* o uno *stimatore non distorto* di λ.

3.2 Media campionaria e varianza campionaria

Definizione 3.2.1. Sia $X_1, \ldots, X_n$ un campione statistico. Chiamiamo **media campionaria** di $X_1, \ldots, X_n$ la statistica

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i,$$

chiamiamo **varianza campionaria** di $X_1, \ldots, X_n$ la statistica

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Proposizione 3.2.1. *Sia $X_1, \ldots, X_n$ un campione statistico di numerosità n con valore atteso μ e varianza σ^2 finiti. Siano $\overline{X}$ e S^2 la media campionaria e la varianza campionaria. Allora*

$$\mathbb{E}\left[\overline{X}\right] = \mu, \qquad \operatorname{Var}\left[\overline{X}\right] = \frac{\sigma^2}{n}, \qquad \mathbb{E}\left[S^2\right] = \sigma^2.$$

Dimostrazione.

$$\mathbb{E}\left[\overline{X}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

$$\operatorname{Var}\left[\overline{X}\right] = \operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}\left[X_i\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Per calcolare il valore atteso di S^2 osserviamo preliminarmente che

$$\begin{aligned} S^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\overline{X}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}n\overline{X} + n\overline{X}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right). \end{aligned}$$

Dunque

$$\begin{aligned} &(n-1)\mathbb{E}\left[S^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu + \mu)^2 - n\left(\overline{X} - \mu + \mu\right)^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu + \mu)^2\right] - n\mathbb{E}\left[\left(\overline{X} - \mu + \mu\right)^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu)^2 + \mu^2 + 2\mu(X_i - \mu)\right] - n\left(\mathbb{E}\left[\left(\overline{X} - \mu\right)^2\right] + \mu^2 + 2\mu\mathbb{E}\left[\overline{X} - \mu\right]\right) \end{aligned}$$

$$= \sum_{i=1}^{n} \left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\,\sigma^2$$

e quindi $\mathbb{E}\left[S^2\right] = \sigma^2$. □

3.2.1 La disuguaglianza di Chebychev e la legge (debole) dei grandi numeri

Enunciamo alcuni importanti risultati asintotici che giustificano l'uso della media campionaria $\overline{X}$ come stima del valore atteso μ del campione.

Teorema 3.2.1 (Disuguaglianza di Chebychev)**.** *Se X è una variabile aleatoria con valore atteso μ e varianza non superiore a σ^2, allora*

$$\mathbb{P}\left(|X - \mu| \geq t\right) \leq \frac{\sigma^2}{t^2} \qquad \forall t > 0.$$

Osservazione 3.2.1. La disuguaglianza di Chebychev può anche essere formulata nel seguente modo: Se X è una variabile aleatoria con valore atteso μ e varianza σ^2 finite, allora

$$\mathbb{P}\left(|X - \mu| > \eta\,\sigma\right) \leq \frac{1}{\eta^2} \qquad \forall \eta > 0.$$

Ovvero: la probabilità che X disti dal suo valore atteso μ più di una frazione η della deviazione standard σ è inferiore a $\frac{1}{\eta^2}$.

Esempio 3.2.1. Sia $X_1, \ \ldots, \ X_n$ un campione statistico di numerosità n. Supponiamo di conoscere la varianza $\sigma^2 = 4$ del campione e che il valore atteso μ sia ignoto. Quanto deve essere grande n per poter affermare che

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| > 1\right) \leq \frac{1}{10}?$$

Sappiamo che

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| > 1\right) \leq \frac{\sigma^2}{n\,1^2} = \frac{4}{n}.$$

è allora sufficiente richiedere $\dfrac{4}{n} \leq \dfrac{1}{10}$ cioè $n \geq 40$.

Dalla disuguaglianza di Chebychev segue facilmente il seguente

Teorema 3.2.2 (Legge debole dei grandi numeri)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con valore atteso μ e varianza σ^2 finiti.*

Per ogni $n \in \mathbb{N}$ sia $\overline{X}_n := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. Allora

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > t\right) = 0 \qquad \forall t > 0.$$

Corollario 3.2.3. *La media campionaria è uno stimatore corretto del valore atteso della distribuzione campionaria.*

Dimostrazione. Sia $(X_i)_{i\in\mathbb{N}}$ una successione di v.a. i.i.d. con valore atteso μ e varianza σ^2, entrambi finiti. Abbiamo già dimostrato che per ogni $n \in N$ la media campionaria $\overline{X}_n$ del campione statistico $X_1, \ldots, X_n$ soddisfa le proprietà

$$\mathbb{E}\left[\overline{X}_n\right] = \mu, \qquad \mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}.$$

Applicando la legge debole dei grandi numeri otteniamo

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \le \frac{\sigma^2}{n\varepsilon^2} \qquad \forall \varepsilon > 0,$$

da cui la tesi. □

La legge debole dei grandi numeri ci *autorizza* a usare il valore di $\overline{X}_n$ come sostituto del valore atteso μ della distribuzione e la disuguaglianza di Chebychev ci dice con precisione quanto è *probabilisticamente accettabile* questa sostituzione.

Esempio 3.2.2. Ho una monetina che potrebbe essere truccata. Voglio scoprire, con un'approssimazione di ± 0.05 e con un grado di certezza del 90% quanto vale la probabilità di ottenere testa in un singolo lancio. Posso formalizzare ogni singolo lancio della monetina con una variabile aleatoria di Bernoulli di parametro p dove p è la probabilità (incognita) di ottenere testa in un singolo lancio. Se lancio la monetina n volte ho allora un campione statistico $X_1, \ldots, X_n$ che segue la distribuzione $B(p)$. Sia $\overline{X}_n$ la media campionaria di questo campione. Allora

$$\mathbb{E}\left[\overline{X}_n\right] = p, \qquad \mathrm{Var}\left[\overline{X}_n\right] = \frac{p(1-p)}{n}.$$

Per la disuguaglianza di Chebychev

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \ge 0.05\right) \le \frac{p(1-p)}{n(0.05)^2} \le \frac{400}{4n} = \frac{100}{n}$$

Voglio

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \le 0.05\right) \ge \frac{90}{100}$$

cioè

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \ge 0.05\right) \le 1 - \frac{90}{100} = \frac{1}{10}$$

Basta allora avere $\frac{100}{n} \le \frac{1}{10}$ cioè $n \ge 1000$. Dunque: tiro la monetina 1000 volte registrando il risultato ad ogni i-esimo lancio ($x_i = 1$) o croce ($x_i = 0$) vedendo questo numero come il valore assunto da una v.a. bernoulliana X_i di parametro p.

Calcolo $\overline{x} = \frac{1}{1000}\sum_{i=1}^{1000} x_i$ e lo vedo come il valore assunto dalla v.a. $\overline{X}$. La probabilità che il valore $\overline{x}$ differisca da p per meno di 0.05 è maggiore-uguale del 90%.

Più in generale

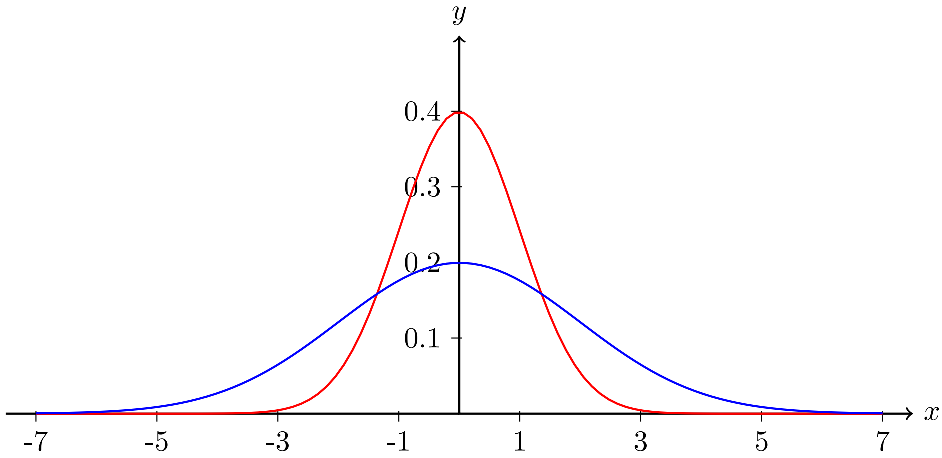

Figura 3.1: Densità associate alle distribuzioni $N(0,1)$ (in rosso) e $N(0,4)$ (in blu)

Esempio 3.2.3. Sia $X_1, \ldots, X_n$ un campione statistico di numerosità n, bernoulliano di parametro (incognito) $p \in [0,1]$. Dunque

$$\mathbb{E}[X_i] = p \qquad \mathrm{Var}[X_i] = p(1-p)$$
$$\mathbb{E}\left[\overline{X}_n\right] = p \qquad \mathrm{Var}\left[\overline{X}_n\right] = \frac{p(1-p)}{n}$$

Allora, per la disuguaglianza di Chebychev

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \le \frac{p(1-p)}{n\,\varepsilon^2} \le \frac{1}{4n\,\varepsilon^2} \qquad \forall \varepsilon > 0.$$

poiché $p(1-p) \le \dfrac{1}{4} \quad \forall p \in [0,1]$.

3.2.2 La distribuzione gaussiana $N(\mu, \sigma^2)$ e il teorema del limite centrale

Ricordiamo che la distribuzione gaussiana di parametri $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, $N(\mu, \sigma^2)$, è la distribuzione assolutamente continua associata alla densità

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

Se una v.a. X segue la distribuzione $N(\mu, \sigma^2)$, allora

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2.$$

Inoltre $f(x) > 0$ per ogni $x \in \mathbb{R}$, quindi la funzione di ripartizione $F_X(x) := \mathbb{P}(X \le x)$ è strettamente monotona crescente. Dunque, per ogni $\alpha \in (0,1)$ esiste uno ed un solo $x = x_\alpha \in \mathbb{R}$ tale $F_X(x_\alpha) = \alpha$. x_α si dice **quantile** di X di livello α. Inoltre, se $\mu = 0$, la densità è una funzione pari, e dunque $F_X(t) + F_X(-t) = 1$ per ogni $t \in \mathbb{R}$; in particolare $x_{1-\alpha} = -x_\alpha$.

Nel caso in cui $\mu = 0$, $\sigma^2 = 1$, la distribuzione $N(0,1)$ si dice *distribuzione gaussiana standard*, la funzione di ripartizione associata si indica con la lettera Φ,

$$\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t, \qquad x \in \mathbb{R}.$$

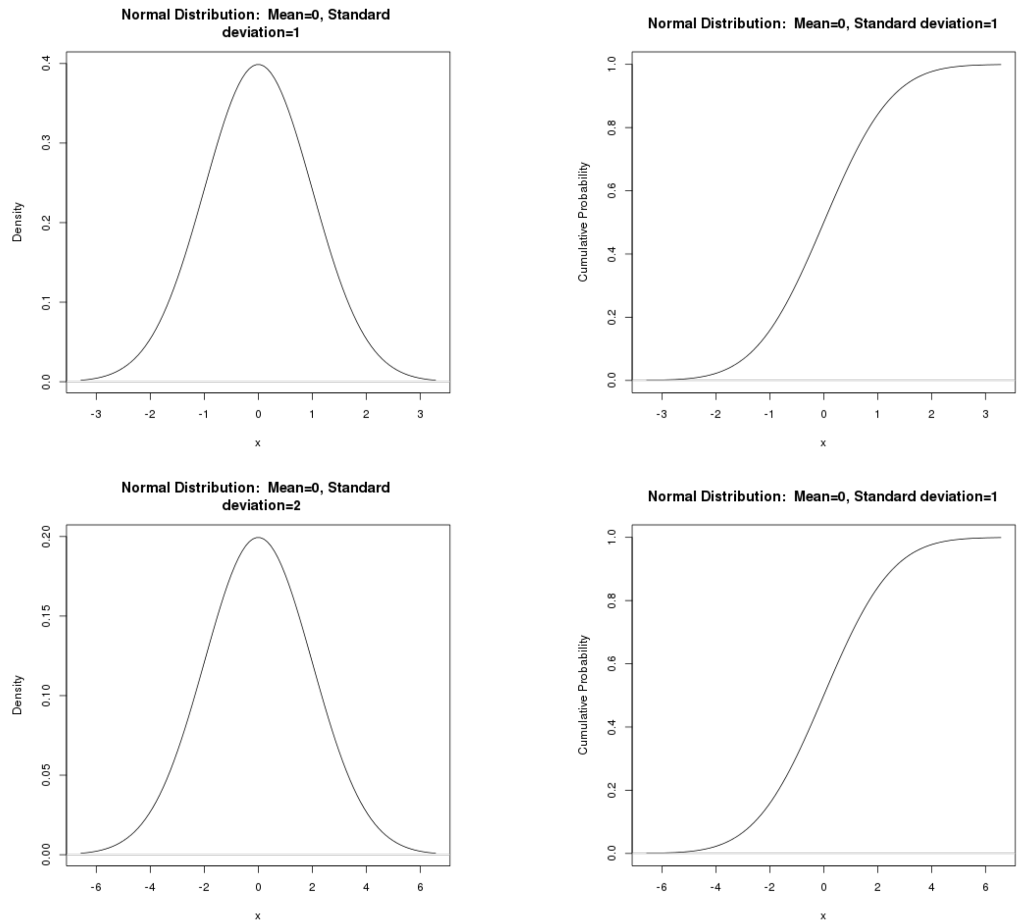

Figura 3.2: $N(0,1)$ e $N(0,4)$, densità e funzione di ripartizione

e per ogni $\alpha \in (0,1)$ il quantile di livello α si indica z_α. Dunque

$$\Phi(x) + \Phi(-x) = 1 \quad \forall x \in \mathbb{R}, \qquad z_{1-\alpha} = -z_\alpha \quad \forall \alpha \in (0,1).$$

Ricordiamo alcune proprietà che abbiamo già visto:

Proprietà 3.2.1. 1. Se X è una v.a. gaussiana di valore atteso μ e varianza σ^2: $\mathbb{P}_X = N\left(\mu, \sigma^2\right)$ e α, β sono due numeri reali, $\alpha \neq 0$, allora la v.a. $\alpha X + \beta$ è gaussiana di valore atteso $\alpha\mu + \beta$ e varianza $\alpha^2\sigma^2$: $\mathbb{P}_{\alpha X+\beta} = N\left(\alpha\mu + \beta, \alpha^2\sigma^2\right)$. In particolare $Y := \dfrac{X-\mu}{\sigma}$ è una v.a. gaussiana standard: $\mathbb{P}_Y = N(0,1)$.

2. Siano $X_1, \ldots, X_n$ v.a. indipendenti con X_i gaussiana di valore atteso μ_i e varianza σ_i^2: $\mathbb{P}_{X_i} = N\left(\mu_i, \sigma_i^2\right)$ $\forall i = 1, \ldots, n$. Allora la v.a. $S_n := X_1 + X_2 + \ldots + X_n$ è gaussiana di valore atteso pari alla somma dei valori attesi e varianza pari alla somma delle varianze:

$$\mathbb{P}_{S_n} = N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

Teorema 3.2.4 (Teorema centrale del limite)**.** *Sia $(X_i)_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con valore atteso μ e varianza σ^2 finiti. Sia $\Phi(t)$ la legge associata alla distribuzione gaussiana standard $N(0,1)$.*

Per ogni $n \in N$ sia $\overline{X}_n$ la media campionaria di $X_1, \ldots, X_n$ e sia $\overline{Z}_n$ la sua standardizzazione:

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

Allora

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{Z}_n \le t\right) = \Phi(t) \qquad \forall t \in \mathbb{R}$$

ed il limite è uniforme in $t \in \mathbb{R}$.

Osservazione 3.2.2. Una formulazione equivalente della tesi del teorema del limite centrale è

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \le t\right) = \Phi(t) \qquad \forall t \in \mathbb{R}.$$

Esempio 3.2.4. Supponiamo di avere un campione statistico di numerosità 25 e deviazione standard 8. Qual è la probabilità che la media campionaria differisca dal valore atteso del campione per più di 4?

Devo calcolare

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| > 4\right)$$

dove $\mu = \mathbb{E}[X_i] \quad \forall i = 1, \ldots, n$ e dunque è anche $\mu = \mathbb{E}\left[\overline{X}\right]$. Applicando la disuguaglianza di Chebychev otteniamo

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| > 4\right) \le \frac{\operatorname{Var}\left[\overline{X}\right]}{4^2} = \frac{64}{25 \cdot 16} = \frac{4}{25} = 0.16$$

Proviamo ad applicare il teorema del limite centrale. Indico con $\overline{Z}$ la standardizzazione della media campionaria. Si ha

$$\begin{aligned}
\mathbb{P}\left(\left|\overline{X} - \mu\right| > 4\right) &= \mathbb{P}\left(\frac{\left|\overline{X} - \mu\right|}{\frac{\sigma}{\sqrt{n}}} > \frac{4}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(\left|\overline{Z}\right| > \frac{4}{\frac{8}{\sqrt{25}}}\right) = \\
&= \mathbb{P}\left(\left|\overline{Z}\right| > \frac{5}{2}\right) = \mathbb{P}\left(\overline{Z} > \frac{5}{2}\right) + \mathbb{P}\left(\overline{Z} < \frac{-5}{2}\right) \\
&\simeq 1 - \Phi(2.5) + \Phi(-2.5) = 2\left(1 - \Phi(2.5)\right) \\
&= 2\left(1 - \Phi(2.5)\right) \simeq 2\left(1 - 0.9938\right) = 0.0124
\end{aligned}$$

Perché questa stima *sembra* tanto migliore di quella ottenuta con la disuguaglianza di Chebychev? Perché non abbiamo un'indicazione sul significato del primo dei $\simeq$. In altre parole, il teorema del limite centrale è appunto un teorema di passaggio al limite e non fornisce una stima dell'errore che si compie sostituendo $\mathbb{P}(Z_n \le t)$ con $\Phi(t)$. A tal proposito vale il seguente

Teorema 3.2.5 (Teorema di Berry–Esseen)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con valore atteso $\mu = 0$, varianza σ^2 e momento terzo*

$\gamma := \mathbb{E}\left[|X_i|^3\right]$ *finiti. Sia* $\Phi(t)$ *la funzione di ripartizione associata alla distribuzione gaussiana standard* $N(0,1)$.

Sia $C := \dfrac{0.8\gamma}{\sigma^3}$. *Allora*

$$\left|\mathbb{P}\left(\frac{\overline{X}_n}{\frac{\sigma}{\sqrt{n}}} \leq t\right) - \Phi(t)\right| \leq \frac{C}{\sqrt{n}} \qquad \forall t \in \mathbb{R}.$$

Dal Teorema di Berry–Esseen, teorema 3.2.5, otteniamo dunque

$$\left|\mathbb{P}\left(\overline{Z}_n \leq t\right) - \Phi(t)\right| \leq \frac{C}{\sqrt{n}} \qquad \forall t \in \mathbb{R}.$$

3.3 Alcune distribuzioni legate alla distribuzione gaussiana

La prima distribuzione che incontreremo sarà la dstribuzione di Pearson che è un caso particolare di distribuzione Γ. Ricordiamo dunque tali distribuzioni.

Definizione 3.3.1. Chiamo distribuzione Gamma di parametri α e λ entrambi reali strettamente positivi e indico con il simbolo $\Gamma(\alpha, \lambda)$, la distribuzione assolutamente continua associata alla densità

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x) & x > 0, \\ 0 & x \leq 0, \end{cases} \tag{3.1}$$

dove $\Gamma(\alpha) := \displaystyle\int_0^{+\infty} x^{\alpha-1}\, \mathrm{e}^{-x} \mathrm{d}x$, $\alpha > 0$.

Osservazione 3.3.1. Verifichiamo che la funzione f definita in (3.1) è una densità di probabilità. Sicuramente $f(x) \geq 0$ per ogni $x \in \mathbb{R}$. Inoltre

$$\begin{aligned} \int_{\mathbb{R}} f(x) \mathrm{d}x &= \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x)\, \mathrm{d}x \qquad \left(\text{con il cambio di variabile } x = \frac{y}{\lambda}\right) \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} \frac{1}{\lambda^\alpha} y^{\alpha-1} \mathrm{e}^{-y} \mathrm{d}y = \frac{1}{\Gamma(\alpha)} \Gamma(\alpha) = 1. \end{aligned}$$

Osservazione 3.3.2. $f(x) > 0$ per ogni $x \in (0, +\infty)$ dunque, se $\mathbb{P}_X = \Gamma(\alpha, \lambda)$, allora X è distribuita su tutta la semiretta $(0, +\infty)$.

Osservazione 3.3.3. Per $\alpha = 1$ otteniamo $\Gamma(1) = \int_0^{+\infty} \mathrm{e}^{-x} \mathrm{d}x = 1$, dunque

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0, \\ 0 & x \leq 0, \end{cases}$$

ovvero $\Gamma(1, \lambda) = \exp(\lambda)$ per ogni $\lambda > 0$.

Osservazione 3.3.4. Per ogni $a > 0$ si ha $\Gamma(a+1) = a\Gamma(a)$. Infatti

$$\Gamma(a+1) = \int_0^{+\infty} x^a \mathrm{e}^{-x} \mathrm{d}x$$

per parti: $f(x) = x^a$, $g'(x) = \mathrm{e}^{-x}$

$$= -x^a\mathrm{e}^{-x}\big|_{x=0}^{x\to+\infty} + \int_0^{+\infty} ax^{a-1}\mathrm{e}^{-x}\mathrm{d}x = 0 + a\Gamma(a).$$

Poiché $\Gamma(1) = 1$ otteniamo, per induzione $\Gamma(n) = (n-1)!$

Inoltre $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Infatti (con la sostituzione $x = y^2$)

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{+\infty} x^{-1/2}\mathrm{e}^{-x}\mathrm{d}x = \int_0^{+\infty} 2\,\mathrm{e}^{-y^2}\mathrm{d}y = \int_{\mathbb{R}} \mathrm{e}^{-y^2}\mathrm{d}y = \sqrt{\pi}.$$

Quindi

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\sqrt{\pi}, \qquad \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3\cdot 1}{2\cdot 2}\sqrt{\pi} = \frac{3!!}{2^2}\sqrt{\pi},$$

$$\ldots\text{per induzione}\ldots \qquad \Gamma\left(\frac{2n+1}{2}\right) = \frac{(2n-1)!!}{2^n}\sqrt{\pi} \quad \text{per ogni intero non-negativo } n.$$

Siamo dunque in grado di calcolare $\Gamma\left(\frac{n}{2}\right)$ per ogni $n \in \mathbb{N}$.

Proposizione 3.3.1. *Se X è una v.a. con distribuzione $\mathbb{P}_X = \Gamma(\alpha, \lambda)$, allora*

$$\mathbb{E}[X] = \frac{\alpha}{\lambda}, \qquad \mathrm{Var}[X] = \frac{\alpha}{\lambda^2}.$$

Dimostrazione. Poiché X è distribuita solo sui reali positivi, $\mathbb{E}[X]$ è ben definito e

$$\begin{aligned}\mathbb{E}[X] &= \int_{\mathbb{R}} xf(x)\mathrm{d}x = \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)}x^\alpha \exp(-\lambda x)\,\mathrm{d}x \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}}\int_0^{+\infty}\frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)}x^{(\alpha+1)-1}\exp(-\lambda x)\,\mathrm{d}x = \frac{1}{\Gamma(\alpha)}\frac{\alpha\Gamma(\alpha)}{\lambda}\cdot 1 = \frac{\alpha}{\lambda}.\end{aligned}$$

Analogamente

$$\mathbb{E}\left[X^2\right] = \int_{\mathbb{R}} x^2 f(x)\mathrm{d}x = \int_0^{+\infty}\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha+1}\exp(-\lambda x)\,\mathrm{d}x$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+2)}{\lambda^{\alpha+2}}\int_0^{+\infty}\frac{\lambda^{\alpha+2}}{\Gamma(\alpha+2)}x^{(\alpha+2)-1}\exp(-\lambda x)\,\mathrm{d}x = \frac{1}{\Gamma(\alpha)}\frac{(\alpha+1)\alpha\Gamma(\alpha)}{\lambda^2}\cdot 1 = \frac{(\alpha+1)\alpha}{\lambda^2}.$$

Dunque $\mathrm{Var}[X] = \frac{(\alpha+1)\alpha}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}$. □

Lemma 3.3.1. *Se X e Y sono due variabili aleatorie indipendenti, con distribuzioni $\mathbb{P}_X = \Gamma(\alpha, \lambda)$, $\mathbb{P}_Y = \Gamma(\beta, \lambda)$, allora la v.a. $X + Y$ ha distribuzione $\Gamma(\alpha+\beta, \lambda)$.*

Dimostrazione. Sappiamo che la distribuzione di $X+Y$ è a.c. con densità $h(x)$ data dal prodotto di convoluzione delle densità associate alle distribuzioni $\Gamma(\alpha,\lambda)$ e $\Gamma(\beta,\lambda)$. Dunque $h(x)=0$ per $x\leq 0$. Per $x>0$ abbiamo invece

$$\begin{aligned}h(x) &= \int_0^x \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \mathrm{e}^{-\lambda y} \frac{\lambda^\beta}{\Gamma(\beta)} (x-y)^{\beta-1} \mathrm{e}^{-\lambda(x-y)} \mathrm{d}y \\ &= \mathrm{e}^{-\lambda x} \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x y^{\alpha-1}(x-y)^{\beta-1} \mathrm{d}y = \qquad\qquad (\text{sostituisco } y = xt) \\ &= \mathrm{e}^{-\lambda x} \frac{\lambda^{\alpha+\beta} x^{\alpha+\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} \mathrm{d}t = C x^{\alpha+\beta-1} \mathrm{e}^{-\lambda x}\end{aligned}$$

dove $C = \dfrac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \displaystyle\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$. Poiché h deve essere una densità di probabilità può solo essere $C = \dfrac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)}$. □

3.3.1 Distribuzione di Pearson (o χ^2) con n gradi di libertà, χ_n^2

Si tratta della distribuzione $\Gamma(\alpha,\lambda)$ dove $\alpha=\frac{n}{2}$, $\lambda=\frac{1}{2}$. È dunque la distribuzione assolutamente continua associata alla densità

$$f(x) := \begin{cases} \dfrac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\dfrac{1}{2}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \exp\left(\frac{-x}{2}\right) & x>0, \\ 0 & x \leq 0. \end{cases}$$

Proprietà 3.3.1. Se X è una v.a. con distribuzione χ^2 a n gradi di libertà, $\mathbb{P}_X = \chi_n^2$, allora

$$\mathbb{E}[X] = n, \qquad \mathrm{Var}[X] = 2n.$$

Dimostrazione. Poiché una v.a. con distribuzione $\Gamma(\alpha,\lambda)$ ha valore atteso α/λ e varianza α/λ^2, in particolare per una v.a. con distribuzione di Pearson abbiamo

$$\mathbb{E}[X] = \frac{\frac{n}{2}}{\frac{1}{2}} = n, \qquad \mathrm{Var}[X] = \frac{\frac{n}{2}}{\left(\frac{1}{2}\right)^2} = 2n.$$

□

Teorema 3.3.2. *Se X e Y sono due variabili di Pearson indipendenti, $\mathbb{P}_X = \chi_n^2$, $\mathbb{P}_Y = \chi_k^2$, allora la v.a. $X+Y$ segue la distribuzione di Pearson a $n+k$ gradi di libertà:*

$$\mathbb{P}_{X+Y} = \chi_{n+k}^2.$$

Dimostrazione. Scegliendo $\alpha=\frac{n}{2}$, $\beta=\frac{k}{2}$, $\lambda=\frac{1}{2}$ nel Lemma 3.3.1, si ottiene la tesi. □

Il seguente teorema dà un legame tra la distribuzione gaussiana e le distribuzioni χ^2:

Teorema 3.3.3. *Se X è una v.a. gaussiana standard, $\mathbb{P}_X = N(0,1)$, allora X^2 segue la distribuzione di Pearson ad un grado di libertà, $\mathbb{P}_{X^2} = \chi_1^2$.*

Dimostrazione. Sappiamo che $\mathbb{P}_X = N(0,1) = f(x)\mathrm{d}x$ con $f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$. Dunque $\mathbb{P}_{X^2} = g(x)\mathrm{d}x$ con

$$g(x) = \begin{cases} 0 & x \leq 0, \\ \dfrac{1}{2\sqrt{x}}\left(f(\sqrt{x}) + f(-\sqrt{x})\right) & x > 0, \end{cases} = \begin{cases} 0 & x \leq 0, \\ \frac{1}{\sqrt{2\pi}} x^{-1/2}\mathrm{e}^{-x/2} & x > 0, \end{cases}$$

cioè $\mathbb{P}_{X^2} = \chi_1^2$. □

Teorema 3.3.4. *Se $X_1, \ldots, X_n$ sono v.a. indipendenti e gaussiane, con X_i di valore atteso μ_i e varianza σ_i^2, $\forall i = 1, \ldots, n$, allora la v.a. $\sum_{i=1}^{n}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ segue la distribuzione di Pearson a n gradi di libertà, χ_n^2.*

Dimostrazione. Poiché la v.a. $\frac{X_i - \mu_i}{\sigma_i}$ ha distribuzione gaussiana standard, applicando i teoremi 3.3.3 e 3.3.2 ed il principio di induzione si ottiene la tesi. □

Corollario 3.3.5. *Se $X_1, \ldots, X_n$ è un campione statistico gaussiano, con valore atteso μ e varianza σ^2, allora la v.a. $\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$ segue una distribuzione χ^2 con n gradi di libertà.*

Esempio 3.3.1. Si vuole localizzare un oggetto puntiforme, misurandone le tre coordinate cartesiane rispetto ad un prefissato sistema di riferimento. L'errore sperimentale, misurato in millimetri per ciascuna delle tre coordinate è una v.a. gaussiana di valore atteso 0 e deviazione standard 2.

Supponendo che i tre errori siano v.a. indipendenti, calcolare la probabilità che la distanza tra la posizione misurata e la posizione reale sia inferiore a 1.2 mm.

Soluzione. Indico con X_1, X_2, X_3, gli errori commessi nella misurazione delle tre coordinate. Per il Teorema di Pitagora la distanza tra le due posizioni è

$$D = \sqrt{X_1^2 + X_2^2 + X_3^2}$$

Vogliamo calcolare $\mathbb{P}(D < 1.2) = \mathbb{P}\left(D^2 < 1.44\right) = \mathbb{P}\left(X_1^2 + X_2^2 + X_3^2 < 1.44\right)$.

Pongo $Z_i := \frac{X_i}{\sigma} = \frac{X_i}{2}$, $i = 1, 2, 3$, da cui $X_i^2 = 4Z_i^2$ e dunque

$$\begin{aligned}\mathbb{P}(D < 1.2) &= \mathbb{P}\left(X_1^2 + X_2^2 + X_3^2 < 1.44\right) = \mathbb{P}\left(4\left(Z_1^2 + Z_2^2 + Z_3^2\right) < 1.44\right) \\ &= \mathbb{P}\left(Z_1^2 + Z_2^2 + Z_3^2 < .36\right).\end{aligned}$$

Basterà dunque controllare (vedi ultima riga del listato a seguire) il valore della funzione di ripartizione delle v.a. di distribuzione χ_3^2 nel punto 0.36 che è (circa) 0.052.

```
> setwd("/home/laura/Documents/didattica/2017-18_analisi_reale/alcuni_appunti")
> .x <- seq(0.015, 18.015, length.out=100)
> plot(.x, dchisq(.x, df=3), xlab="x", ylab="Density",
+ main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")
> plot(.x, pchisq(.x, df=3), xlab="x", ylab="Density",
```

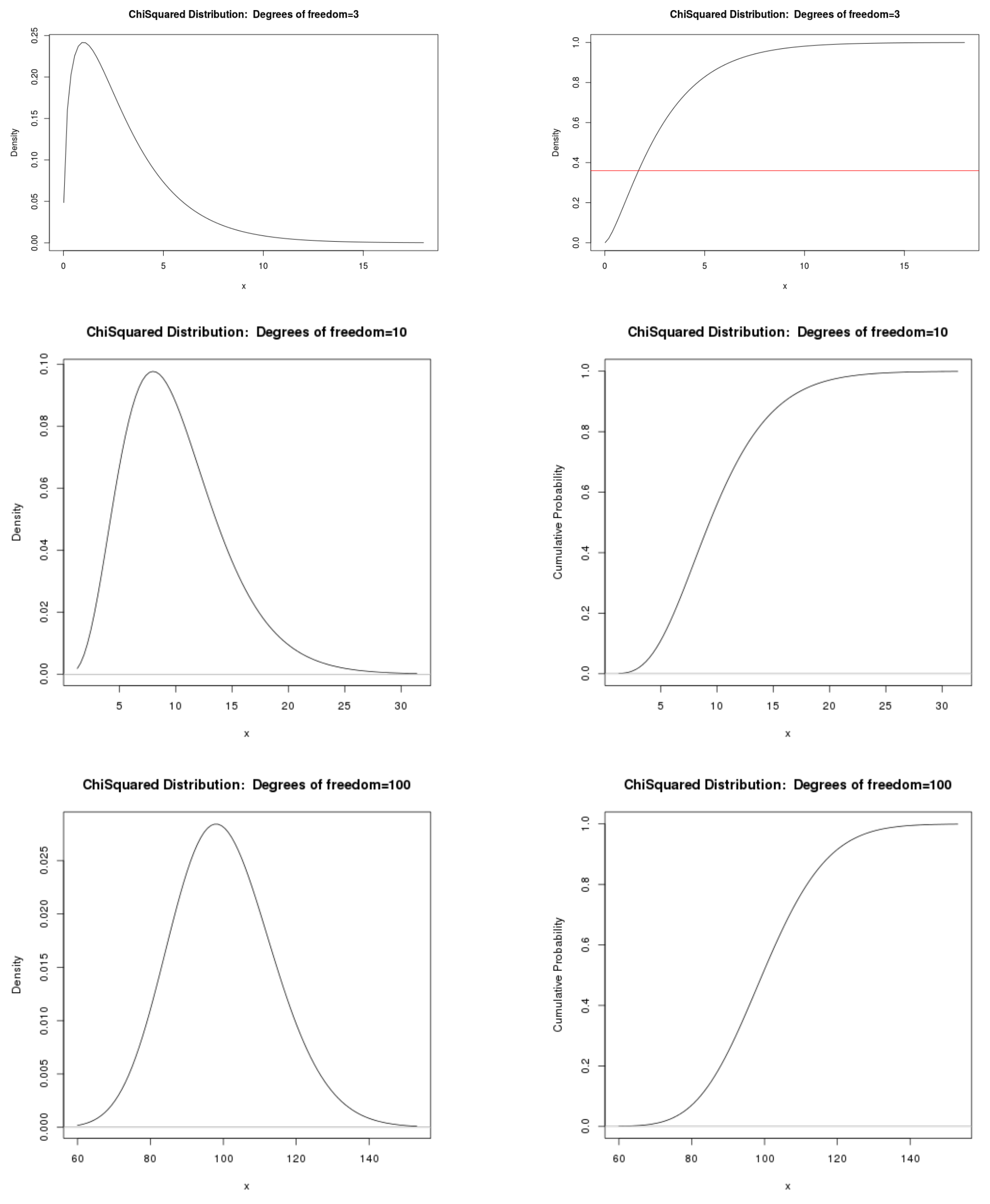

Figura 3.3: χ^2_3, χ^2_{10} e χ^2_{100}, densità e funzione di ripartizione

```
+ main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")
> abline(h=0.36, col="red")
> pchisq(c(0.36), df=3, lower.tail=TRUE)
[1] 0.05162424
```

Il seguente teorema raccoglie alcune importanti proprietà dei campioni statistici gaussiani e delle loro media e varianza campionarie.

Teorema 3.3.6. *Sia $X_1, \ldots, X_n$ un campione statistico gaussiano di numerosità n, valore atteso μ e varianza σ^2.*
Allora, la media campionaria $\overline{X}$ e la varianza campionaria S^2 sono v.a. indipendenti.

Sia $Z_1, Z_2, \ldots Z_n$ la standardizzazione del campione statistico $X_1, \ldots, X_n$ i.e.

$$Z_i := \frac{X_i - \mu}{\sigma} \qquad \forall i = 1, \ldots, n$$

e sia $\overline{Z}$ la media campionaria del campione normalizzato $Z_1, \ldots, Z_n$.
Allora $\overline{Z} = \dfrac{\overline{X} - \mu}{\sigma}$ e la v.a. $\displaystyle\sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2$ sono indipendenti e quest'ultima segue una distribuzione χ^2 con $n-1$ gradi di libertà.

Dimostrazione. **1. n = 2.** Sappiamo che $\mathbb{P}_{X_1+X_2} = N\left(2\mu, 2\sigma^2\right)$ e $\mathbb{P}_{\overline{X}} = N\left(\mu, \sigma^2/2\right)$. Inoltre

$$S^2 = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 = \frac{1}{2}(X_1 - X_2)^2.$$

Dunque $\overline{X}$ e S^2 sono indipendenti se e solo se $X_1 + X_2$ e $X_1 - X_2$ sono indipendenti. Poiché $\mathbb{P}_{-X_2} = N\left(-\mu, \sigma^2\right)$ abbiamo che $\mathbb{P}_{X_1-X_2} = N\left(0, 2\sigma^2\right)$.

Per provare che $U := X_1 + X_2$ e $V := X_1 - X_2$ sono indipendenti ne calcoliamo la densità congiunta e mostriamo che è uguale al prodotto delle densità marginali. Abbiamo già visto che $\mathbb{P}_{X_1+X_2} = N\left(2\mu, 2\sigma^2\right)$. Inoltre, poiché $\mathbb{P}_{-X_2} = N\left(-\mu, \sigma^2\right)$ abbiamo che $\mathbb{P}_{X_1-X_2} = N\left(0, 2\sigma^2\right)$. Posto

$$\varphi\colon (x, y) \in \mathbb{R}^2 \mapsto (x + y, x - y) \in \mathbb{R}^2$$

abbiamo

$$(U, V) = \varphi \circ (X_1, X_2)$$

dunque, per ogni funzione boreliana non-negativa $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ abbiamo

$$\int_{\mathbb{R}^2} \psi(u, v)\mathbb{P}_{U,V}(\mathrm{d}u\mathrm{d}v) = \int_{\mathbb{R}^2} \psi(x + y, x - y)\mathbb{P}_{X_1,X_2}(\mathrm{d}x\mathrm{d}y)$$
$$= \int_{\mathbb{R}^2} \psi(x + y, x - y)\frac{1}{2\pi\sigma^2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) \mathrm{d}x\mathrm{d}y$$

con il cambiamento di variabile $u = x + y$, $v = x - y$

$$= \int_{\mathbb{R}^2} \psi(u, v)\frac{1}{2\pi(\sqrt{2}\sigma)^2} \exp\left(\frac{-(u-2\mu)^2}{2(\sqrt{2}\sigma)^2}\right) \exp\left(\frac{-v^2}{2(\sqrt{2}\sigma)^2}\right) \mathrm{d}u\mathrm{d}v$$

ovvero la densità congiunta è il prodotto delle densità marginali

$$f_{X_1+X_2}(u) = \frac{1}{\sqrt{2\pi(\sqrt{2}\sigma)^2}} \exp\left(\frac{-(u-2\mu)^2}{2(\sqrt{2}\sigma)^2}\right), \quad f_{X_1-X_2}(v) = \frac{1}{\sqrt{2\pi(\sqrt{2}\sigma)^2}} \exp\left(\frac{-v^2}{2(\sqrt{2}\sigma)^2}\right).$$

Inoltre, se Z_1 e Z_2 sono gaussiane standard indipendenti abbiamo:

$$(Z_1 - \overline{Z})^2 + (Z_2 - \overline{Z})^2 = \frac{1}{2}(Z_1 - Z_2)^2 = \left(\frac{Z_1 - Z_2}{\sqrt{2}}\right)^2.$$

La v.a. $Z_1 - Z_2$ ha distribuzione $N(0,2)$, dunque $\dfrac{Z_1 - Z_2}{\sqrt{2}}$ ha distribuzione $N(0,1)$. Applicando il Teorema 3.3.3 otteniamo la tesi.

2. $\mathbf{n \geq 3}$. Procediamo per induzione, supponendo che $\overline{X}_{n-1}$ e S^2_{n-1} siano indipendenti. Osserviamo che

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\big((n-1)\overline{X}_{n-1} + X_n\big) = \frac{n-1}{n}\overline{X}_{n-1} + \frac{1}{n}X_n \tag{3.2}$$

e dunque

$$\overline{X}_n - \overline{X}_{n-1} = \frac{1}{n}\left(X_n - \overline{X}_{n-1}\right).$$

Abbiamo dunque

$$\begin{aligned}
S^2_n &= \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_{n-1} + \overline{X}_{n-1} - \overline{X}_n\right)^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^n \left(X_i - \overline{X}_{n-1}\right)^2 + 2\sum_{i=1}^n \left(\overline{X}_{n-1} - \overline{X}_n\right)\left(X_i - \overline{X}_{n-1}\right) + \sum_{i=1}^n \left(\overline{X}_{n-1} - \overline{X}_n\right)^2\right) \\
&= \frac{1}{n-1}\left((n-2)S^2_{n-1} + \left(X_n - \overline{X}_{n-1}\right)^2 + 2\left(\overline{X}_{n-1} - \overline{X}_n\right) n \left(\overline{X}_n - \overline{X}_{n-1}\right) + n\left(\overline{X}_{n-1} - \overline{X}_n\right)^2\right) \\
&= \frac{1}{n-1}\left((n-2)S^2_{n-1} + \left(X_n - \overline{X}_{n-1}\right)^2 - \frac{2}{n}\left(X_n - \overline{X}_{n-1}\right)\left(X_n - \overline{X}_{n-1}\right) + \frac{1}{n}\left(X_n - \overline{X}_{n-1}\right)^2\right) \\
&= \frac{1}{n-1}\left((n-2)S^2_{n-1} + \frac{n-1}{n}\left(X_n - \overline{X}_{n-1}\right)^2\right)
\end{aligned} \tag{3.3}$$

Per la (3.2) e l'ipotesi di induzione $\overline{X}_n$ è indipendente da S^2_{n-1}. Avremo dunque che S^2_n e $\overline{X}_n$ sono indipendenti se e solo se $\overline{X}_n$ e $X_n - \overline{X}_{n-1}$ sono indipendenti.
Sappiamo che $\mathbb{P}_{X_n} = N\left(\mu, \dfrac{\sigma^2}{n}\right)$, dunque

$$\mathbb{P}_{\overline{X}_n} = N\left(\mu, \frac{\sigma^2}{n}\right), \quad \mathbb{P}_{\overline{X}_{n-1}} = N\left(\mu, \frac{\sigma^2}{n-1}\right), \quad \mathbb{P}_{X_n - \overline{X}_{n-1}} = N\left(0, \sigma^2\frac{n}{n-1}\right),$$

Devo provare che $U := \dfrac{n-1}{n}\overline{X}_{n-1} + \dfrac{1}{n}X_n$ e $V = X_n - \overline{X}_{n-1}$ sono indipendenti. Osserviamo che

$$(U,V) = \varphi \circ \left(\overline{X}_{n-1}, X_n\right), \qquad \varphi(x,y) = \left(\frac{n-1}{n}x + \frac{1}{n}y, y - x\right).$$

Sia dunque $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ una funzione di Borel non negativa. Abbiamo

$$\begin{aligned}
\int_{\mathbb{R}^2} \psi(u,v)\mathbb{P}_{U,V}(\mathrm{d}u\mathrm{d}v) &= \int_{\mathbb{R}^2} \psi\left(\frac{n-1}{n}x + \frac{1}{n}y, y - x\right)\mathbb{P}_{\overline{X}_{n-1},X_n}\mathrm{d}x\mathrm{d}y \\
&= \int_{\mathbb{R}^2} \psi\left(\frac{n-1}{n}x + \frac{1}{n}y, y - x\right)\frac{\sqrt{n-1}}{2\pi\sigma^2}\exp\left(\frac{-(n-1)(x-\mu)^2 - (y-\mu)^2}{2\sigma^2}\right)\mathrm{d}x\mathrm{d}y
\end{aligned}$$

con il cambiamento di variabile $u = \dfrac{n-1}{n}x + \dfrac{1}{n}y$, $v = y - x$

$$= \int_{\mathbb{R}^2} \psi(u,v) \frac{\sqrt{n-1}}{2\pi\sigma^2} \exp\left(\frac{-(u-\mu)^2 \left(\sqrt{n}\right)^2}{2\sigma^2} \right) \exp\left(\frac{-v^2 \left(\sqrt{\frac{n-1}{n}}\right)^2}{2\sigma^2} \right) \mathrm{d}u\mathrm{d}v$$

$$= \int_{\mathbb{R}^2} \psi(u,v) \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} \exp\left(\frac{-(u-\mu)^2}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2} \right) \frac{1}{\sqrt{2\pi\sigma^2\frac{n}{n-1}}} \exp\left(\frac{-v^2}{2\left(\sigma\sqrt{\frac{n-1}{n}}\right)^2} \right) \mathrm{d}u\mathrm{d}v$$

ovvero la densità congiunta è il prodotto delle densità marginali. Questo prova l'indipendenza di U e V e dunque la prima parte della tesi.

Per dimostrare la seconda parte della tesi, osserviamo che essa è sicuramente vera per $n-1$, grazie al Teorema 3.3.3. Procediamo per induzione e riconsideriamo ora la formula (3.3) e supponiamo che essa non sia relativa al campione $X_1, \ldots, X_n$ ma alla sua versione standardizzata $Z_1, \ldots, Z_n$:

$$\sum_{i=1}^{n} \left(Z_i - \overline{Z}_n\right)^2 = (n-1)S_n^2 = (n-2)S_{n-1}^2 + \left(\sqrt{\frac{n-1}{n}} \left(Z_n - \overline{Z}_{n-1}\right) \right)^2 .$$

Poiché il campione $Z_1, \ldots, Z_n$ è gaussiano standard, $\mathbb{P}_{Z_n - \overline{Z}_{n-1}} = N\left(0, \dfrac{n}{n-1}\right)$ dunque la v.a. $\sqrt{\dfrac{n-1}{n}} \left(Z_n - \overline{Z}_{n-1}\right)$ è gaussiana standard e quindi il suo quadrato segue una distribuzione di Pearson con un grado di libertà. D'altra parte, per induzione, $\sum_{i=1}^{n-1} \left(Z_i - \overline{Z}_{n-1}\right)^2 = (n-2)S_{n-1}^2(Z)$ segue una distribuzione di Pearson a $n-2$ gradi di libertà. Per il Teorema 3.3.2 otteniamo la tesi. □

Corollario 3.3.7. *Sia $X_1, \ldots, X_n$ un campione statistico gaussiano di numerosità n, valore atteso μ e varianza σ^2 e sia S^2 la sua varianza campionaria. Allora la v.a. $V := (n-1)\dfrac{S^2}{\sigma^2}$ segue una distribuzione χ^2 con $n-1$ gradi di libertà.*

Dimostrazione. Si ha infatti

$$V = (n-1)\frac{S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n} \left((\mu + \sigma Z_i) - (\mu + \sigma\overline{Z})\right)^2 = \sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2$$

□

3.3.2 Distribuzione t di Student con n gradi di libertà, $t(n)$

Si chiama così la distribuzione associata alla densità

$$\tau_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{\frac{-(n+1)}{2}} \qquad x \in \mathbb{R}.$$

Proprietà 3.3.2. Se X è una v.a. con distribuzione t di Student a n gradi di libertà, allora

$$\mathbb{E}[X] = 0, \qquad \operatorname{Var}[X] = \begin{cases} \dfrac{n}{n-2} & \text{se } n \geq 3, \\ +\infty & \text{se } n = 1,\ 2. \end{cases}$$

Osservazione 3.3.5. Il quantile di livello $\alpha \in (0,1)$ associato alla distribuzione $t(n)$ si indica $t_{n,\alpha}$. Poiché la densità τ_n è una funzione pari, se $\mathbb{P}_X = t(n)$, allora $F_X(x) + F_X(-x) = 1$. Dunque per i quantili della distribuzione $t(n)$ si ha $t_{n,\alpha} = -t_{n,1-\alpha}$ per ogni $\alpha \in (0,1)$.

Teorema 3.3.8. *Se Z è una v.a. gaussiana standard, $\mathbb{P}_Z = N(0,1)$, se Y segue la distribuzione χ^2 con n gradi di libertà, $\mathbb{P}_Y = \chi^2_n$ e se Z e Y sono indipendenti, allora la v.a. $T := \dfrac{Z\sqrt{n}}{\sqrt{Y}}$ segue la distribuzione t di Student a n gradi di libertà: $\mathbb{P}_T = t(n)$.*

Dimostrazione. Possiamo scrivere $T = \varphi \circ (Y, Z)$ dove $\varphi\colon (y,z) \in \mathbb{R}^2 \mapsto \begin{cases} \dfrac{z\sqrt{n}}{y} & y > 0 \\ 0 & y \leq 0 \end{cases} \in \mathbb{R}$.

Sia dunque $\psi\colon \mathbb{R} \to \mathbb{R}$ una funzione di Borel non negativa.

$$\begin{aligned} \int_{\mathbb{R}} \psi(t)\mathbb{P}_T(\mathrm{d}t) &= \int_{y>0, z\in\mathbb{R}} \psi\left(\frac{z\sqrt{n}}{\sqrt{y}}\right) \mathbb{P}_{Y,Z}(\mathrm{d}y\mathrm{d}z) \\ &= \int_{y>0, z\in\mathbb{R}} \psi\left(\frac{z\sqrt{n}}{\sqrt{y}}\right) \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{\frac{n}{2}} y^{\frac{n}{2}-1} \exp\left(\frac{-y}{2}\right) \exp\left(\frac{-z^2}{2}\right) \mathrm{d}y\mathrm{d}z \end{aligned}$$

con il cambio di variabile $t = \dfrac{z\sqrt{n}}{\sqrt{y}}$, $z = \dfrac{t\sqrt{y}}{\sqrt{n}}$, $\mathrm{d}z = \dfrac{\sqrt{y}}{\sqrt{n}}\mathrm{d}t$,

$$= \int_{\mathbb{R}} \psi(t) \frac{1}{\sqrt{2n\pi}} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{\frac{n}{2}} \left(\int_0^{+\infty} y^{\frac{1}{2}} y^{\frac{n}{2}-1} \exp\left(\frac{-y}{2}\right) \exp\left(\frac{-yt^2}{2n}\right)\right) \mathrm{d}t$$

con il cambio di variabile $u = \dfrac{y}{2}\left(1 + \dfrac{t^2}{n}\right)$, $y = 2u\left(1 + \dfrac{t^2}{n}\right)^{-1}$, $\mathrm{d}y = 2\left(1 + \dfrac{t^2}{n}\right)^{-1} \mathrm{d}u$,

$$\begin{aligned} &= \int_{\mathbb{R}} \psi(t) \frac{1}{\sqrt{2n\pi}} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{\frac{n}{2}} \left(\int_0^{+\infty} (2u)^{\frac{n+1}{2}-1} \exp(-u) \left(1 + \frac{t^2}{n}\right)^{\frac{-(n+1)}{2}} \mathrm{d}u \right) \mathrm{d}t \\ &= \int_{\mathbb{R}} \psi(t) \frac{1}{\sqrt{2n\pi}} \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{\frac{n}{2}} \left(1 + \frac{t^2}{n}\right)^{\frac{-(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right) \mathrm{d}t \end{aligned}$$

da cui la tesi. □

Corollario 3.3.9. *Se $X_1, \ \ldots, \ X_n$ è un campione statistico gaussiano di numerosità n, valore atteso μ e varianza σ^2, allora*

$$T := \frac{\left(\overline{X} - \mu\right)\sqrt{n}}{S}$$

segue la distribuzione t di Student con $n-1$ gradi di libertà: $\mathbb{P}_T = t(n-1)$.

Dimostrazione. Basta applicare il teorema 3.3.8 con $Z = \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ e $Y = V = (n-1)\dfrac{S^2}{\sigma^2}$. □

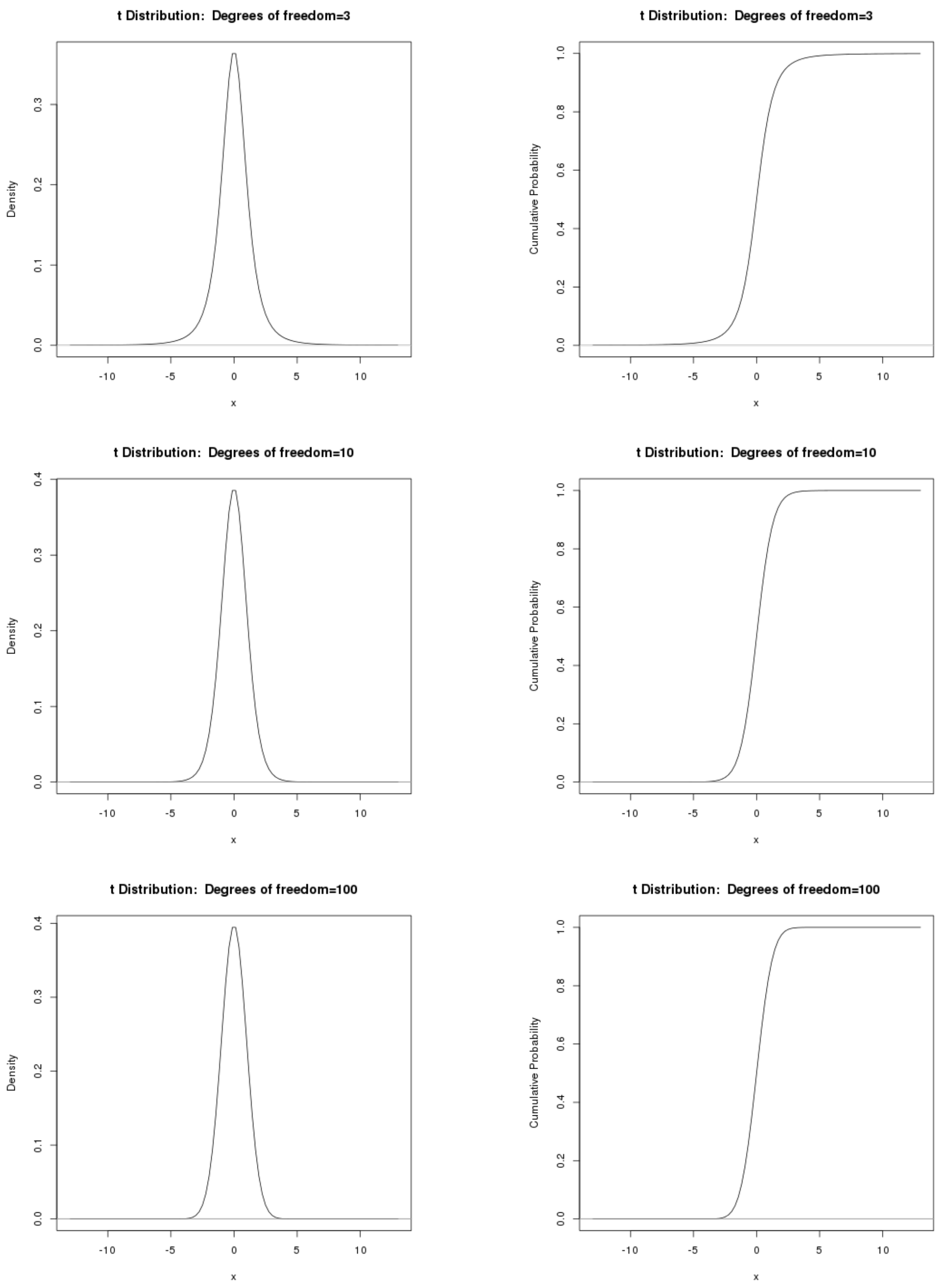

Figura 3.4: $t(3)$, $t(10)$, $t(100)$, densità e funzione di ripartizione