Parte II

Statistica inferenziale

5. Campioni statistici

5.1. 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 $\mathcal{D}$ e che (molto spesso) si possono supporre indipendenti.

- La distribuzione $\mathcal{D}$ è (parzialmente) incognita; si cercano informazioni su $\mathcal{D}$ a partire dai dati rilevati. Le informazioni ricavate sulla distribuzione $\mathcal{D}$ sono di natura probabilistica. Per esempio, non riusciremo ad ottenere informazioni del tipo *la media della distribuzione* $\mathcal{D}$ *è* 50 ma informazioni del tipo *la media della distribuzione* $\mathcal{D}$ *è compresa tra 49.8 e 50.2 con probabilità del 90%.*

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

Definizione 5.1.1 (Campione statistico)**.** Una famiglia di variabili aleatorie

$$X_1,\ X_2,\ \ldots,\ X_n$$

si dice un *campione statistico di numerosità* n se le v.a. $X_1,\ X_2,\ \ldots,\ X_n$ sono indipendenti ed identicamente distribuite.
Se f è la comune densità delle v.a. $X_1,\ X_2,\ \ldots,\ X_n$, allora la v.a. vettoriale
$X := (X_1,\ X_2,\ \ldots,\ X_n)$ ha densità congiunta

$$g_{X_1,\ X_2,\ \ldots,\ X_n}(x_1,\ x_2,\ \ldots,\ x_n) = f(x_1)f(x_2)\ldots f(x_n).$$

La comune distribuzione delle X_i si dice *distribuzione campionaria di* $X_1,\ X_2,\ \ldots,\ X_n$.

Osservazione 5.1.1. Poiché le v.a. $X_1,\ X_2,\ \ldots,\ X_n$ seguono la stessa distribuzione, esse hanno anche la stessa media e la stessa varianza (se queste quantità esistono).

Definizione 5.1.2 (Statistica)**.** Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico. Una funzione (non dipendente da parametri) di $X_1,\ X_2,\ \ldots,\ X_n$ si dice una statistica.

Osservazione 5.1.2. Chiariamo cosa si intende per statistica: $3X_1 - 2X_2$ è una statistica; $\max\{X_1,\ X_2,\ \ldots,\ X_n\}$ è una statistica. $X_1 - \mu\ \mu \in \mathbb{R}$ *non* è una statistica.

5.2. Media campionaria e varianza campionaria

Definizione 5.2.1. Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico. Chiamiamo **media campionaria** di $X_1,\ X_2,\ \ldots,\ X_n$ la statistica

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

chiamiamo **varianza campionaria** di $X_1,\ X_2,\ \ldots,\ X_n$ la statistica

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

Proposizione 5.2.1. *Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico di numerosità n con media μ e varianza σ^2 finite. Siano $\overline{X}$ e S^2 la media campionaria e la varianza campionaria. Allora*

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

Dimostrazione.

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

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

Per calcolare la media 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

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

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

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

5.2.1. La disuguaglianza di Chebyshev e la legge (debole) dei grandi numeri

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

Teorema 5.2.1 (Disuguaglianza di Chebyshev)**.** *Se X è una variabile aleatoria con media μ e varianza σ^2 finite, allora*

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

Dimostrazione. Consideriamo la v.a. $Y := \begin{cases} 0 & \text{se } |X-\mu| < t, \\ t & \text{se } |X-\mu| \geq t. \end{cases}$
Sicuramente $0 \leq Y \leq |X-\mu|$, quindi $Y^2 \leq (X-\mu)^2$ e dunque

$$\mathbb{E}\left[Y^2\right] \leq \mathbb{E}\left[(X-\mu)^2\right] = \operatorname{Var}\left[X\right].$$

D'altra parte

$$\mathbb{E}\left[Y^2\right] = 0\,\mathbb{P}\left(|X-\mu| < t\right) + t^2\,\mathbb{P}\left(|X-\mu| \geq t\right) = t^2\,\mathbb{P}\left(|X-\mu| \geq t\right),$$

da cui la tesi. □

Osservazione 5.2.1. La disuguaglianza di Chebyshev può anche essere formulata nel seguente modo: Se X è una variabile aleatoria con media 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 dalla sua media μ più di una frazione η della deviazione standard σ è inferiore a $\frac{1}{\eta^2}$.

Esempio 5.2.1. Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico di numerosità n. Supponiamo di conoscere la varianza $\sigma^2 = 4$ del campione e che la media μ sia ignota. 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 Chebyshev segue facilmente il seguente

Teorema 5.2.2 (Legge debole dei grandi numeri)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con media μ e varianza σ^2 finite.*

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

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

Dimostrazione. Poiché $\mathbb{E}\left[\overline{X}_n\right] = \mu$ e $\mathrm{Var}\left[\overline{X}_n\right] = \dfrac{\sigma^2}{n}$, per la disuguaglianza di Chebyshev si ha

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > t\right) \leq \frac{\sigma^2}{nt^2} \qquad \forall n \in \mathbb{N}.$$

La tesi segue passando a limite. □

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

Esempio 5.2.2. 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 $X_1,\ X_2,\ \ldots,\ X_n$ 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 Chebyshev

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

Voglio

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

cioè

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

Basta allora avere $\frac{100}{n} \leq \frac{1}{10}$ cioè $n \geq 1000$. 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

Esempio 5.2.3. Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico di numerosità n, bernoulliano di parametro (incognito) $p \in [0,1]$. Dunque

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

Allora, per la disuguaglianza di Chebyshev

$$\mathbb{P}\left(\left|\overline{X} - p\right| > t\right) \leq \frac{p(1-p)}{n\,t^2} \leq \frac{1}{4n\,t^2} \qquad \forall t > 0. \tag{5.1}$$

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

5.2.2. La distribuzione gaussiana $\mathcal{N}\left(\mu, \sigma^2\right)$ e il teorema del limite centrale

Ricordiamo che la distribuzione gaussiana di parametri $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, $\mathcal{N}\left(\mu, \sigma^2\right)$, 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 $\mathcal{N}\left(\mu, \sigma^2\right)$, 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 $F_X(x) := \mathbb{P}(X \leq x)$ 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 $\mathcal{N}((,0),1)$ si dice *distribuzione gaussiana standard*, la funzione di ripartizione associata si indica con la lettera Φ,

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

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

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

Si possono inoltre dimostrare le seguenti proprietà

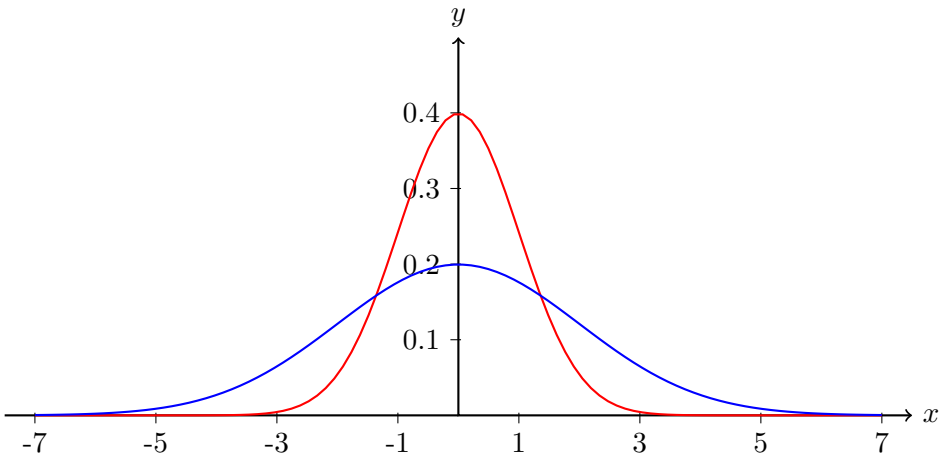

Figura 5.1: Densità associate alle distribuzioni $\mathcal{N}((,0),1)$ (in rosso) e $\mathcal{N}((,0),4)$ (in blu)

Proprietà 5.2.1. 1. Se X è una v.a. gaussiana di media μ e varianza σ^2: $X \sim \mathcal{N}(\mu, \sigma^2)$ e α, β sono due numeri reali, $\alpha \neq 0$, allora la v.a. $\alpha X + \beta$ è gaussiana di media $\alpha\mu + \beta$ e varianza $\alpha^2\sigma^2$: $\alpha X + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$. In particolare $Y := \dfrac{X - \mu}{\sigma}$ è una v.a. gaussiana standard: $Y \sim \mathcal{N}((,0),1)$.

2. Siano X_1, X_2, …, X_n v.a. indipendenti. Supponiamo che ciascuna v.a. X_i sia gaussiana di media μ_i e varianza σ_i^2: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ $\forall i = 1, \ldots, n$. Allora la v.a. $X_1 + X_2 + \ldots + X_n$ è gaussiana di media pari alla somma delle medie e varianza pari alla somma delle varianze:

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Teorema 5.2.3 (Teorema del limite centrale)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con media μ e varianza σ^2 finite. Sia $\Phi(t)$ la funzione di ripartizione associata alla distribuzione gaussiana standard $\mathcal{N}(0,1)$.*

Per ogni $n \in N$ sia $\overline{X}_n$ la media campionaria di X_1, X_2, …, X_n e sia $\overline{Z}_n$ la sua standardizzazione:

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

Allora

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

Osservazione 5.2.2. Una formulazione equivalente della tesi del teorema del limite centrale è

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

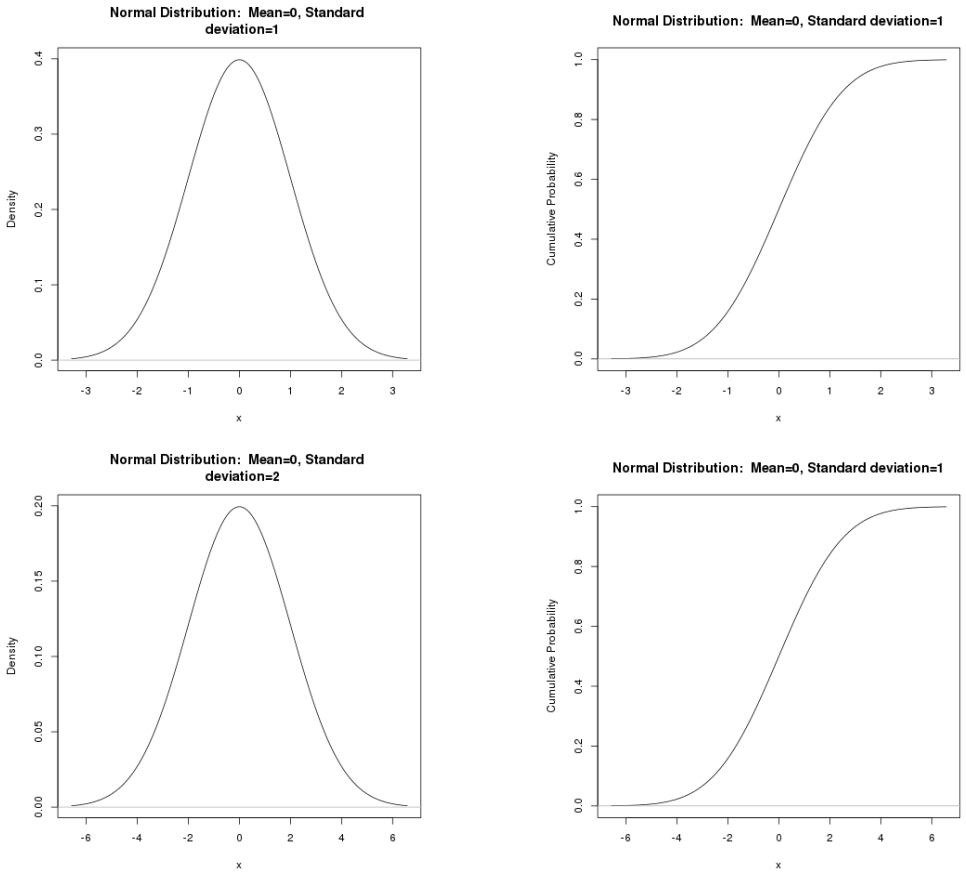

Figura 5.2: $\mathcal{N}((,0),1)$ e $\mathcal{N}((,0),4)$, densità e funzione di ripartizione

Esempio 5.2.4. Supponiamo di avere un campione statistico di numerosità 25 e deviazione standard 8. Qual è la probabilità che la media campionaria differisca dalla media 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 Chebyshev otteniamo

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

Proviamo ad applicare il teorema del limite centrale. Indico con $\overline{Z}$ la standardizzazione

della media campionaria. Si ha

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

Perché questa stima *sembra* tanto migliore di quella ottenuta con la disuguaglianza di Chebyshev? Perché non abbiamo un'indicazione sul significato di quei $\simeq$. 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 $\mathbb{P}\left(Z_n \leq t\right)$ con $\Phi(t)$. A tal proposito vale il seguente

Teorema 5.2.4 (Teorema di Berry–Esseen)**.** *Sia $\{X_i\}_{i=1}^{\infty}$ una successione di v.a. indipendenti, identicamente distribuite, con media $\mu = 0$, 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 $\mathcal{N}\left((,0),1\right)$.*

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 5.2.4, otteniamo dunque

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

5.3. Alcune distribuzioni legate alla distribuzione gaussiana

5.3.1. Distribuzione di Pearson (o χ^2) con n gradi di libertà, χ^2_n

Si chiama così la distribuzione associata alla densità

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

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

Osservazione 5.3.1. Si può dimostrare che $\forall a > 0$ si ha $\Gamma(a+1) = a\Gamma(a)$ e che $\Gamma(1) = 1$, $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$. Quindi

$\Gamma(2) = 1{\cdot}1$, $\quad\Gamma(3) = 2\,\Gamma(2) = 2{\cdot}1 = 2!\quad\ldots\quad\Gamma(n) = (n-1)!$ per ogni intero positivo n

mentre

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\sqrt{\pi}, \qquad \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3\cdot 1}{2\cdot 2}\sqrt{\pi} = \frac{3!!}{2^2}\sqrt{\pi},$$
$$\ldots \qquad \Gamma\left(\frac{2k+1}{2}\right) = \frac{(2k-1)!!}{2^k}\sqrt{\pi} \quad \text{per ogni intero non-negativo } k.$$

Si possono calcolare media e varianza di una v.a. che segua una distribuzione di Pearson:

Proprietà 5.3.1. Se X è una v.a. con distribuzione χ^2 a n gradi di libertà, $X \sim \chi^2_n$, allora

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

Proprietà 5.3.2. Se X e Y sono due variabili di Pearson indipendenti, $X \sim \chi^2_n, Y \sim \chi^2_k$, allora la v.a. $X + Y$ segue la distribuzione di Pearson a $n + k$ gradi di libertà:

$$X + Y \sim \chi^2_{n+k}.$$

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

Teorema 5.3.1. *Se $X_1,\ X_2,\ \ldots,\ X_n$ sono v.a. indipendenti e gaussiane, con X_i di media μ_i e varianza σ_i^2, $\forall i = 1, \ldots, n$, allora la v.a. $\displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$ segue la distribuzione di Pearson a n gradi di libertà, χ^2_n.*

Corollario 5.3.2. *Se $X_1,\ X_2,\ \ldots,\ X_n$ è un campione statistico gaussiano, con media μ e varianza σ^2, allora la v.a. $\chi^2 := \displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ segue una distribuzione χ^2 con n gradi di libertà.*

Esempio 5.3.1. 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 media 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 $Z_i := \dfrac{X_i}{\sigma} = \dfrac{X_i}{2}$, $i = 1, 2, 3$, 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^2_3 nel punto 0.36 che è (circa) 0.052.

```
> setwd("/home/laura/Documents/didattica/2012-13_elaborazioni_B194")

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, dchisq(.x, df=3), xlab="x", ylab="Density",
 main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")

> abline(h=0, col="gray")

> remove(.x)

> dev.copy(png,'densitachiquadro3.png');dev.off()
png
  3
X11cairo
       2

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, pchisq(.x, df=3), xlab="x", ylab="Cumulative Probability",
 main=paste("ChiSquared Distribution:  Degrees of freedom=3"), type="l")

> abline(h=0, col="gray")

> remove(.x)

> pchisq(c(0.36), df=3, lower.tail=TRUE)
[1] 0.05162424

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, dchisq(.x, df=3), xlab="x", ylab="Density",
 main=paste("ChiSquared   Distribution:  Degrees of freedom=3"),type="l")

> abline(h=0, col="gray")

> remove(.x)

> dev.copy(png,'densitachiquadro3.png');dev.off()
png
  3
```

```
X11cairo
       2

> .x <- seq(0.015, 17.73, length.out=100)

> plot(.x, pchisq(.x, df=3), xlab="x", ylab="Cumulative Probability",
 main=paste("ChiSquared Distribution:  Degrees of freedom=3"),type="l")

> abline(h=0, col="gray")

> remove(.x)

> dev.copy(png,'ripartizionechiquadro3.png');dev.off()
png
  3
X11cairo
       2

> pchisq(c(0.36), df=3, lower.tail=TRUE)
[1] 0.05162424
```

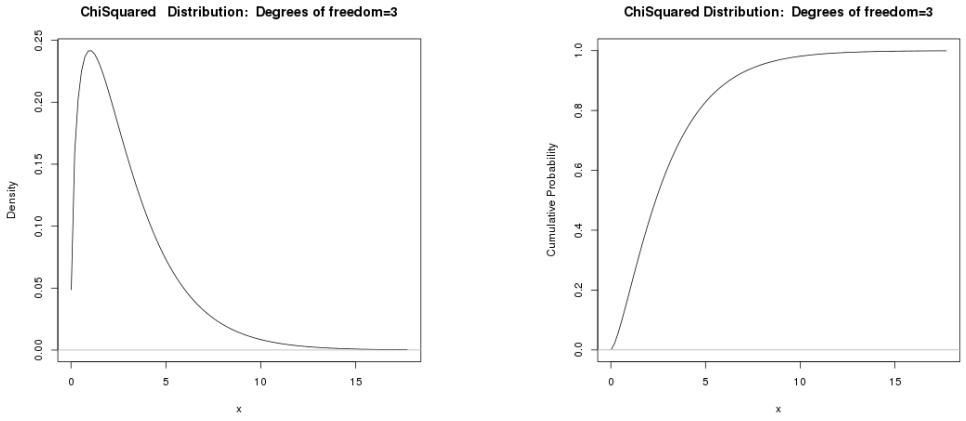

Figura 5.3: χ^2_3, densità e funzione di ripartizione

A titolo di confronto, visualizziamo anche densità e funzione di ripartizione delle distribuzioni χ^2_{10} e χ^2_{100}. Il seguente teorema raccoglie alcune importanti proprietà dei campioni statistici gaussiani e delle loro media e varianza campionarie.

Teorema 5.3.3. *Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico gaussiano di numerosità n, media μ e varianza σ^2.*

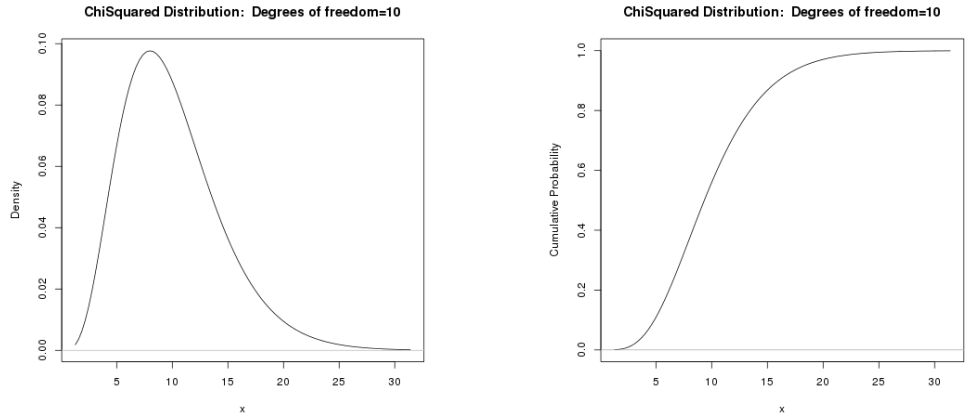

Figura 5.4: χ^2_{10}, densità e funzione di ripartizione

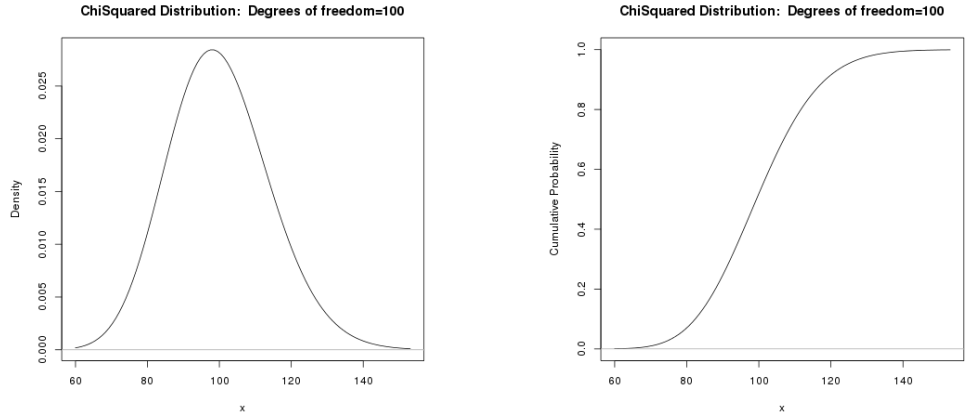

Figura 5.5: χ^2_{100}, densità e funzione di ripartizione

Allora, la media campionaria $\overline{X} := \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *e la varianza campionaria* $S^2 := \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *sono v.a. indipendenti.*

Sia $Z_1,\ Z_2,\ \ldots Z_n$ *la standardizzazione del campione statistico* $X_1,\ X_2,\ \ldots,\ X_n$ *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* $Z_1,\ Z_2,\ \ldots,\ Z_n$.

Allora $\overline{Z} = \dfrac{\overline{X} - \mu}{\sigma}$ *e la v.a.* $\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2$ *segue una distribuzione* χ^2 *con* $n-1$ *gradi di libertà.*

Corollario 5.3.4. *Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico gaussiano di numerosità n, media 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

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

□

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

Si chiama così la distribuzione associata alla densità

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

Proprietà 5.3.3. Se X è una v.a. con distribuzione t di Student a n gradi di libertà, allora

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

Osservazione 5.3.2. 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 $X \sim t(n)$, 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 5.3.5. *w Se Z è una v.a. gaussiana standard, $Z \sim \mathcal{N}(0,1)$, se Y segue la distribuzione χ^2 con n gradi di libertà, $Y \sim \chi^2_n$ 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:*

$$T := \frac{Z\sqrt{n}}{\sqrt{Y}} \sim t(n).$$

Corollario 5.3.6. *Se $X_1,\ X_2,\ \ldots,\ X_n$ è un campione statistico gaussiano di numerosità n, media 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:

$$T \sim t(n-1).$$

Dimostrazione. Basta applicare il teorema 5.3.5 con $Z = \dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ e $Y = (n-1)\dfrac{S^2}{\sigma^2}$. □

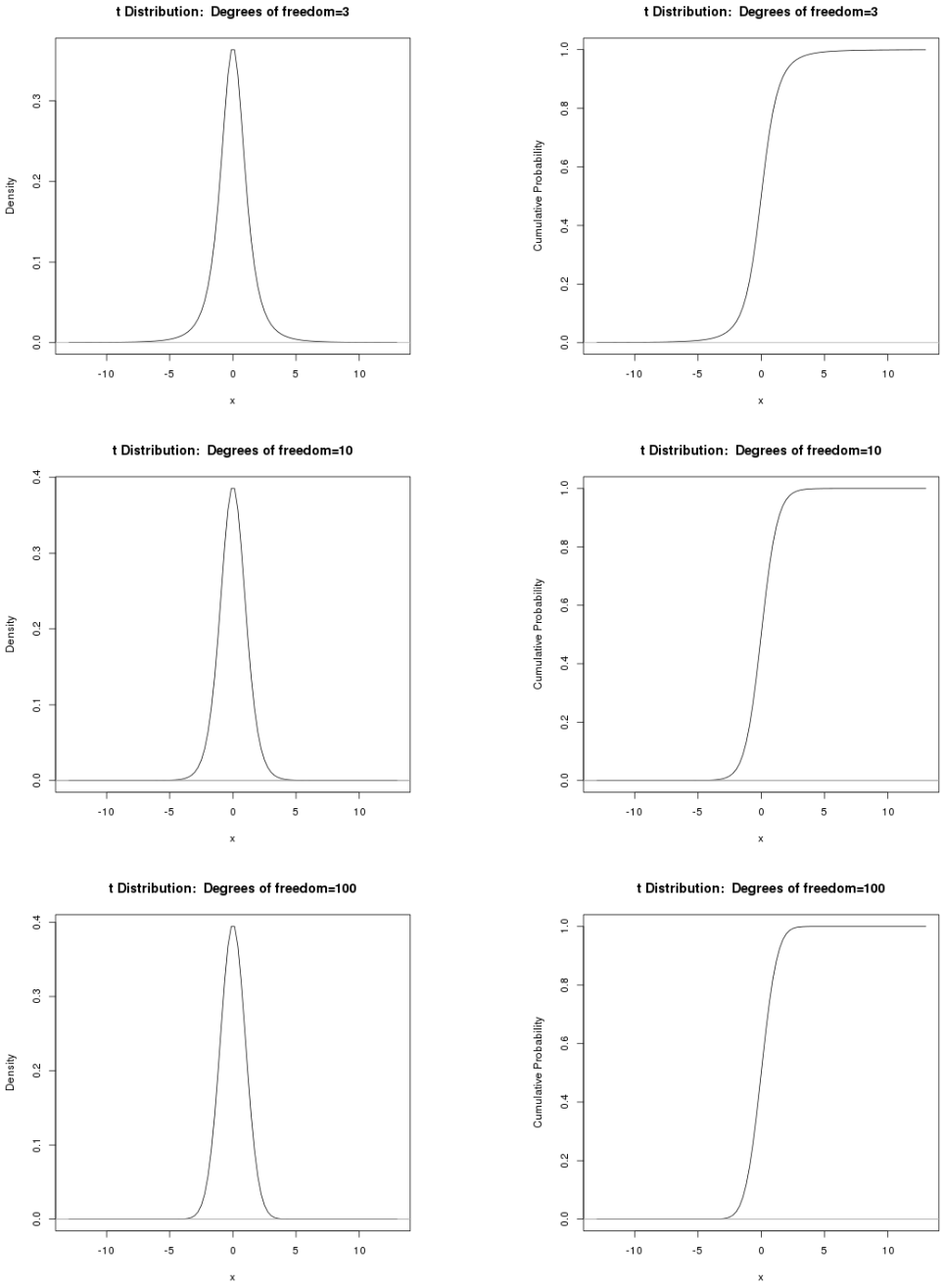

Figura 5.6: $t(3)$, $t(10)$, $t(100)$, densità e funzione di ripartizione