METODI STATISTICI PER IL MANAGEMENT

Integrazioni al libro di testo

Domenico Piccolo

Università degli Studi della Basilicata

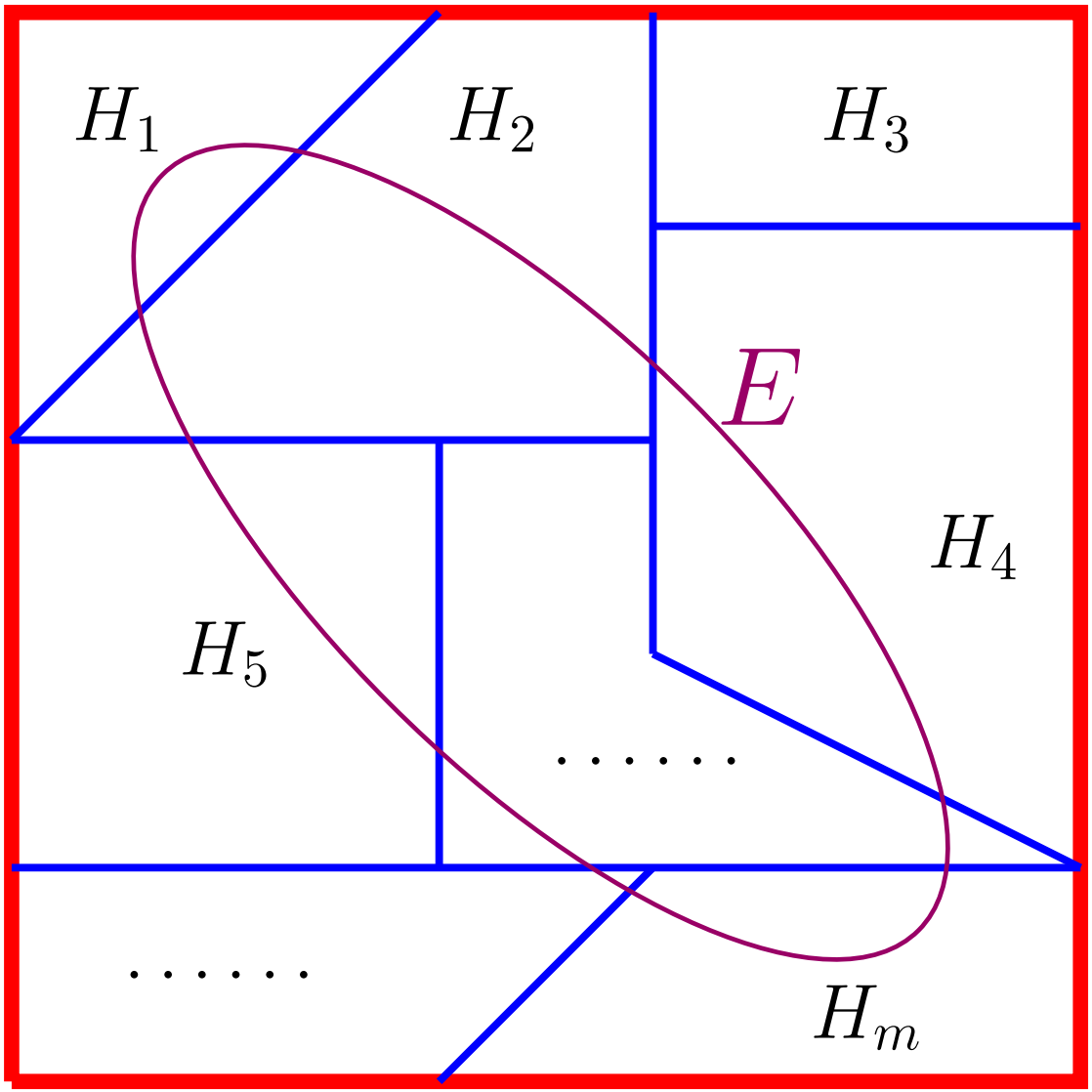
domenico.piccolo@unibas.it

Indice del materiale integrativo

Integrazione al Teorema di Bayes

➤ ***La Figura successiva va collocata immediatamente dopo il riquadro del Teorema di Bayes, nella sezione 8.13, a pag.221 del testo di riferimento***.

Rappresentazione della partizione di Ω



- Questo materiale completa l'elenco delle variabili casuali continue di uso più comune nelle applicazioni mediante la presentazione della variabile casuale Esponenziale Negativa.
- Inoltre, si aggiungono definizioni dei quantili su v.c. continue per agevolare l'uso delle Tavole.
- Infine, si presenta in modo compatto la rappresentazione della distribuzione di probabilità di una v.c. discreta.

La variabile casuale **Esponenziale Negativa**

➤ In alcune prove, i risultati osservati possono assumere tutti i **valori reali non-negativi** (cioè maggiori o uguali a zero).

➤ Spesso, queste prove derivano da esperimenti soggetti ad incertezza per i quali interessa studiare il ***tempo***:

- Durata di una conversazione telefonica
- Tempo di fermata ad un semaforo
- Attesa al telefono prima che risponda un operatore
- Intervallo di tempo fra due chiamate ad un centralino, fra due incidenti stradali su un tratto autostradale, etc.
- .....

➤ Un variabile casuale molto utilizzata in questi ambiti è la v.c. *Esponenziale Negativa*.

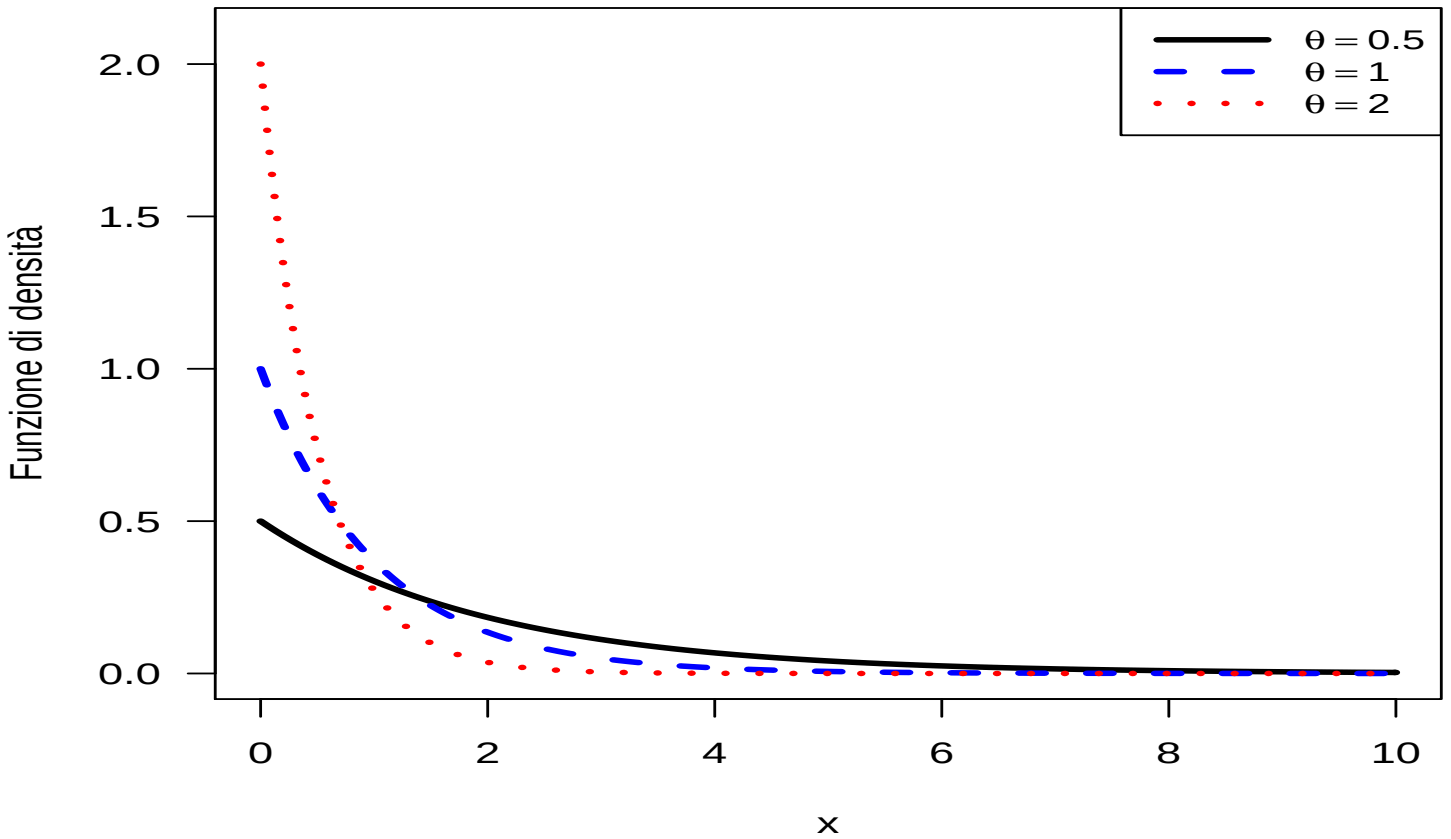
Integrazione alla **Lezione 05**

- ***Collocare prima della sezione 10.2, a pag.286***

La famiglia delle v.c. Esponenziali Negative

➤ Una v.c. **continua** X appartiene alla famiglia delle v.c. ***Esponenziali Negative***, e si indica con $X \sim En(\theta)$, se la sua funzione di densità è definita da:

$$\mathbf{f(x\,,\,\theta)} \;=\; \begin{cases} \theta\, e^{-\theta\, x} & \text{, se } x \geq 0\,; \\ 0 & \text{, altrove}\,; \end{cases} \qquad \boldsymbol{\theta > 0}\,.$$



Caratterizzazioni della v.c. Esponenziale Negativa

- Il parametro $\theta > 0$ caratterizza l'intera distribuzione. Al crescere di θ aumenta la probabilità che si verificano valori bassi di X.

- Per esempio, con riferimento alla Figura precedente, la probabilità che la v.c. X assuma un valore minore di 1 è:

$$\begin{cases} Pr\left(0 \leq X \leq 1 \mid \theta = 0.5\right) & = 0.393\,; \\ Pr\left(0 \leq X \leq 1 \mid \theta = 1.0\right) & = 0.632\,; \\ Pr\left(0 \leq X \leq 1 \mid \theta = 2.0\right) & = 0.865\,. \end{cases}$$

- **Valore medio** e **varianza** sono inversamente legati al parametro θ:

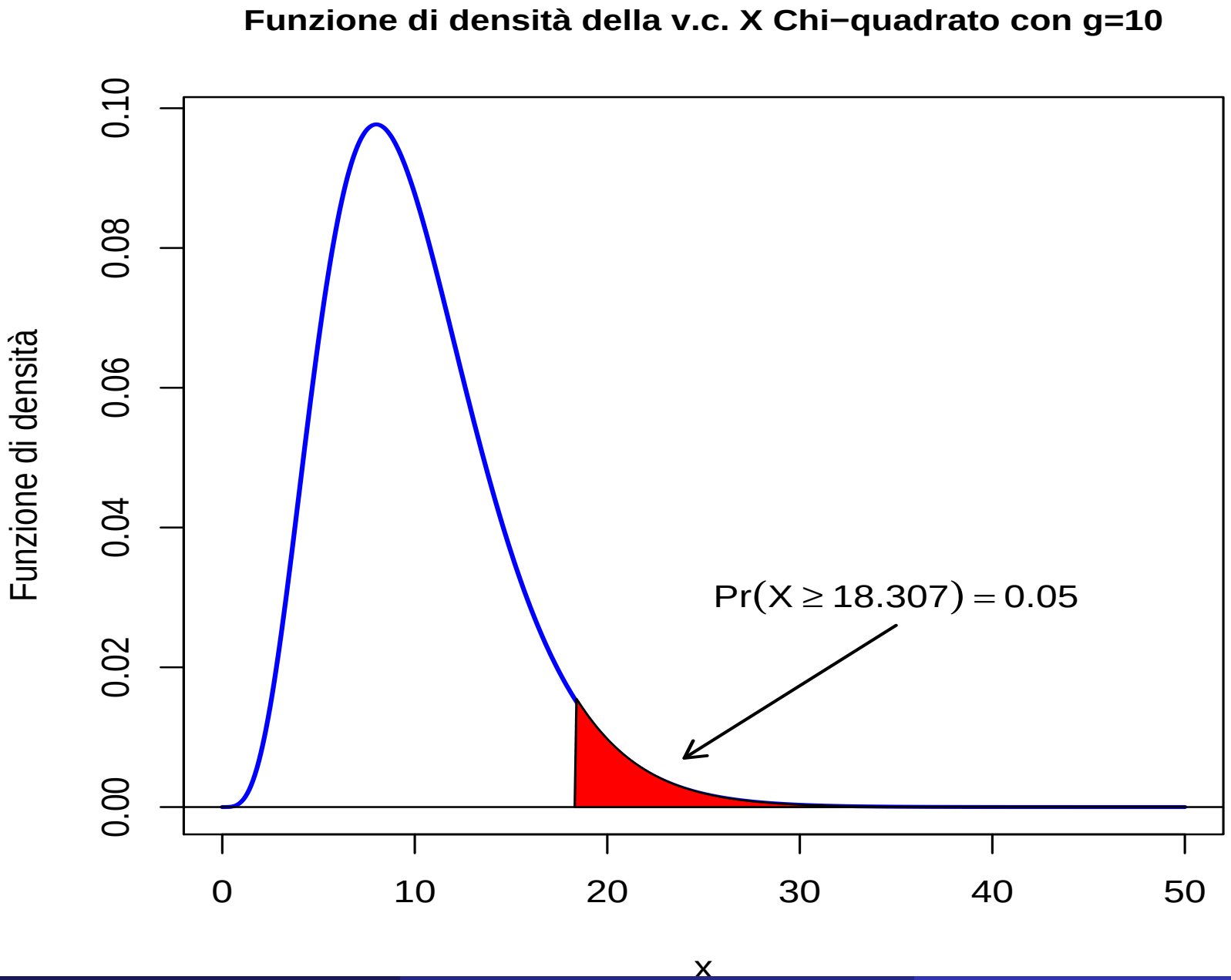
$$\mathbb{E}\left(X\right) = \frac{1}{\theta}\,; \qquad Var(X) = \frac{1}{\theta^2}\,.$$

- É importante osservare che se $X \sim U(0,1)$ (*Uniforme continua* fra 0 e 1), allora la v.c. $-\log\left(\dfrac{1-X}{\theta}\right) \sim En(\theta)$ (*Esponenziale Negativa* con parametro θ).

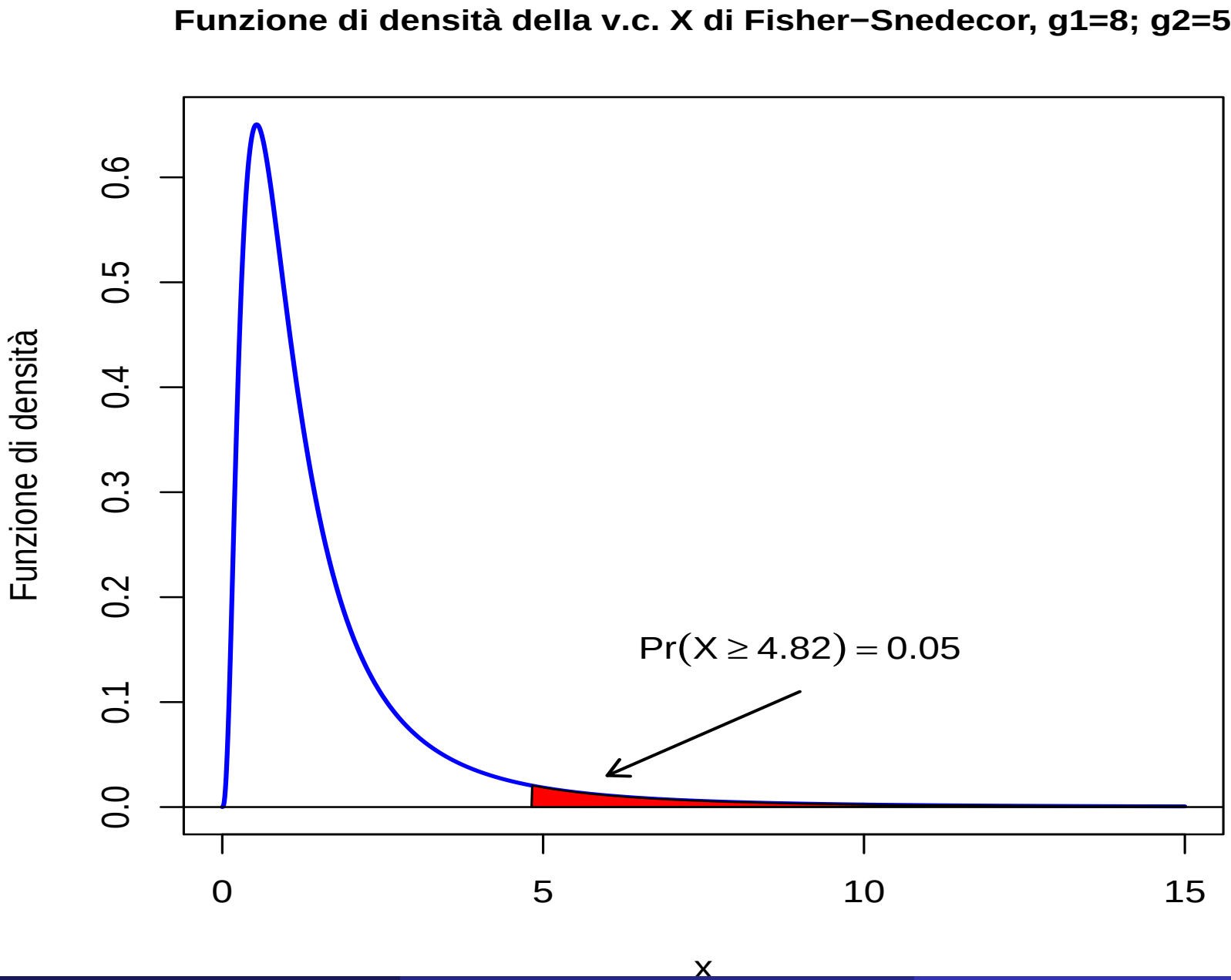
Integrazione alla **Lezione 05**

- ***Collocare nella sezione 10.4, alle pagg.295-297***

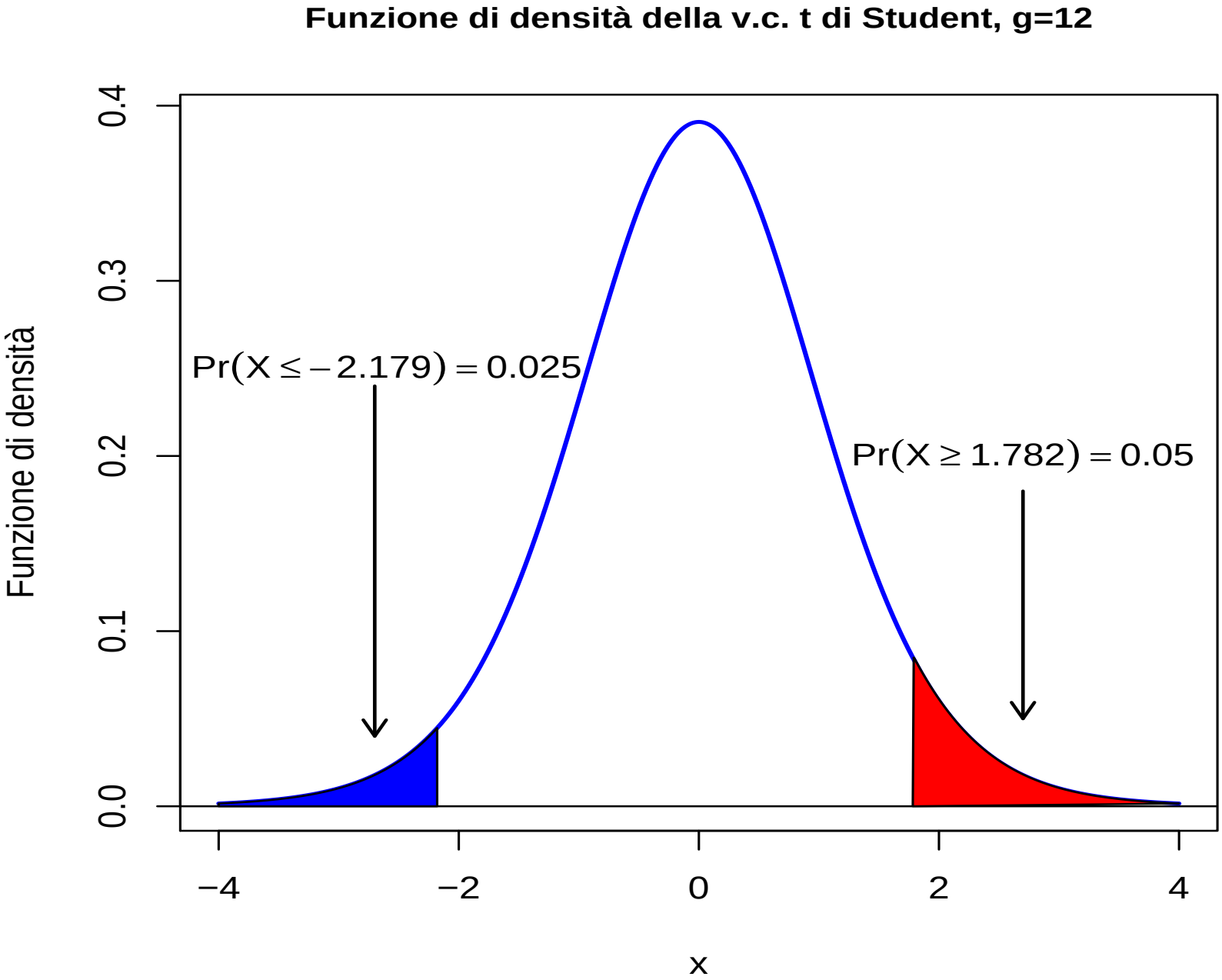
Esempio di quantile di una variabile casuale χ^2



Esempio di quantile di una variabile casuale F



Esempi di quantili di una variabile casuale t



Integrazione alla **Lezione 05**

- ***Collocare prima dell'inizio della sezione 10.5, a pag.297***

Esprimere una distribuzione di probabilità mediante una funzione

➤ Si supponga che una v.c. discreta X ben definita che assume due soli valori, per esempio 1 e 2, con probabilità $\frac{1}{4}$ e $\frac{3}{4}$.

Valori di X	Probabilità
$(X = 1)$	$\frac{1}{4}$
$(X = 2)$	$\frac{3}{4}$
Totale	1

➤ Se si introduce la funzione **indicatore** $I(A)$, che vale 1 se l'affermazione A è vera e vale 0 se l'affermazione A è falsa, la precedente tabella, e quindi l'intera distribuzione di probabilità, si può sintetizzare mediante la funzione:

$$p(x) = Pr\,(X = x) = \left(\frac{1}{4}\right)^{I(x=1)} \left(\frac{3}{4}\right)^{I(x=2)}, \qquad x = 1,\, 2\,.$$

Si controlla subito che:

$$Pr\,(X = 1) = \left(\frac{1}{4}\right)^{I(x=1)} \left(\frac{3}{4}\right)^{I(x=2)} = \left(\frac{1}{4}\right)^{1} \left(\frac{3}{4}\right)^{0} = \frac{1}{4}$$

e, similmente, per $Pr\,(X = 2)$.

Espressione compatta della distribuzione di una v.c. discreta

- Una v.c. discreta è nota se si conosce la sua distribuzione di probabilità:

Valori di $X \rightarrow$	x_1	x_2	$\ldots$	x_i	$\ldots$
Probabilità $\rightarrow$	p_1	p_2	$\ldots$	p_i	$\ldots$

- Utilizzando la **funzione indicatore** $I(x = x_i)$, che vale 1 quando x assume il valore reale x_i e vale 0 quando x assume qualsiasi valore reale differente da x_i, l'intera distribuzione di probabilità si può esprimere mediante la funzione:

$$p(x) = Pr\,(X = x) = p_1^{I(x=x_1)}\; p_2^{I(x=x_2)}\; \ldots\; p_i^{I(x=x_i)}\; \ldots, \quad x = x_1, x_2, \ldots, x_i, \ldots$$

- Così, per esempio, se $x = x_2$ allora: $I(x = x_2) = 1$ e, quindi,

$$\begin{aligned} Pr\,(X = x_2) &= p_1^{I(x=x_1)}\; p_2^{I(x=x_2)}\; \ldots\; p_i^{I(x=x_i)}\; \ldots \\ &= p_1^0\; p_2^1\; \ldots\; p_i^0\; \ldots\; = p_2\,. \end{aligned}$$

- Questo modo compatto di esprimere l'intera distribuzione di probabilità sarà necessario negli sviluppi successivi.

➤ Questo materiale presenta l'importante teorema di fattorizzazione e sottolinea il suo ruolo nella individuazione di una statistica sufficiente per il parametro a partire dalla funzione di verosimiglianza.

➤ Successivamente, sono mostrate in dettaglio alcune applicazioni dell'uso di tale teorema.

➤ **Questi argomenti vanno studiati "dopo" aver esaminato con attenzione:**

- il ***concetto di sufficienza***
- la ***funzione di verosimiglianza***

Importanza e vincoli per stimatori sufficienti

➤ La sufficienza di uno stimatore T_n per θ è proprietà molto importante.

➤ La sufficienza richiede di prendere in considerazione la sufficienza se non si assume che la v.c. X appartenga ad una famiglia di v.c. ben definita perché la definizione di sufficienza richiede la conoscenza della famiglia delle distribuzioni della v.c. X.

➤ D'altra parte, non sempre esiste uno stimatore sufficiente per il parametro di interesse per cui la sola conoscenza della famiglia a cui appartiene la v.c. X non è un criterio conclusivo.

➤ D'altra parte, il problema successivo è quello di pervenire alla costruzione di uno stimatore sufficiente.

➤ La risposta a questi problemi è offerta dal **Teorema di fattorizzazione** (di Neyman e Fisher) che consente di verificare se esiste uno stimatore sufficiente e come ottenerlo.

➤ **Questo teorema mette in relazione la funzione di verosimiglianza con gli stimatori sufficienti**.

Teorema di fattorizzazione per stimatori sufficienti

Condizione necessaria e sufficiente **perché la statistica T_n sia sufficiente per θ è che la *funzione di verosimiglianza* $\mathcal{L}(\theta;\, \underline{x})$ si possa esprimere mediante il prodotto di due funzioni $g(T_n\,;\theta)$ e $h(X_1, X_2, \ldots, X_n)$ di cui la prima sia funzione del parametro θ e del campione casuale solo attraverso la statistica T_n mentre la seconda sia funzione del campione casuale (ma non del parametro θ):**

$$\mathcal{L}(\boldsymbol{\theta};\, \underline{\mathbf{x}}) \;=\; \mathbf{g}\left(\mathbf{T_n};\, \boldsymbol{\theta}\right) \times \mathbf{h}(\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})\,.$$

➤ Si noti che $h(X_1, X_2, \ldots, X_n)$ potrebbe essere anche una costante nota perché, per esempio: $X_1^0 \times X_2^0 \times \cdots \times X_n^0 = 1$.

- Se $X \sim Ber(\theta)$, la funzione di verosimiglianza di $(X_1, X_2, \ldots, X_n)$ è:

$$\mathcal{L}(\theta; \underline{x}) = \theta^{\sum X_i}(1-\theta)^{3-\sum X_i} = \theta^{T_n}(1-\theta)^{3-T_n}$$

dove, si è indicato con $T_n = \sum X_i$.

- Ebbene, si può applicare il Teorema di fattorizzazione perché:

$$\mathcal{L}(\theta; \underline{x}) = \underbrace{\theta^{T_n}(1-\theta)^{3-T_n}}_{g(T_n;\theta)} \times \underbrace{1}_{h(X_1,X_2,\ldots,X_n)} .$$

- Come si vede, il primo fattore include sia il parametro θ che il campione casuale mediante la "sintesi" T_n che è, allora, la statistica sufficiente per θ perché il secondo fattore non include il parametro ma solo il campione casuale (in effetti è la costante 1).

- Quindi, $T_n = \sum X_i$ è la statistica sufficiente per θ per la famiglia delle v.c. di Bernoulli.

Secondo esempio del Teorema di fattorizzazione

- Un v.c. X ha funzione di densità ben definita sull'intervallo $(0, 1)$:

$$f(x) = \begin{cases} \theta\, x^{\theta-1}\,, & 0 < x < 1\,; \\ 0\,, & \text{altrove.} \end{cases}$$

- La funzione di verosimiglianza per $(X_1, X_2, \ldots, X_n)$ è:

$$\mathcal{L}(\theta;\, \underline{x}) = \theta\, X_1^{\theta-1}\ \theta\, X_2^{\theta-1}\ \ldots\ \theta\, X_n^{\theta-1} = \theta^n\ (X_1 \times X_2 \times \cdots \times X_n)^{\theta-1}\ .$$

- Se si pone: $T_n = (X_1 \times X_2 \times \cdots \times X_n)$, la funzione di verosimiglianza si fattorizza nel modo seguente:

$$\mathcal{L}(\theta;\, \underline{x}) = \underbrace{\theta^n\ (T_n)^{\theta-1}}_{g(T_n;\,\theta)} \times \underbrace{1}_{h(X_1, X_2, \ldots, X_n)}\ .$$

- Quindi, $T_n = (X_1 \times X_2 \times \cdots \times X_n)$ è la statistica sufficiente per questa famiglia di v.c.

- Sia X una v.c. Normale di valore medio θ che abbia varianza nota e pari a 1, quindi $X \sim N(\theta, 1)$.
- La sua funzione di densità è:

$$f(x) = \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}(x-\theta)^2}\,, \qquad -\infty < x < \infty\,.$$

- La funzione di verosimiglianza per $(X_1, X_2, \ldots, X_n)$ è:

$$\mathcal{L}(\theta;\, \underline{x}) = \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}(X_1-\theta)^2}\, \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}(X_2-\theta)^2}\, \ldots\, \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}(X_n-\theta)^2}$$

$$= \left(\frac{1}{\sqrt{2\,\pi}}\right)^n\, e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2}$$

- Sviluppando il quadrato all'esponente, la funzione di verosimiglianza si presenta nel modo seguente.

$$
\begin{aligned}
\mathcal{L}(\theta;\, \underline{x}) &= \left(\frac{1}{\sqrt{2\,\pi}}\right)^{n} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2} \\
&= \left(\frac{1}{\sqrt{2\,\pi}}\right)^{n} e^{-\frac{1}{2}\sum_{i=1}^{n}\left[X_i^2+\theta^2-2\,\theta\,X_i\right]} \\
&= \left(\frac{1}{\sqrt{2\,\pi}}\right)^{n} e^{-\frac{1}{2}\sum_{i=1}^{n}X_i^2}\, e^{-\frac{1}{2}\sum_{i=1}^{n}\theta^2-\frac{1}{2}(-2\,\theta)\sum_{i=1}^{n}X_i} \\
&= \left(\frac{1}{\sqrt{2\,\pi}}\right)^{n} e^{-\frac{1}{2}\sum_{i=1}^{n}X_i^2}\, e^{-\frac{n}{2}\theta^2\;+\;\theta\sum_{i=1}^{n}X_i}\,.
\end{aligned}
$$

➤ L'ultima espressione della funzione di verosimiglianza si può fattorizzare:

$$\mathcal{L}(\theta;\, \underline{x}) \quad = \quad \left[\left(\frac{1}{\sqrt{2\,\pi}} \right)^n \quad e^{-\frac{1}{2}\sum_{i=1}^{n} X_i^2} \right] \quad \left[e^{-\frac{n}{2}\theta^2 \;+\; \theta \sum_{i=1}^{n} X_i} \right]$$

$$= \quad \underbrace{\left[e^{-\frac{n}{2}\theta^2 \;+\; \theta\, T_n} \right]}_{g(T_n;\,\theta)} \times \underbrace{\left[\left(\frac{1}{\sqrt{2\,\pi}} \right)^n \quad e^{-\frac{1}{2}\sum_{i=1}^{n} X_i^2} \right]}_{h(X_1, X_2, \ldots, X_n)},$$

dove si è posto: $T_n = \sum_{i=1}^{n} X_i$.

➤ Secondo il *teorema di fattorizzazione*, T_n è una statistica sufficiente per il parametro θ nella famiglia delle v.c. $X \sim N(\theta, 1)$.

- Si consideri la v.c. X discreta che ha la seguente distribuzione:

x	$Pr(X = x)$
0	$1 - 3\,\theta$
1	θ
2	$2\,\theta$

- Tale v.c. θ è ben definita se $\theta \in \left[0,\, \frac{1}{3}\right]$. Inoltre, è agevole ottenere:

$$\mathbb{E}(X) = 5\,\theta; \qquad Var(X) = \theta\,(9 - 25\,\theta)\,.$$

- In forma compatta, la distribuzione di probabilità di tale v.c. può essere espressa nel modo seguente:

$$f(x;\,\theta) \;=\; Pr(X = x) \;=\; (1 - 3\theta)^{I(x=0)}\;(\theta)^{I(x=1)}\;(2\theta)^{I(x=2)}\;, \quad x = 0, 1, 2.$$

- Dato il campione casuale $(X_1, X_2, \ldots, X_n)$, la funzione di verosimiglianza è:

$$\begin{aligned}
\mathcal{L}(\theta;\, \underline{x}) &= f(X_1;\, \theta)\, f(X_2;\, \theta)\, \ldots\, f(X_n;\, \theta) \\
&= \prod_{i=1}^{n} f(X_i;\, \theta) \\
&= \prod_{i=1}^{n} (1-3\theta)^{I(X_i=0)}\ (\theta)^{I(X_i=1)}\ (2\theta)^{I(X_i=2)} \\
&= (1-3\theta)^{\sum_{i=1}^{n} I(X_i=0)}\ (\theta)^{\sum_{i=1}^{n} I(X_i=1)}\ (2\theta)^{\sum_{i=1}^{n} I(X_i=2)} \\
&= (1-3\theta)^{N_0}\ (\theta)^{N_1}\ (2\theta)^{N_2} = (1-3\theta)^{N_0}\ (\theta)^{N_1}\ (2)^{N_2}\ (\theta)^{N_2} \\
&= (1-3\theta)^{N_0}\ (\theta)^{N_1+N_2}\, (2)^{N_2} = (1-3\theta)^{N_0}\ (\theta)^{n-N_0}\ (2)^{N_2} \\
&= \left(\frac{1-3\,\theta}{\theta}\right)^{N_0}\ (\theta)^{n}\ (2)^{N_2}
\end{aligned}$$

essendo:

$$N_0 = \sum_{i=1}^{n} I(X_i = 0)\,; \quad N_1 = \sum_{i=1}^{n} I(X_i = 1)\,; \quad N_2 = \sum_{i=1}^{n} I(X_i = 2)\,.$$

- Inoltre: $n = N_0 + N_1 + N_2$ e quindi: $N_1 + N_2 = n - N_0$.

- Allora, la funzione di verosimiglianza si fattorizza nel modo seguente:

$$\mathcal{L}(\theta;\, \underline{x}) \;=\; \underbrace{\left(\frac{1-3\,\theta}{\theta}\right)^{N_0} (\theta)^n}_{g(T_n;\,\theta)} \;\times\; \underbrace{(2)^{N_2}}_{h(X_1,X_2,\ldots,X_n)} \quad .$$

- Essendo

$$N_0 = \sum_{i=1}^{n} I(X_i = 0) = \{\textit{Quanti in } (X_1, X_2, \ldots, X_n) \textit{ sono uguali a } 0\}$$

una funzione del campione casuale, la *statistica sufficiente* per θ per questa famiglia di v.c. è $T_n = N_0$.

- Il risultato è molto ragionevole se si considera la particolare distribuzione di probabilità di queste v.c.

➤ Questo materiale presenta un esempio di confronto tra stimatori alternativi per lo stesso parametro e permettere di sottolineare il ruolo e l'importanza delle differenti proprietà degli stimatori.

➤ Si dispone di un campione casuale $(X_1, X_2, \ldots, X_n)$ estratto da una v.c. Normale $X \sim N(\theta, 1)$ di cui, per semplicità, si suppone di conoscere la varianza $\sigma^2 = 1$.

➤ Per la stima del parametro θ (che è il valore medio della popolazione) si cerca uno stimatore lineare:

$$T_n = \sum_{i=1}^{n} a_i \, X_i$$

dove a_i, per $i = 1, 2, \ldots, n$ sono delle costanti note.

➤ Per la stima di θ si propongono due **stimatori lineari** U_n e V_n caratterizzati, rispettivamente, dai coefficienti:

$$a_i^{(U)} = \frac{1}{n}\,; \qquad a_i^{(V)} = \frac{2\,i}{n(n+1)}\,; \qquad i = 1, 2, \ldots, n\,.$$

➤ Occorre, allora, confrontare i due seguenti stimatori lineari alternativi per stimare il valore medio θ:

$$\begin{cases} U_n = \displaystyle\sum_{i=1}^{n} a_i^{(U)} X_i = \sum_{i=1}^{n} \left(\frac{1}{n}\right) X_i = \frac{1}{n} \sum_{i=1}^{n} X_i \,; \\ V_n = \displaystyle\sum_{i=1}^{n} a_i^{(V)} X_i = \sum_{i=1}^{n} \left(\frac{2\,i}{n(n+1)}\right) X_i = \frac{2}{n(n+1)} \sum_{i=1}^{n} i\, X_i \,. \end{cases}$$

➤ Lo stimatore U_n attribuisce un coefficiente costante ad ogni elemento del campione mentre lo stimatore V_n attribuisce un coefficiente variabile (crescente).

➤ Si può dimostrare che entrambi questi stimatori **lineari** sono **non-distorti** per θ, **consistenti** e **asintoticamente Normali**.

➤ Per gli sviluppi successivi, occorre ricordare che:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \,; \qquad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \,.$$

- Si ricordi che: $\mathbb{E}(X_i) = \theta$; $Var(Xi) = 1$, per ogni $i = 1, 2, \ldots, n$.

- **Valore medio degli stimatori**:

$$
\begin{cases}
\mathbb{E}(U_n) &= \mathbb{E}\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\right) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\theta = \dfrac{1}{n}\,n\,\theta = \theta\,; \\
\mathbb{E}(V_n) &= \mathbb{E}\left(\displaystyle\sum_{i=1}^{n}\dfrac{2\,i}{n(n+1)}\,X_i\right) = \displaystyle\sum_{i=1}^{n}\left(\dfrac{2\,i}{n(n+1)}\right)\mathbb{E}(X_i) \\
&= \dfrac{2}{n(n+1)}\displaystyle\sum_{i=1}^{n} i\,\mathbb{E}(X_i) = \dfrac{2}{n(n+1)}\,\theta\displaystyle\sum_{i=1}^{n} i \\
&= \dfrac{2}{n(n+1)}\,\theta\,\dfrac{n(n+1)}{2} = \theta\,.
\end{cases}
$$

- Entrambi gli stimatori U_n e V_n sono **non-distorti** per θ .

■ **Varianza degli stimatori**:

$$\begin{cases} Var(U_n) &= \dfrac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \dfrac{1}{n^2}\sum_{i=1}^{n} 1 = \dfrac{1}{n^2}\, n = \dfrac{1}{n}\,; \\ Var(V_n) &= \displaystyle\sum_{i=1}^{n}\left(\frac{2\,i}{n(n+1)}\right)^2 Var(X_i) = \frac{4}{n^2\,(n+1)^2}\sum_{i=1}^{n} i^2 \quad (1) \\ &= \dfrac{4}{n^2\,(n+1)^2}\,\dfrac{n(n+1)(2n+1)}{6} = \dfrac{2\,(2n+1)}{3\,n(n+1)}\,. \end{cases}$$

➤ Poiché entrambi gli stimatori sono *non-distorti per* θ, per controllare la consistenza di tali stimatori basta verificare il comportamento della loro varianza al crescere di n. Ebbene, si ha:

$$\begin{cases} \lim_{n\to\infty} Var(U_n) &= \lim_{n\to\infty} \dfrac{1}{n} = 0\,; \\ \lim_{n\to\infty} Var(V_n) &= \lim_{n\to\infty} \dfrac{2\,(2n+1)}{3\,n\,(n+1)} = 0\,; \end{cases}$$

➤ Gli stimatori U_n e V_n sono entrambi **consistenti** per θ.

■ **Distribuzione asintotica degli stimatori**:

$$\begin{cases} U_n = \sum_{i=1}^{n} \left(\frac{1}{n}\right) X_i \,; \\ V_n = \sum_{i=1}^{n} \left(\frac{2\,i}{n(n+1)}\right) X_i \,. \end{cases}$$

➤ I due stimatori U_n e V_n sono entrambi combinazione lineare di v.c. Normali e indipendenti per cui, per la *proprietà riproduttiva delle v.c. Normali*, per qualsiasi valore di n, i due stimatori si distribuiscono ***esattamente*** come v.c. Normali.

➤ Evidentemente, se U_n e V_n sono Normali per ogni n sono anche **asintoticamente Normali** per θ .

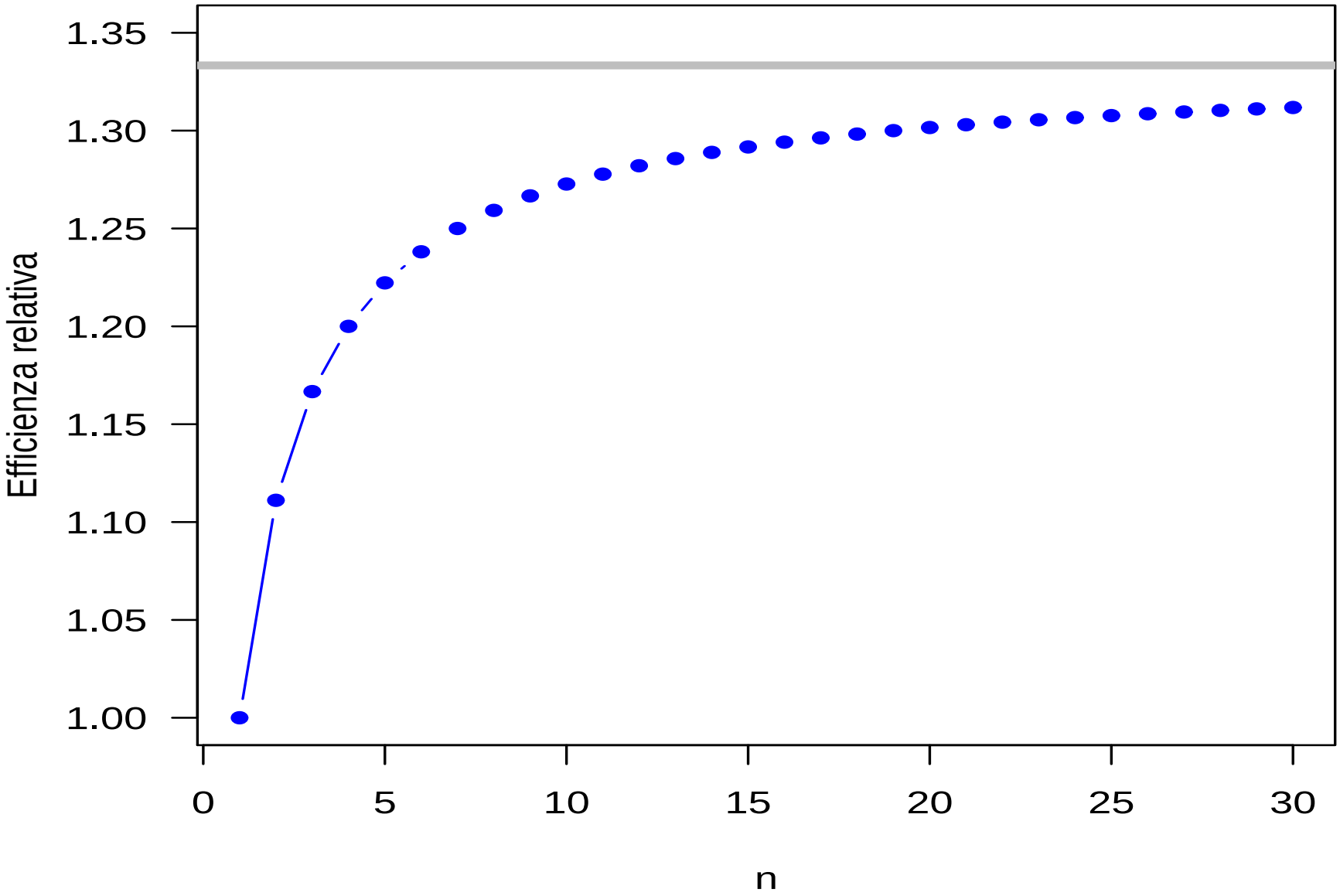
- Un confronto tra i due stimatori può avvenire in termini di **efficienza relativa** (che è una proprietà finita).
- Essendo entrambi non-distorti, l'efficienza relativa si ottiene confrontando la varianza dei due stimatori.
- Conoscendo la varianza di U_n e di V_n, si ha:

$$eff\,(U_n \mid V_n) = \frac{Var(V_n)}{Var(U_n)} = \frac{\frac{2\,(2n+1)}{3\,n(n+1)}}{\frac{1}{n}} = \frac{2}{3}\,\frac{2n+1}{n+1} = 1 + \frac{1}{3}\,\frac{1-\frac{1}{n}}{1+\frac{1}{n}} \to 1+\frac{1}{3} > 1\,.$$

- Per ogni $n > 1$, lo stimatore U_n è **più efficiente** dello stimatore V_n.

- La Figura seguente mostra come tale efficienza varia al crescere di n e tenda verso il valore limite $1.333\ldots$

- Al crescere di n, lo stimatore V_n richiede **un terzo in più di numerosità campionaria** rispetto allo stimatore U_n per raggiungere la stessa precisione (affidabilità, accuratezza, . . .).


Efficienza relativa
1.35
1.30
1.25
1.20
1.15
1.10
1.05
1.00
0
5
10
15
20
25
30
n

➤ Questo materiale introduce tali modelli mediante l'utilizzo di variabili dicotomiche, definite *dummy* (al plurale: *dummy variables*).

Variabili qualitative nei modelli di regressione

➤ Nei modelli di regressione sinora considerati sia la variabile dipendente Y che le variabili esplicative $X_1, X_2, \ldots, X_p$ potevano essere solo quantitative essendo necessario, in tutte le elaborazioni, applicare calcoli sulle modalità di tali variabili, che quindi devono essere valori numerici.

➤ I **problemi reali** richiedono, invece, di considerare situazioni in cui sia la variabile dipendente che le variabili esplicative potrebbero essere legittimamente **variabili qualitative**.

➤ Il problema si affronta in modo differente, a seconda che la variabile qualitativa sia quella dipendente oppure una o più fra le variabili esplicative.

➤ Per ragioni di semplicità didattica si affronta prima il secondo caso e, poi, si considera il primo.

- Studio dei fenomeni tramite **misurazioni**:
  - **Astronomia** (*Gauss*, *Legendre*)
  - **Antropometria, Genetica** (*Galton*)
  - **Scienze agrarie** (*Fisher*), etc.

- **Relazioni deterministiche tra fenomeni (variabili) di tipo quantitativo** modificate da **errori accidentali** (di misurazione, per lo più) richiedono l'introduzione delle v.c. errori ϵ_i.

- A partire dagli anni $'20$ del secolo scorso, con la crescita di interesse verso gli studi psicologici, sociali ed economici (***scienze umane e sociali***), ci si pone la domanda se non sia possibile costruire modelli che mettano in relazione anche le **risposte delle persone** a *stimoli esterni* (meccanici, biologici, ambientali, psicologici, etc.).

- In tali ambiti, quasi sempre, le *risposte* sono **qualitative** e le *variabili esplicative* possono essere, a seconda dei casi, *quantitative* oppure *qualitative*.

Variabili esplicative qualitative

- **Presenza nei modelli di variabili qualitative** o verificarsi un certo evento/condizione come "spiegazione razionale" di variabili dipendenti, anche quantitative.
  - **Variabili nominali**: genere, stato civile, titolo di studio, professione, regione di residenza, auto preferita, orientamento politico, religione, abitudine al fumo, stato di salute, assunzione di stupefacenti, etc.
  - **Variabili ordinali**: stress, paure, preoccupazioni, gusti, preferenze, gradimento, valutazioni, soddisfazione, felicità, percezioni, etc.
  - **Circostanze**: risultato ottenuto, esito di una prova, caratteristica posseduta, situazione specifica, occasione temporale, etc.
- Approccio mediante ***variabili dummy*** capaci di registrare mediante 0 oppure 1, rispettivamente, l'assenza oppure la presenza di una certa categoria per la variabile qualitativa.

Definizione di una **variabile dummy**

- Sia E un *evento* di cui interessa solo se si verifica (presenza) o non si verifica (assenza).
- Si definisce la ***variabile dummy*** D_i, per $i = 1, 2, \ldots, n$, tale che:

$$D_i = \begin{cases} 0\,, & \text{se per l'unità } i\text{-esima si verifica } \overline{E}\,; \\ 1\,, & \text{se per l'unità } i\text{-esima si verifica } E\,. \end{cases}$$

- Lo schema si adatta a qualsiasi **situazione dicotomica**, che può essere:
  - **genuinamente dicotomica**: Fumo, Guarigione, Vittoria, Genere, Laurea magistrale, Dipendente, etc.
  - **resa dicotomica**: Componenti familiari in numero maggiore di 3, Arrivo entro le ore 12:00, Reddito annuo superiore a $50\,000$ Euro, etc.

➤ Questo materiale introduce tali modelli mediante l'utilizzo della variabile casuale discreta di Bernoulli.

➤ L'approccio dei modelli di regressione si modifica in modo notevole quando la variabile dipendente è di tipo qualitativo.

➤ nel seguito verrà affrontata la soluzione più semplice ma anche più diffusa che conduce al ***modello di regressione logistica***.

Esemplificazioni di risposte qualitative

- ***Psicologia:*** lo **stato di stress** di una persona dipende certamente da genere, età, livelli di reddito, salute fisica, fattori ereditari, stato civile, urbanizzazione del territorio, relazioni amicali, etc.
- ***Demografia:*** la **sopravvivenza ad una certa età** oppure **la presenza di figli in una famiglia** dipendono dal livello di reddito, dal lavoro, dalle abitudini di vita, dalle condizioni di salute, etc.
- ***Sociologia:*** la **felicità** dipende certamente da genere, età, stato civile, livello di reddito, presenza di figli, traumi personali e/o familiari, abitudini di vita, relazioni sociali, ambiente di vita, etc.
- ***Medicina:*** l'insorgenza e gli stadi di una **malattia** sono collegati a genere, età, abitudini alimentari, fattori genetici, igiene personale ed ambientale, tipologia del lavoro svolto, stagioni dell'anno, assunzione di sostanze (fumo, alcool, droga, etc.), etc.
- ***Politica:*** la scelta di **votare un partito** dipende da età, titolo di studio, residenza, livello di reddito, tipologia del lavoro, scelte ideologiche, tradizioni familiari, impegno sociale, etc.
- ***Economia e Marketing:*** l'uso dei **mezzi di trasporto** così come la scelta di **acquistare un prodotto** dipendono dall'età, dal livello di reddito, dallo stato civile, dalla tipologia del lavoro, dai costi comparati del servizio/prodotto, etc.
- .....

La risposta come variabile casuale di Bernoulli

➤ Il modello deve "spiegare" la risposta Y_i (corrispondente a **SI** oppure **NO**) in funzione delle caratteristiche $(x_{i1}, x_{i2}, \ldots, x_{ip})$ del soggetto i-esimo.

➤ La risposta Y_i è una *variabile casuale discreta* che assume i valori: 0 ("Insuccesso") e 1 ("Successo"), cioè una ***variabile casuale di Bernoulli***:

Eventi	*Valori di* Y_i	*Probabilità*
$\overline{E}$ = Insuccesso	0	$1-\theta_i$
E = Successo	1	θ_i

dove $\theta_i \in [0,1]$, per ogni $i = 1, 2, \ldots, n$.

➤ In forma compatta, si può scrivere:

$$\begin{cases} Pr\,(Y_i = 0) & = 1-\theta_i \\ \\ Pr\,(Y_i = 1) & = \theta_i \end{cases} \iff \boxed{Pr\,(Y_i = y) \;=\; \theta_i^{y}\,(1-\theta_i)^{1-y}\,, \;\; y = 0,\, 1.}$$

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

➤ Valore medio e varianza delle v.c. Y_i sono:

$$\mathbb{E}\,(Y_i) = Pr\,(Y_i = 1) = \theta_i\,; \qquad Var(Y_i) = Pr\,(E)\,Pr\left(\overline{E}\right) = \theta_i\,(1-\theta_i)\,.$$