METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 4: MODELLI DI VARIABILI ALEATORIE, INDICI DI FORMA

A.A. 2024-2025

Prof. Martina Vettoretti

VARIABILI ALEATORIE IID

Le variabili aleatorie X_1, X_2, …, X_n si dicono **indipendenti ed identicamente distribuite (iid)** se sono tra loro indipendenti, ovvero:

$$P(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

$$\forall\, A_1 \subset \mathbb{R}, A_2 \subset \mathbb{R}, \dots, A_n \subset \mathbb{R}$$

$$F_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot \ \dots \cdot F_{X_n}(x_n)$$

e hanno la stessa distribuzione, ovvero:

$$F_{X_1}(x) = F_{X_k}(x) \qquad \forall k \in \{1, \dots, n\}\ \forall x$$

Un esempio di variabili iid sono n variabili che rappresentano una quantità osservata in n esperimenti ripetuti nelle stesse condizioni.

MODELLI DI VARIABILI ALEATORIE

Alcuni tipi di variabili aleatorie risultano spesso adatti per descrivere diverse grandezze che incontriamo comunemente negli esperimenti.

- Variabile aleatoria di Bernoulli
- Variabile aleatoria binomiale
- Variabile aleatoria di Poisson

} Variabili aleatorie discrete

- Variabile aleatoria uniforme
- Variabile aleatoria esponenziale
- Variabile aleatoria di tipo gamma
- Variabile aleatoria di Weibull
- Variabile aleatoria normale o gaussiana
- Variabili aleatorie che derivano dalla normale (chi-quadro, t, F)

} Variabili aleatorie continue

VARIABILE ALEATORIA DI BERNOULLI

La variabile aleatoria di Bernoulli è una variabile aleatoria X che può assumere solo **due valori: 1** con probabilità p; **0** con probabilità 1-p.

$X \sim \mathcal{B}(p)$

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$
$$0 \leq p \leq 1$$

$\Rightarrow$

- $E[X] = p$
- $Var(X) = p(1 - p)$

La variabile aleatoria di Bernoulli è utilizzata per rappresentare l'esito di una prova che può avere solo due esiti: «**successo**» (X=1) o «**fallimento**» (X=0).

- Esempio. Un'azienda produce componenti elettronici che possono essere funzionanti con probabilità 0.95 o difettosi con probabilità 0.05. L'esito funzionante/difettoso può essere descritto con una variabile aleatoria di Bernoulli di parametro p=0.95.

VARIABILE ALEATORIA BINOMIALE (1/2)

Supponiamo di realizzare **n ripetizioni indipendenti** di un esperimento, ciascuna che può concludersi in un «**successo**» con probabilità **p**, o in un «**fallimento**» con probabilità 1-p. Sia X il numero totale di successi nelle n prove.

→ X si dice **variabile aleatoria binomiale di parametri (n,p)** $X \sim \mathcal{B}(n, p)$

Funzione di massa:

$$P(X = k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}, k = 0, 1, \dots, n$$

- $E[X] = n \cdot p$
- $Var(X) = n \cdot p \cdot (1-p)$

$$\binom{n}{k} := \frac{n!}{k!\,(n-k)!}$$

Il **coefficiente binomiale** n su k rappresenta il numero di combinazioni diverse che possiamo ricavare scegliendo k elementi da un insieme di n oggetti.

VARIABILE ALEATORIA BINOMIALE (2/2)

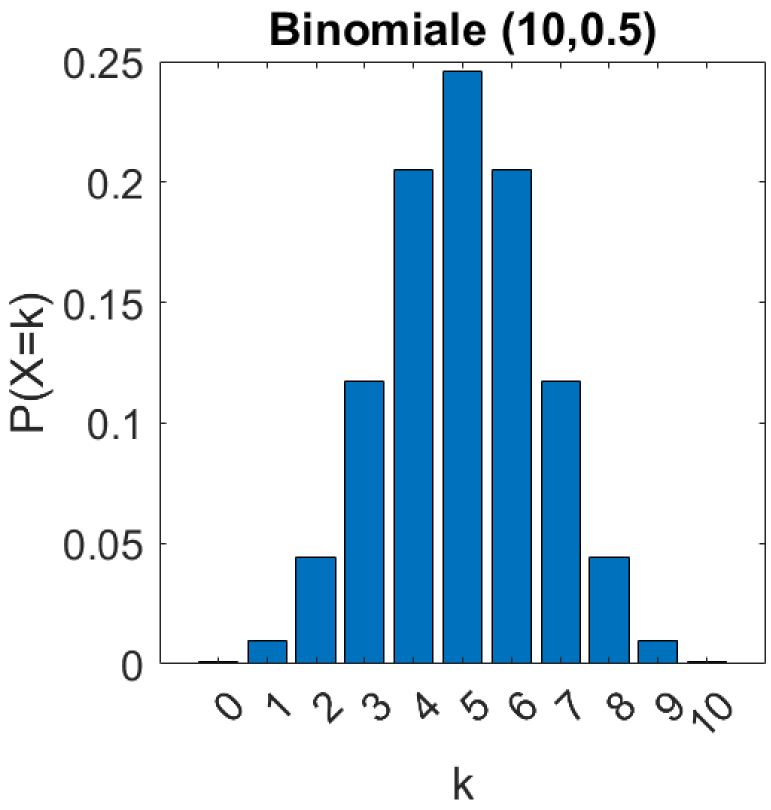
- Una variabile aleatoria binomiale X di parametri (n,p) si può scrivere come **somma di n variabili aleatorie bernoulliane iid** X_i di parametro p.

$$X = \sum_{i=1}^{n} X_i \qquad X_i \sim \mathcal{B}(p) \qquad X \sim \mathcal{B}(n,p)$$

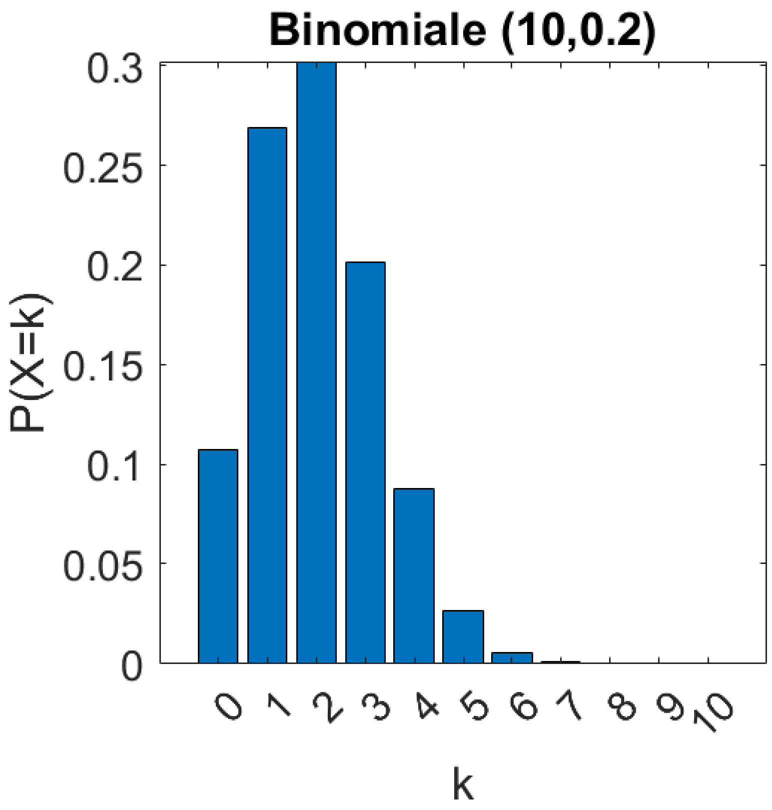
$$X_i = \begin{cases} 1 & \text{se la prova i–esima ha successo} \\ 0 & \text{altrimenti} \end{cases}$$

- Esempio. Un'azienda produce componenti elettronici che possono essere funzionanti con probabilità 0.95 o difettosi con probabilità 0.05. Questi componenti sono venduti in lotti ciascuno contenente 50 pezzi. → Il numero di componenti funzionanti in ciascun lotto è una variabile aleatoria binomiale di parametri (n,p)=(50,0.95).

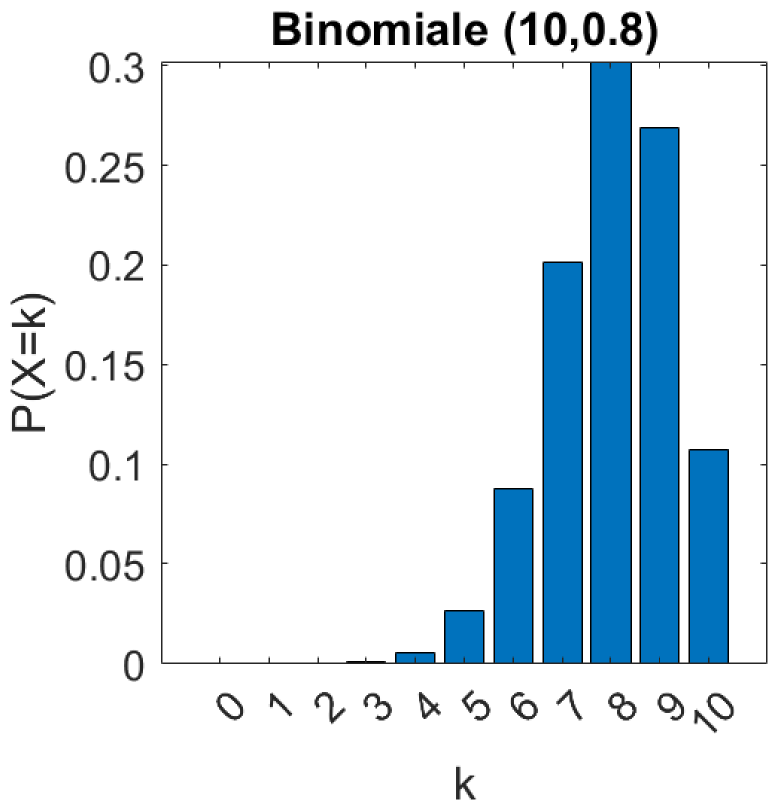
ESEMPI DI FUNZIONI DI MASSA BINOMIALI



Simmetrica. Con p=0.5 è più probabile che in n prove ripetute il numero di successi sia simile al numero di insuccessi.



Asimmetrica con coda verso destra. Con p piccolo è più probabile che in n prove ripetute il numero di successi sia << n.



Asimmetrica con coda verso sinistra. Con p grande è più probabile che in n prove ripetute il numero di successi sia vicino ad n.

VARIABILE ALEATORIA DI POISSON

Una variabile aleatoria X si dice di Poisson con parametro λ, λ>0, se assume i valori 0, 1, 2, …, e la sua funzione di massa di probabilità è data da:

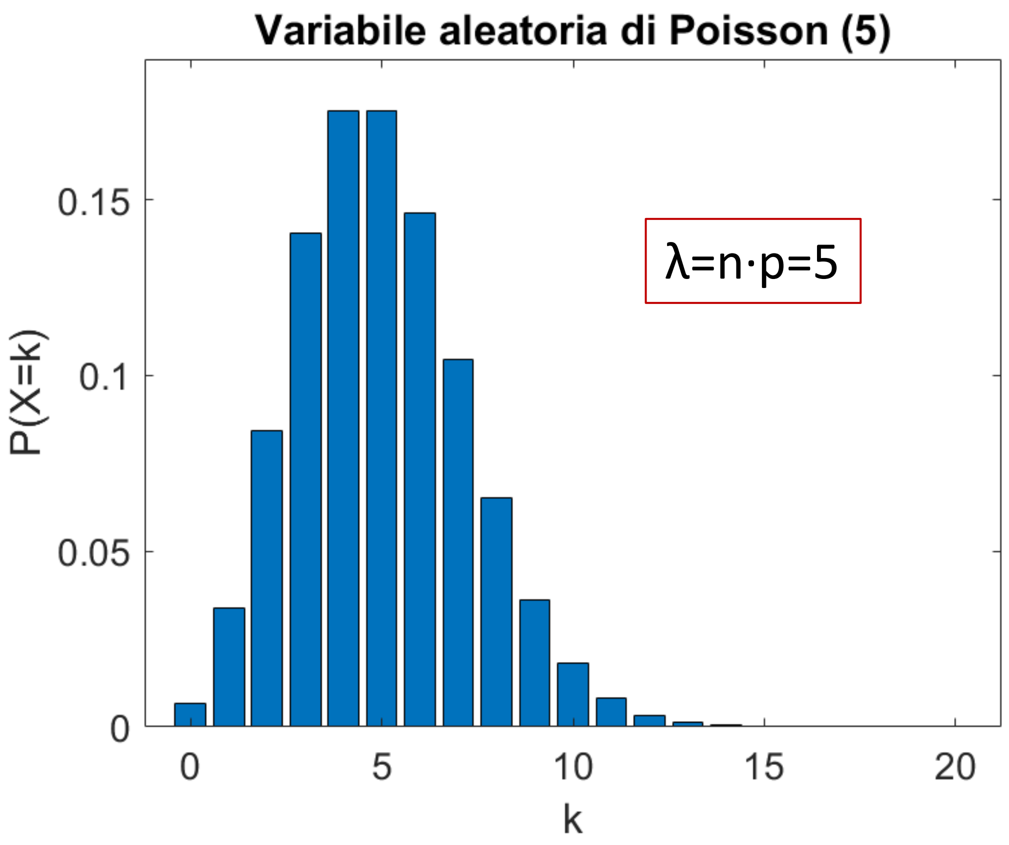
$$X \sim P(\lambda) \qquad P(X=k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \qquad k = 0, 1, 2, \ldots$$

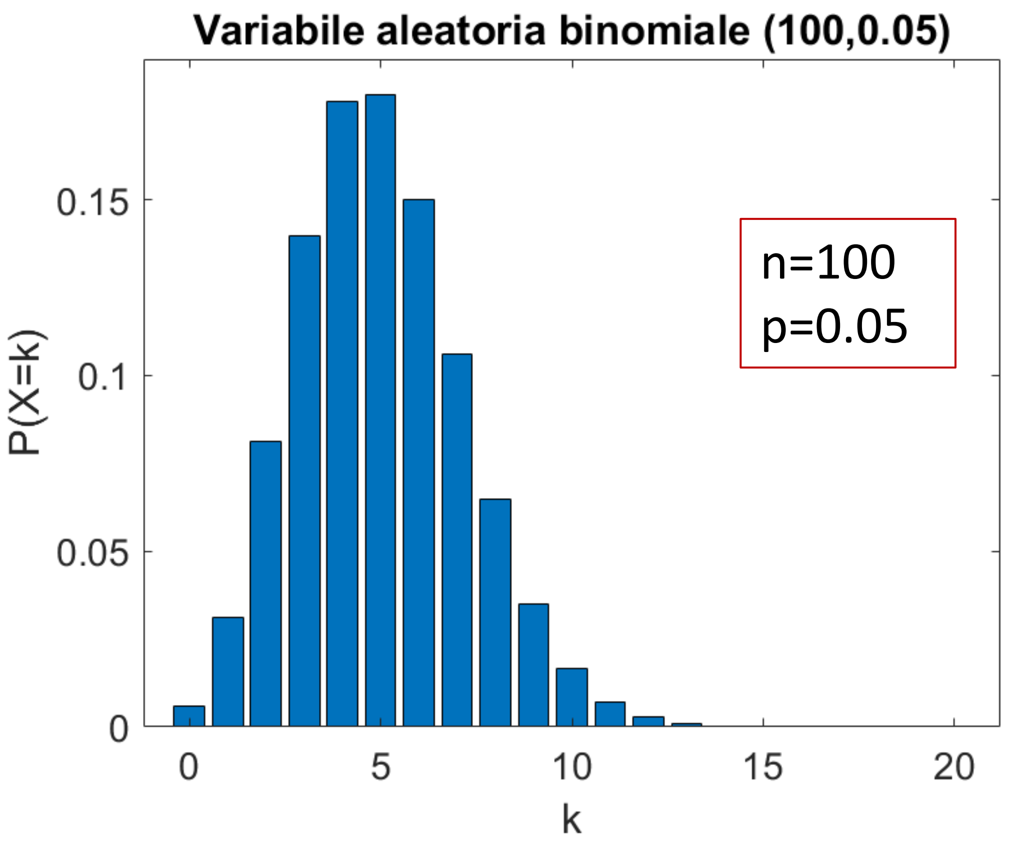
Si può dimostrare che: E[X] = λ, Var[X] = λ.

La variabile aleatoria di Poisson di parametro λ=n·p viene usata come approssimazione di una binomiale di parametri (n,p) quando n è grande e p è piccolo.

Esempio. Il numero di individui, in una certa popolazione molto ampia, che raggiunge i 100 anni di età si potrebbe rappresentare con una variabile aleatoria di Poisson.

POISSON VS. BINOMIALE



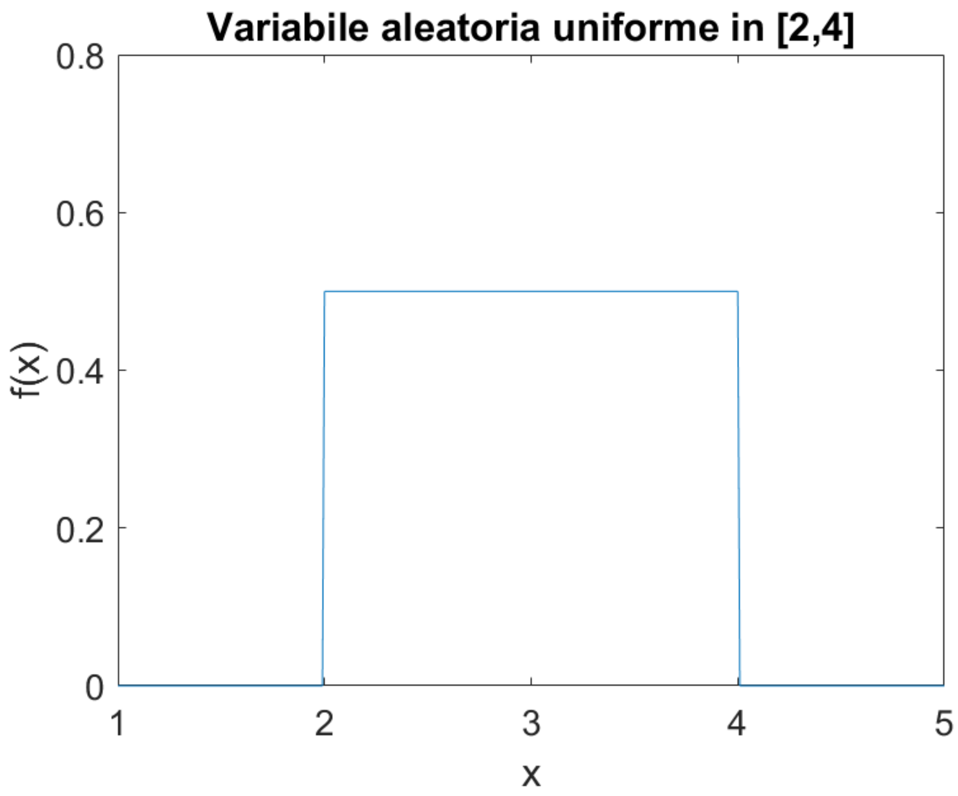


VARIABILE ALEATORIA CONTINUA UNIFORME

Una variabile aleatoria continua X si dice uniforme sull'intervallo [a, b] se la sua funzione densità di probabilità è:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & se\ a \leq x \leq b \\ 0 & altrimenti \end{cases}$$

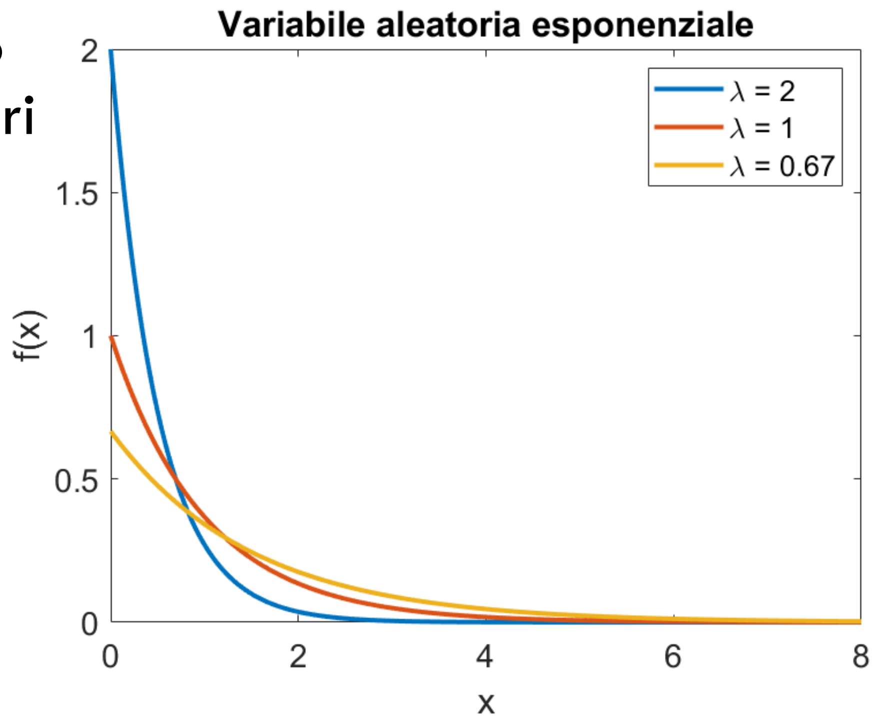
$$X \sim U(a, b)$$



- $E[X] = \frac{a+b}{2}$
- $Var[X] = \frac{(b-a)^2}{12}$

VARIABILE ALEATORIA ESPONENZIALE

Una variabile aleatoria X si dice esponenziale di parametro λ, con λ>0, se la sua funzione di densità di probabilità è pari a:



$$X \sim \mathcal{E}(\lambda) \qquad f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & se\ x \geq 0 \\ 0 & se\ x < 0 \end{cases}$$

- $E[X] = \frac{1}{\lambda}$
- $Var[X] = \frac{1}{\lambda^2}$

Questa variabile aleatoria viene generalmente impiegata per rappresentare il tempo di attesa prima che si verifichi un certo evento casuale (es. il tempo prima che un certo componente si guasti), sotto l'ipotesi che la «probabilità istantanea» dell'evento sia costante nel tempo (nell'esempio citato, non tiene conto dell'usura dei materiali del componente).

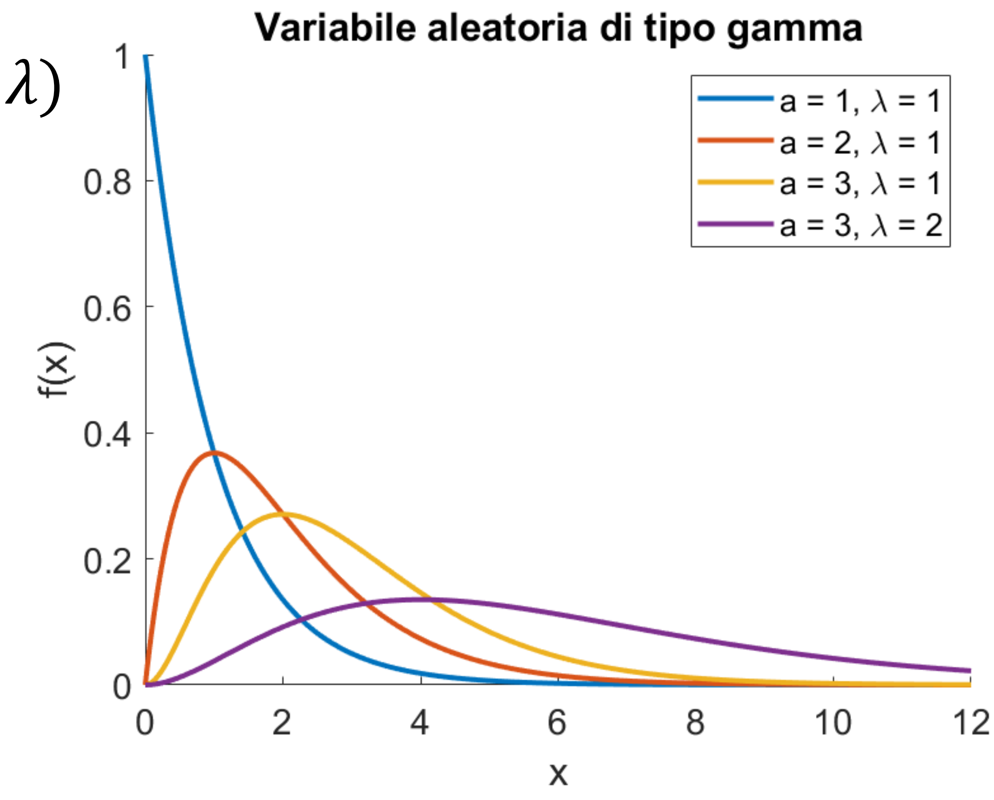
Una variabile aleatoria è di tipo gamma di parametri (a, λ), con a>0 e λ>0, se la sua funzione di densità di probabilità è data da:

$$f(x) = \begin{cases} \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} & se\ x > 0 \\ 0 & se\ x \leq 0 \end{cases}, \quad X \sim Gamma(a, \lambda)$$

$\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy$ → Funzione Gamma di Eulero

- $E[X] = \frac{a}{\lambda}$
- $Var[X] = \frac{a}{\lambda^2}$

Nota: con a=1, la variabile aleatoria di tipo gamma è equivalente ad una variabile aleatoria esponenziale di parametro λ.



VARIABILE ALEATORIA DI TIPO GAMMA (2/2)

La somma di n variabili aleatorie esponenziali di parametro λ è una variabile aleatoria di tipo gamma di parametri (n, λ).

Esempio. Se il tempo di vita di una batteria è descritto da una variabile aleatoria esponenziale di parametro λ, allora volendo far funzionare un dispositivo che richiede una sola batteria, e avendo a disposizione n batterie, il tempo totale di funzionamento del dispositivo potrà essere descritto con una variabile aleatoria gamma di parametri n e λ.

VARIABILE ALEATORIA DI WEIBULL (1/2)

Una variabile aleatoria di Weibull di parametri γ e λ, con γ>0 e λ>0, è una variabile aleatoria X avente la seguente funzione di densità di probabilità:

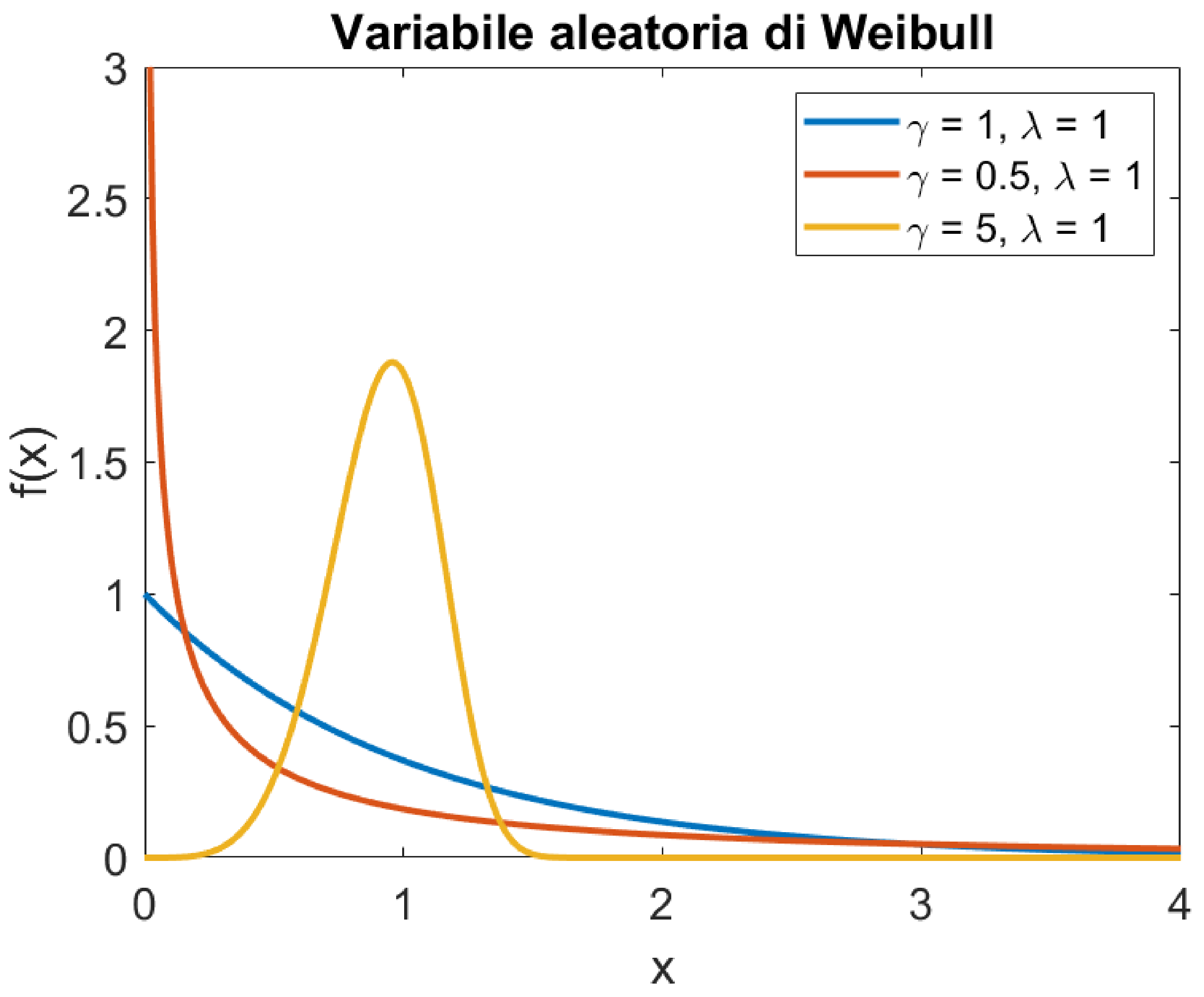
$$f(x) = \begin{cases} \gamma \cdot \lambda \cdot (\lambda \cdot x)^{\gamma-1} \cdot e^{-(\lambda \cdot x)^{\gamma}} & se\ x \geq 0 \\ 0 & se\ x < 0 \end{cases}$$

γ = parametro di forma

λ = parametro di scala

- La variabile aleatoria di Weibull si usa tipicamente per rappresentare il tempo ad un guasto per casi in cui la «probabilità istantanea» di guasto varia nel tempo (aumenta nel tempo con γ>1, diminuisce nel tempo con γ<1, è costante nel tempo con γ=1).
- Con γ=1, la variabile aleatoria di Weibull è equivalente ad una variabile aleatoria esponenziale di parametro λ.

VARIABILE ALEATORIA DI WEIBULL (2/2)



VARIABILE ALEATORIA NORMALE O GAUSSIANA

Una variabile aleatoria X si dice normale o gaussiana di parametri μ e σ^2 se X ha funzione di densità di probabilità data da:

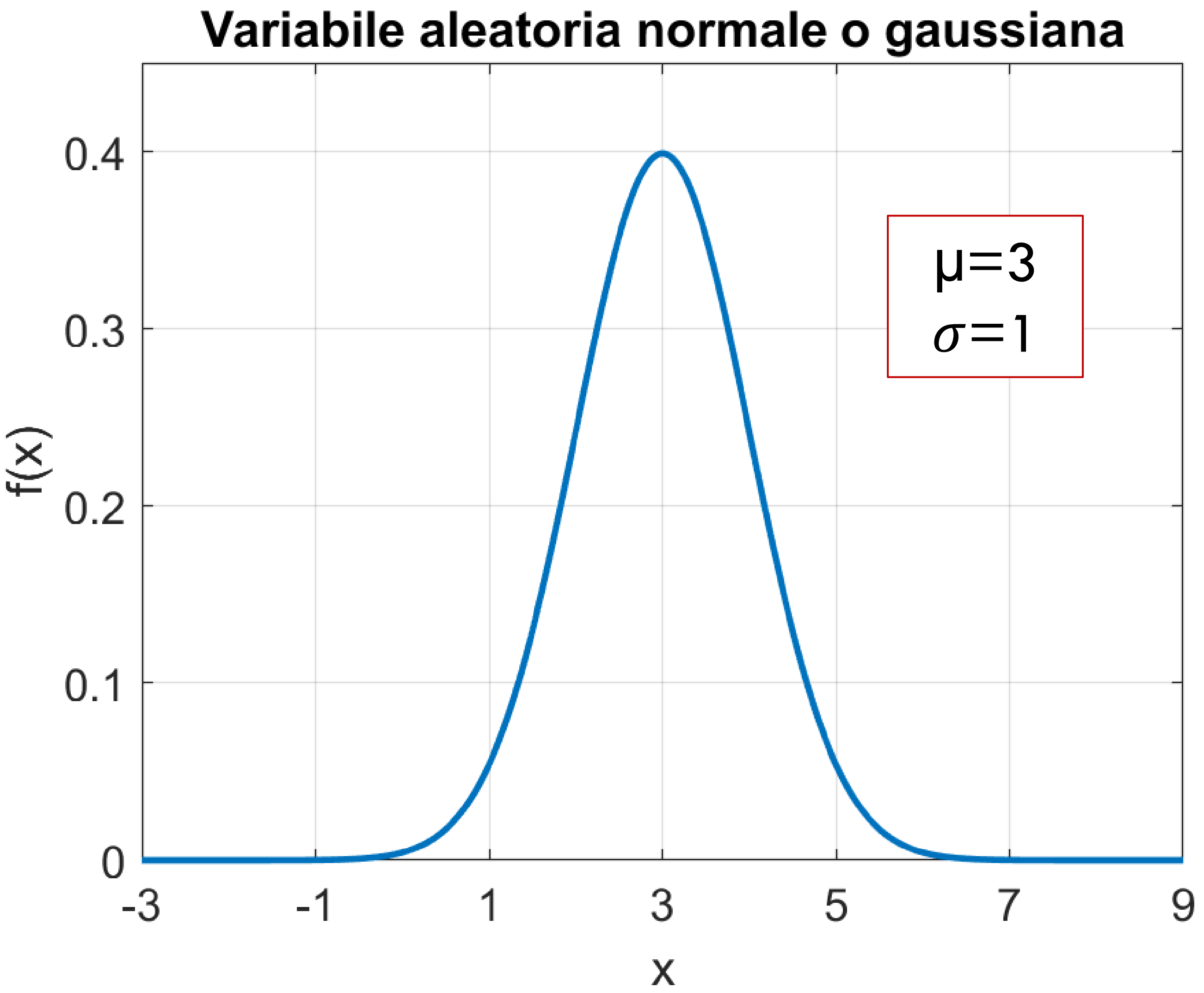
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}$$

- $E[X] = \mu$
- $Var(X) = \sigma^2$

Per indicare che X è normale con parametri μ e σ^2 scriviamo $X \sim N(\mu, \sigma^2)$

ESEMPIO DI DENSITA' DI PROBABILITA' NORMALE

- La densità di una variabile aleatoria normale è una curva a campana simmetrica rispetto all'asse x= μ.
- La ddp normale è unimodale, cioè presenta un solo picco.
- Il massimo della ddp normale vale $(\sigma\sqrt{2\pi})^{-1} \approx 0.399/\sigma$.



TRASFORMAZIONE LINEARE DI UNA VARIABILE ALEATORIA NORMALE

La trasformazione lineare di una variabile aleatoria normale è anch'essa una variabile aleatoria normale.

$$\left.\begin{array}{l} \blacktriangleright\ X \sim N(\mu, \sigma^2) \\ \blacktriangleright\ Y = a \cdot X + b \end{array}\right\} \Rightarrow \quad Y \sim N(a \cdot \mu + b, a^2 \cdot \sigma^2)$$

NORMALE STANDARD

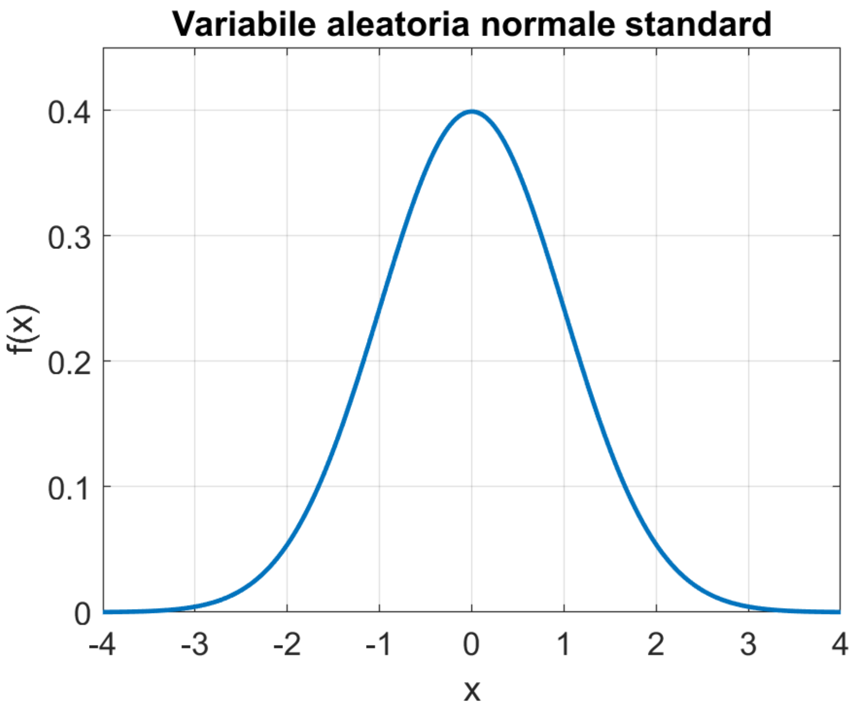
Se standardizziamo una variabile aleatoria normale, X, con valore atteso μ e deviazione standard σ, otteniamo una variabile aleatoria normale, Z, con valore atteso 0 e deviazione standard 1 che viene detta **normale standard**:

$$Z := \frac{X - \mu}{\sigma}$$

- $X \sim N(\mu, \sigma^2)$
- $Z \sim N(0, 1)$

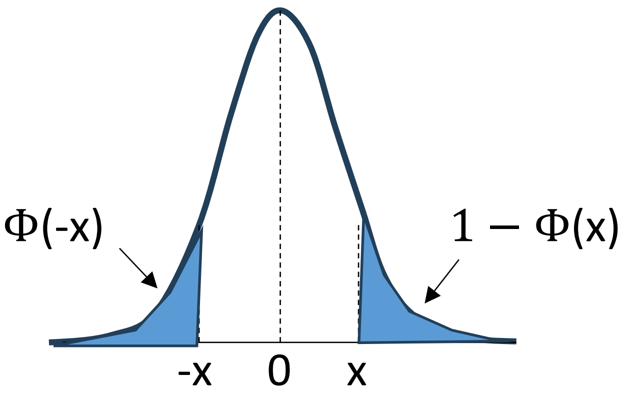
La funzione di ripartizione di una normale standard si indica con $\Phi(x)$ ed è:

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy \qquad \forall x \in \mathbb{R}$$



PROBABILITA' DI UNA VARIABILE ALEATORIA NORMALE

- Possiamo esprimere le probabilità di una variabile aleatoria normale, X, in termini delle probabilità di una variabile aleatoria normale standard, Z.
  - $P(X < d) = P\left(Z < \frac{d-\mu}{\sigma}\right) = \Phi\left(\frac{d-\mu}{\sigma}\right)$
  - $P(c < X < d) = P\left(Z < \frac{d-\mu}{\sigma}\right) - P\left(Z < \frac{c-\mu}{\sigma}\right) = \Phi\left(\frac{d-\mu}{\sigma}\right) - \Phi\left(\frac{c-\mu}{\sigma}\right)$
- Operativamente come calcoliamo queste probabilità? L'integrale che definisce Φ non presenta una soluzione analitica. Tuttavia Φ(x) si può calcolare usando delle approssimazioni.
  - Tabelle che riportano i valori approssimati di Φ(x)
  - Funzioni al calcolatore (in Matlab: normcdf)
- Nota: grazie alla simmetria, $\Phi(-x) = 1 - \Phi(x)$



LA PROBABILITA' DI ALCUNI UTILI INTERVALLI

Data una variabile aleatoria normale X con media μ e deviazione standard σ qualsiasi è possibile dimostrare che:

- $P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$ → Circa il 68% dei valori di una variabile aleatoria normale stanno nell'intervallo media $\pm$ deviazione standard
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$ → Circa il 95% dei valori di una variabile aleatoria normale stanno nell'intervallo media $\pm$ 2 deviazioni standard
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997$ → Circa il 99.7% dei valori di una variabile aleatoria normale stanno nell'intervallo media $\pm$ 3 deviazioni standard

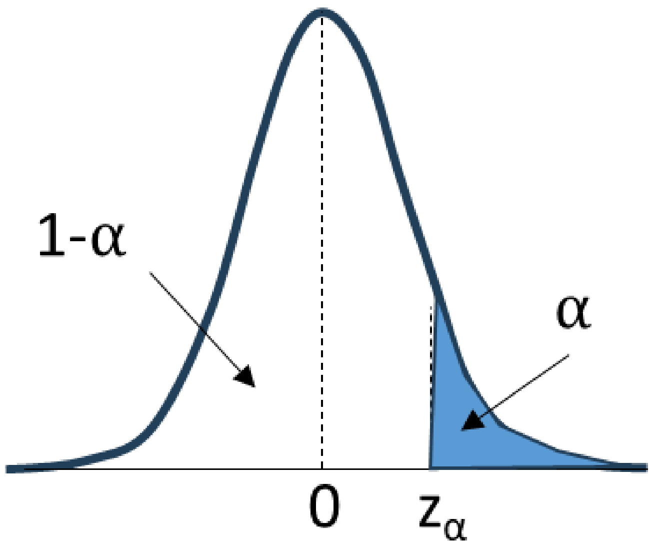
QUANTILI GAUSSIANI

Per ogni α in [0, 1], definiamo il quantile α della normale standard, Z, come la quantità:

$$z_\alpha := \Phi^{-1}(1-\alpha)$$

Ovvero quel valore z_α tale che:

$$P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = \alpha$$



$z_{0.05} \approx 1.645$

$z_{0.025} \approx 1.96$

$z_{0.01} \approx 2.33$

VARIABILE ALEATORIA CHI-QUADRO

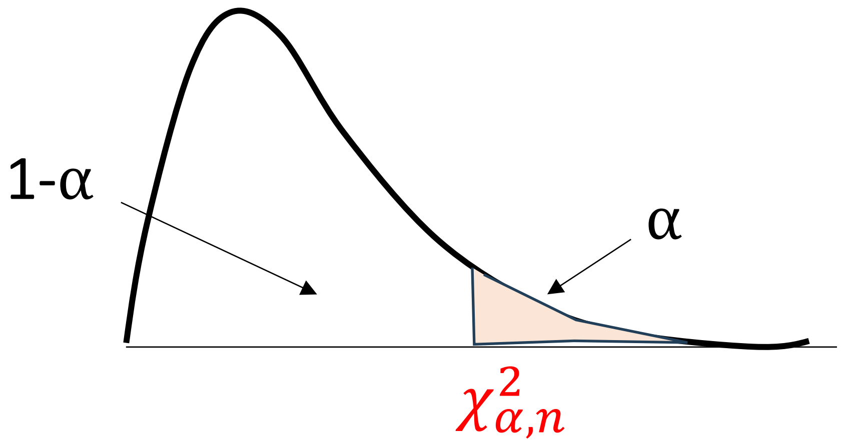
Se $Z_1, Z_2, \ldots, Z_n$ sono variabili aleatorie normali standard indipendenti, allora la somma dei loro quadrati, X, è una variabile aleatoria **chi-quadro ad n gradi di libertà:**

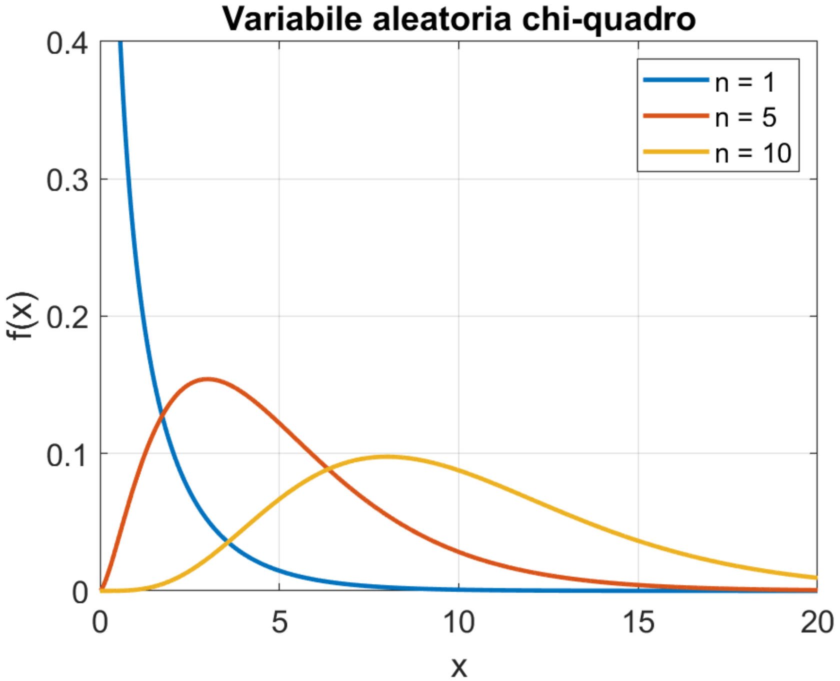
$$X = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_n}^2$$

$$X \sim {\chi_n}^2$$

Quantili:

$\chi^2_{\alpha,n}$: $\quad P\left(X > \chi^2_{\alpha,n}\right) = \alpha$





VARIABILE ALEATORIA t DI STUDENT (1/2)

Se Z è una variabile aleatoria normale standard e C_n è una variabile aleatoria chi-quadro con n gradi di libertà, e se Z e C_n sono indipendenti, allora la variabile aleatoria T_n definita come:

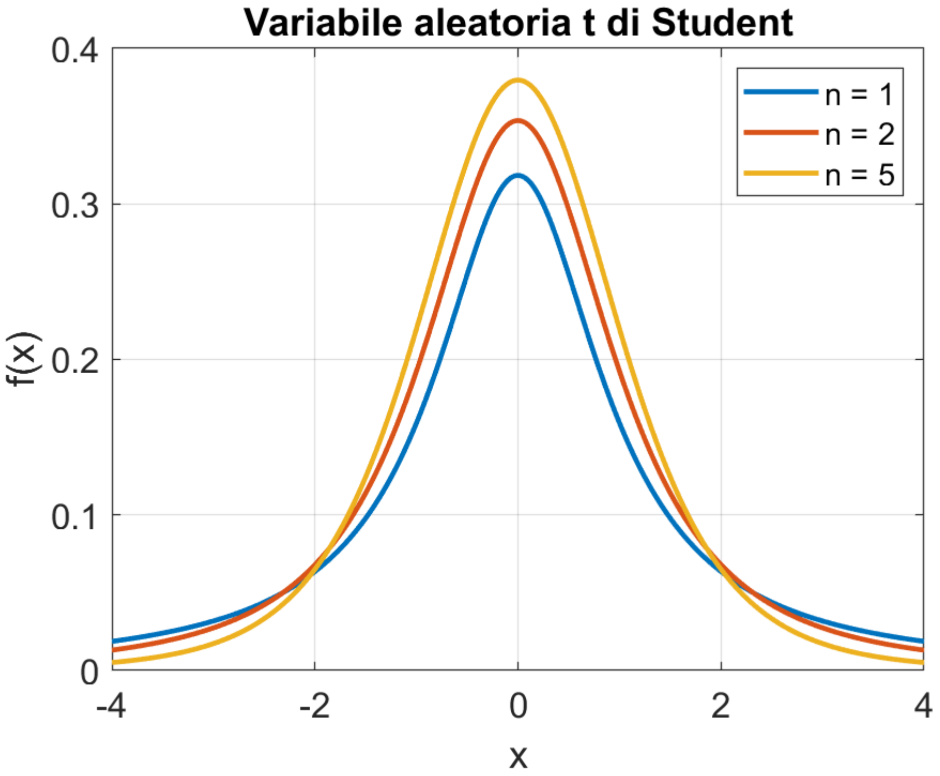
$$T_n := \frac{Z}{\sqrt{\frac{C_n}{n}}}$$

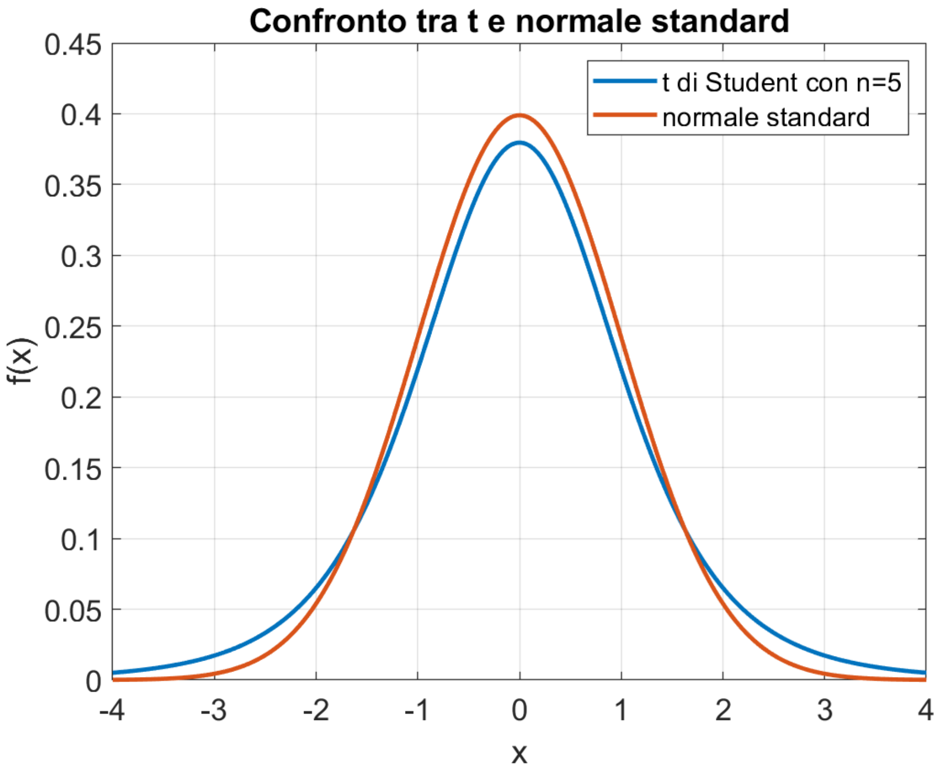
è una variabile aleatoria **t, o t di Student, con n gradi di libertà,** e si indica con:

$$T_n \sim t_n$$

- $E[T_n] = 0$ per n $\geq$ 2, indefinito altrimenti
- $Var(T_n) = \frac{n}{n-2}$ se n ≥ 3, infinita altrimenti

VARIABILE ALEATORIA T DI STUDENT (2/2)





La ddp della t di Student ha forma a campana ed è simmetrica rispetto all'asse x=0. Con n=5, la forma è simile a quella della normale standard, con la t di Student caratterizzata da code più spesse, ad indicare una maggiore variabilità rispetto alla normale standard.

Anche per la t di Student possiamo definire i quantili: $t_{\alpha,n}$: $\quad P\left(T_n > t_{\alpha,n}\right) = \alpha$

VARIABILE ALEATORIA F

Se C_n e C_m sono variabili aleatorie di tipo chi-quadro con n ed m gradi di libertà, rispettivamente, allora la variabile aleatoria $F_{n,m}$ definita da:

$$F_{n,m} = \frac{C_n/n}{C_m/m}$$

È una **variabile aleatoria di tipo F, o F di Fisher, con n ed m gradi di libertà.**

Anche per la F di Fisher possiamo definire i quantili:

$$F_{\alpha,n,m}: \quad P\left(F_{n,m} > F_{\alpha,n,m}\right) = \alpha$$

TEOREMA DEL LIMITE CENTRALE

Siano X_1, X_2, …, X_n delle variabili aleatorie iid, con distribuzione qualsiasi ma uguale tra loro, tutte con media μ e deviazione standard σ^2. Allora se n è grande, la variabile aleatoria Y definita come:

$$Y = X_1 + X_2 + \cdots + X_n$$

converge in distribuzione ad una variabile aleatoria normale con media $n \cdot \mu$ e varianza $n \cdot \sigma^2$.

→ Y si può approssimare con una normale con media $n\mu$ e varianza $n\sigma^2$

Se standardizziamo, la variabile aleatoria standardizzata Z:

$$Z = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

per n grande converge in distribuzione ad una normale standard.

CAMPIONE ALEATORIO

Nella **statistica descrittiva** abbiamo definito un **campione statistico** come un insieme di dati, x_1, x_2, …, x_n, raccolti per una certa variabile su di un sottoinsieme della popolazione.

Nella **statistica inferenziale** i dati di un campione statistico vengono utilizzati per ricavare delle conoscenze sulla popolazione, s<u>fruttando dei modelli di probabilità</u>.

Ipotizziamo che i valori della variabile osservata seguano una stessa distribuzione con funzione di ripartizione F(x).

Allora possiamo rappresentare il campione statistico con n variabili aleatorie X_1, X_2, …, X_n iid aventi la stessa distribuzione F(x), di cui noi osserviamo una realizzazione (x_1 è realizzazione di X_1, …, x_n è realizzazione di X_n). → In questo caso si parla di **campione aleatorio**

INFERENZA PARAMETRICA E NON PARAMETRICA

In pratica la distribuzione F(x) non è mai completamente nota. Si usano appunto i dati del campione statistico per fare delle inferenze su F(x).

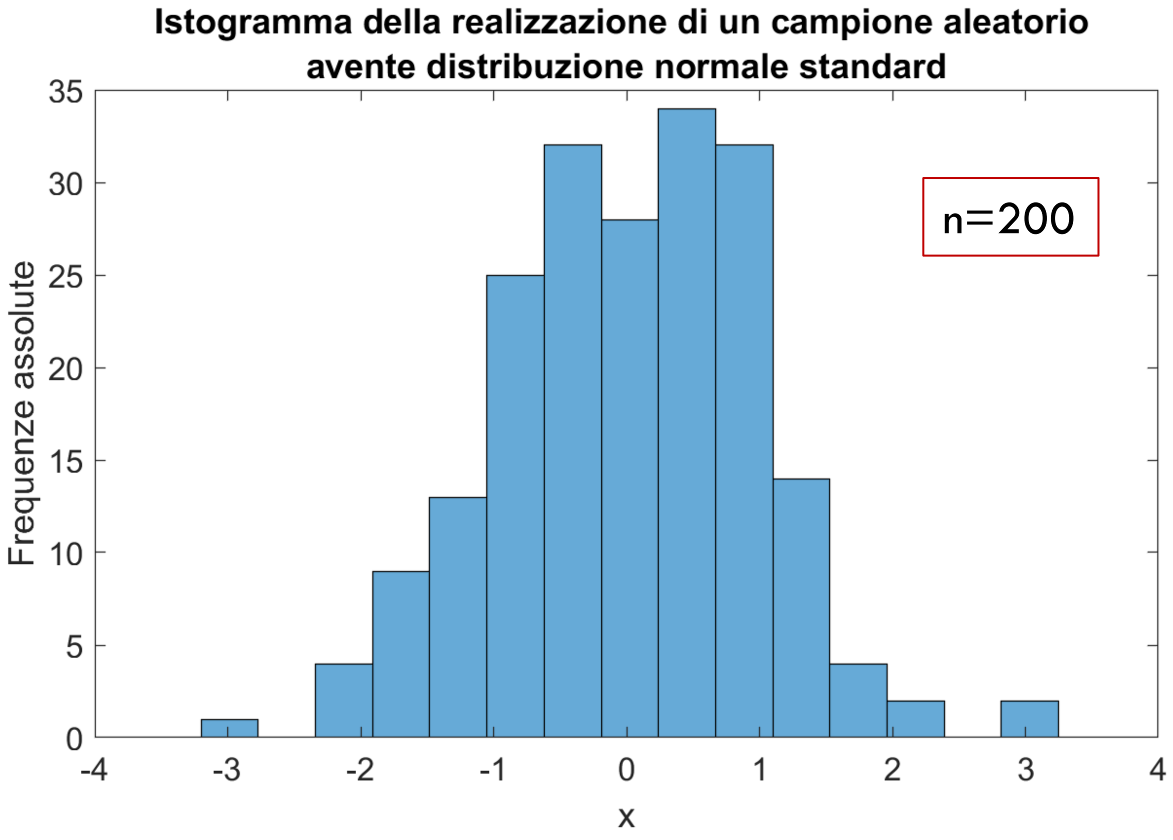
- Se F(x) è parzialmente nota, eccetto alcuni parametri incogniti, (ad esempio F(x) è una normale con media e varianza incognite) si parla di **inferenza parametrica.**
- Se F(x) è del tutto sconosciuta, si parla di **inferenza non parametrica.**

CAMPIONE NORMALE

Si dice **normale** un campione aleatorio di n variabili aleatorie normali con media μ e varianza σ^2.

Nella pratica molti campioni statistici possono essere rappresentati come campioni normali.

L'**istogramma** di una realizzazione di un campione normale avrà una forma a campana che ricorda quella della densità di probabilità normale.



INDICI DI FORMA NELLA STATISTICA DESCRITTIVA

Ora che abbiamo definito che cos'è un campione normale, torniamo un momento alla **statistica descrittiva** e definiamo gli indici di forma.

Gli **indici di forma** sono indici statistici che ci forniscono un'indicazione sulla forma che assume la distribuzione di un campione statistico, in particolare dandoci un'idea di quanto questa si discosti dalla distribuzione di un campione normale.

- Indice di asimmetria o skewness campionaria
- Indice di curtosi campionaria

INDICE DI ASIMMETRIA O SKEWNESS

L'**indice di asimmetria o skewness**, γ_1, di una variabile aleatoria X è definito come il rapporto tra il suo momento centrale di ordine 3, μ_3, e la sua deviazione standard, σ, al cubo:

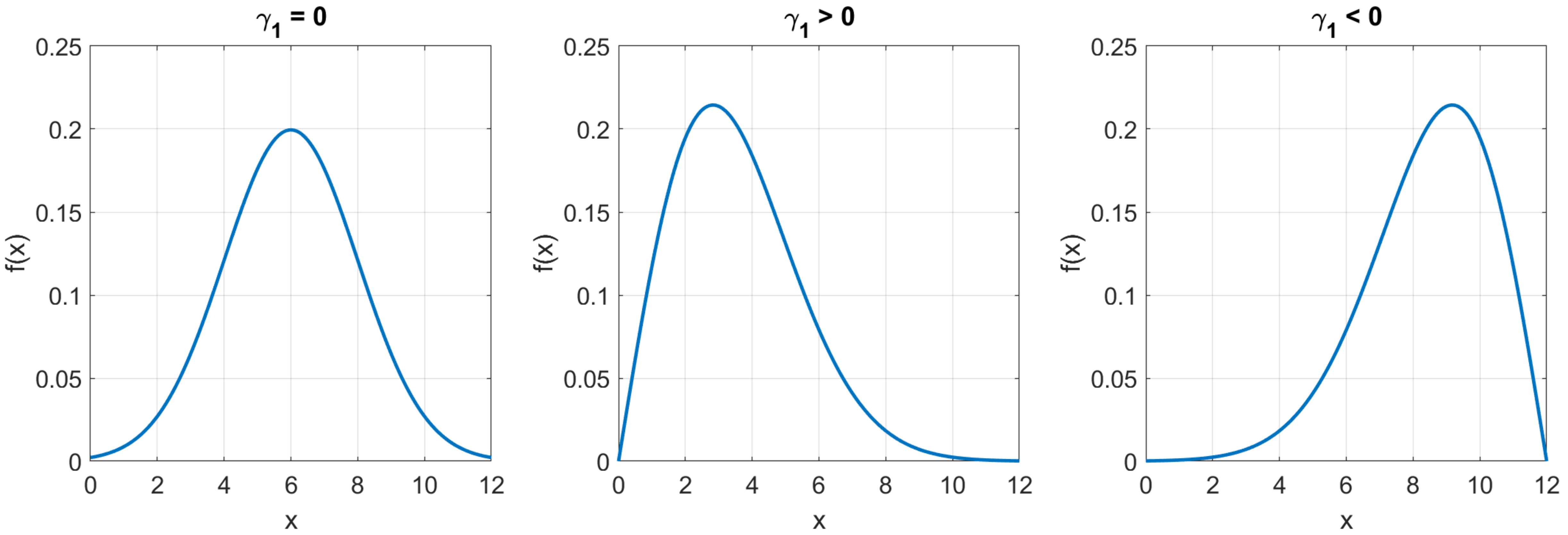
$$\gamma_1 = \frac{\mu_3}{\sigma^3},$$

$$\mu_3 = E[(X - E[X])^3], \qquad \sigma = \sqrt{E[(X - E[X])^2]}$$

L'indice γ_1 quantifica il grado di assimmetria della funzione di densità di probabilità, se X è continua, o la funzione di massa, se X è discreta.

- $\boldsymbol{\gamma_1}$**= 0** → se la densità di probabilità (o la funzione di massa) è **simmetrica**
- $\boldsymbol{\gamma_1}$**>0** → se la densità di probabilità (o la funzione di massa) è asimmetrica con una coda verso destra → **asimmetria positiva**
- $\boldsymbol{\gamma_1}$**<0** → se la densità di probabilità (o la funzione di massa) è asimmetrica con una coda verso sinistra → **asimmetria negativa**

ESEMPI



Densità simmetrica

Densità con coda verso destra → asimmetria positiva

Densità con coda verso sinistra → asimmetria negativa

INDICE DI ASIMMETRIA O SKEWNESS CAMPIONARIA

Nella pratica se disponiamo di una campione statistico di n realizzazioni di variabili aleatorie distribuite come X, x_1,x_2, …, x_n, possiamo calcolare l'indice di asimmetria campionaria come:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^{3/2}} \qquad \text{o} \qquad G_1 = \frac{\sqrt{n\cdot(n-1)}}{n-2}\cdot g_1$$

μ_3 e σ^3 calcolati con gli stimatori distorti

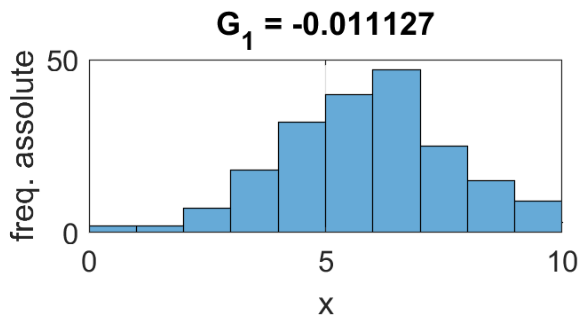
μ_3 e σ^3 calcolati con gli stimatori corretti

L'interpretazione degli indici g_1 e G_1 è analoga a quella illustrata per γ_1.

RELAZIONE TRA MEDIA E MEDIANA CAMPIONARIA NELLE DISTRIBUZIONI ASIMMETRICHE UNIMODALI

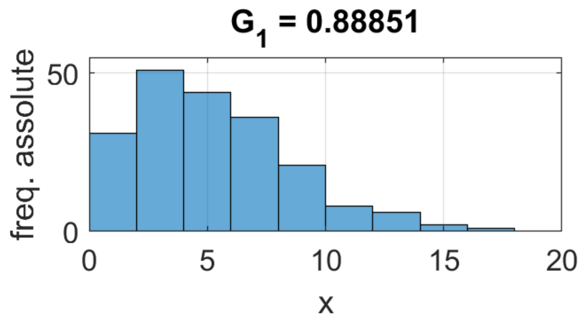
Quando la distribuzione dei dati è <u>unimodale</u> (c'è un solo valore a frequenza massima, ovvero la moda esiste), <u>generalmente</u> si ha che:

- Se la distribuzione è simmetrica → media e mediana campionaria sono molto simili tra loro



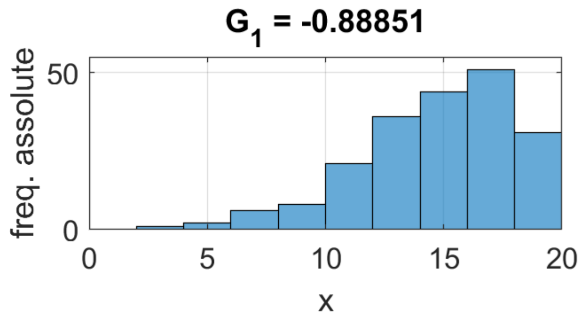
Media = 5.91
Mediana = 5.98

- Se la distribuzione ha asimmetria positiva → mediana < media



Media = 5.28
Mediana = 4.68

- Se la distribuzione ha asimmetria negativa → media < mediana



Media = 14.73
Mediana = 15.32

Nota: questa regola non vale sempre!

INDICE DI CURTOSI

L'**indice di curtosi**, γ_2, di una variabile aleatoria X è definito come il rapporto tra il suo momento centrale di ordine 4, μ_4, e la sua varianza al quadrato:

$$\gamma_2 = \frac{\mu_4}{\sigma^4}$$

$$\mu_4 = E[(X - E[X])^4], \qquad \sigma = \sqrt{E[(X - E[X])^2]}$$

L'indice γ_2 quantifica quanto la funzione di densità di probabilità, se X è continua, o la funzione di massa, se X è discreta, è «appuntita». <u>Ha senso solo per distribuzioni unimodali</u>.

- $\boldsymbol{\gamma_2}$**= 3** → se la densità di probabilità (o la funzione di massa) è **normocurtica**, ovvero «appuntita» come quella di una normale
- $\boldsymbol{\gamma_2}$**>3** → se la densità di probabilità (o la funzione di massa) è **leptocurtica**, ovvero più «appuntita» di quella di una normale
- $\boldsymbol{\gamma_2}$**<3** → se la densità di probabilità (o la funzione di massa) è **platicurtica**, ovvero meno «appuntita» di quella di una normale

INDICE DI CURTOSI CAMPIONARIA

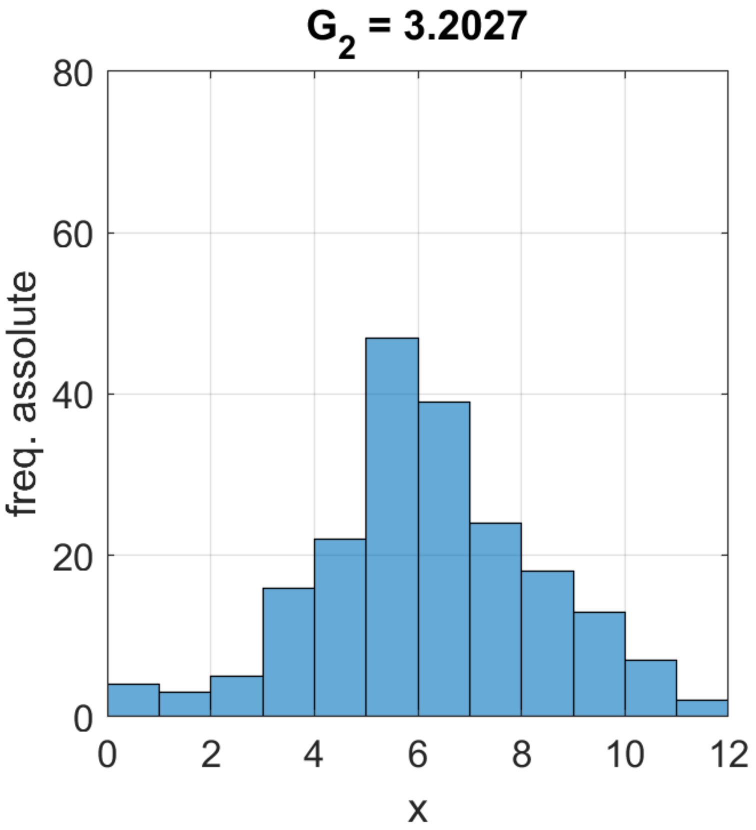
Nella pratica se disponiamo di una campione statistico di n realizzazioni di variabili aleatorie distribuite come X, x_1,x_2, …, x_n, possiamo calcolare l'indice di curtosi campionaria come:

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \quad \text{o} \quad G_2 = \frac{n-1}{(n-2)\cdot(n-3)} \cdot [(n+1)g_2 - 3(n-1)] + 3$$

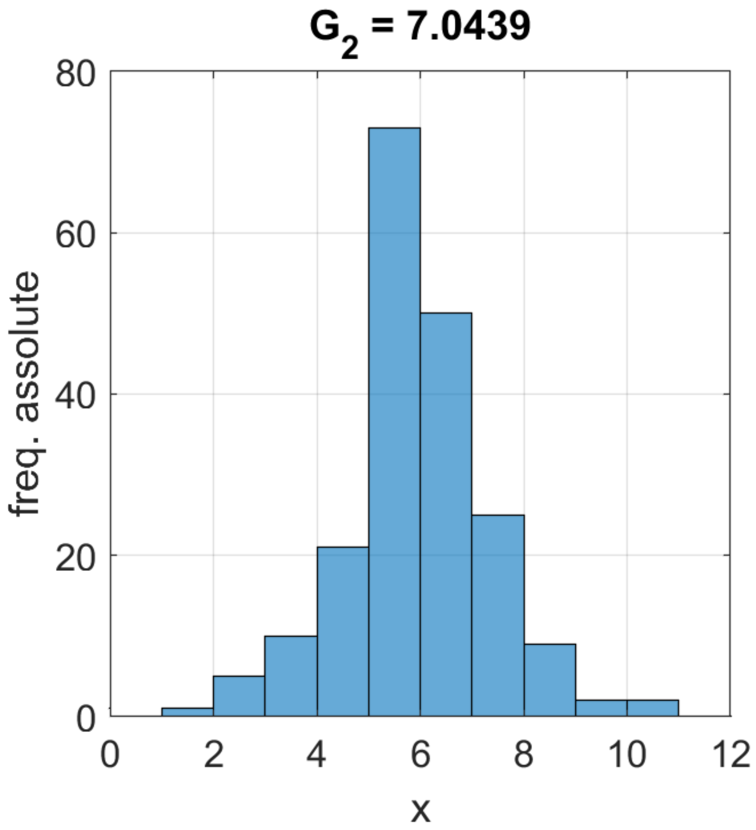
μ_4 e σ^4 calcolati con gli stimatori distorti

μ_4 e σ^4 calcolati con gli stimatori corretti

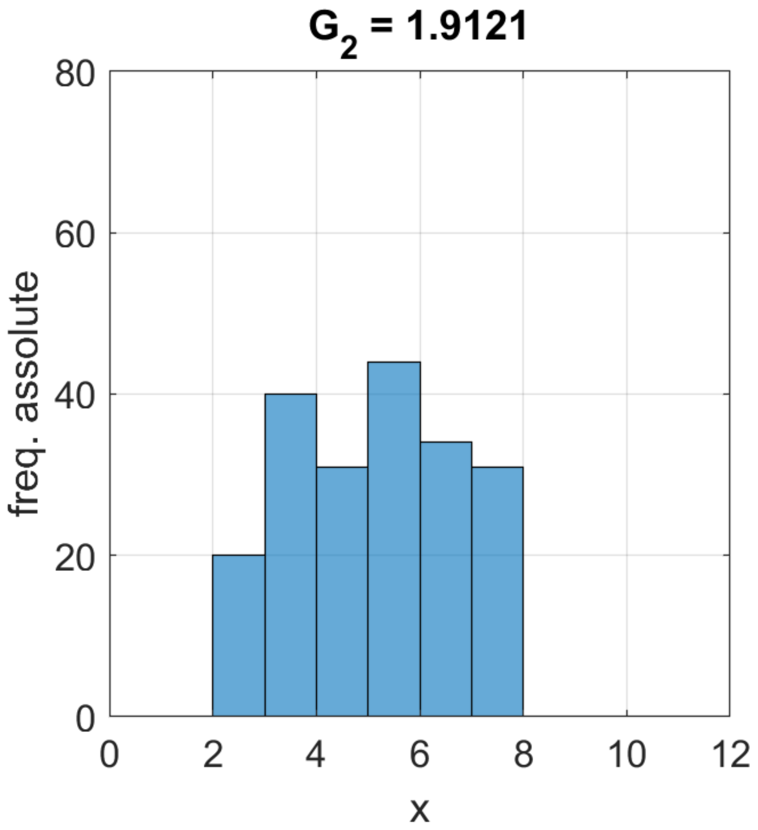
ESEMPI



Distribuzione approssimativamente normocurtica



Distribuzione leptocurtica



Distribuzione platicurtica