Esercizi di Statistica

Indrjo Dedej

Ultimo aggiornamento: 6 dicembre 2024.

Sommario

Questi sono alcuni esercizi svolti: possono essere problemi presenti negli esami scritti passati, esercitazioni durante le ore di tutorato oppure esercizi tratti da libri.

Questo documento e tutto il sorgente LaTeX sono disponibili all'indirizzo:

`https://github.com/indrjo/esercizi-statistica.git`

Sono ben accetti i contributi di chiunque, in particolare degli studenti del corso di Statistica dell'Università degli Studi di Pavia.

In caso di *fork* e *pull request*, aggiungere il proprio nome alla lista degli autori presente in `\author{...}` nel file `main.tex`, separandolo da quello degli altri con una virgola. Se invece preferite inviare i vostri svolgimenti ai proprietari della repo, provvederanno loro ad aggiungere i nomi.

In ogni caso, sono gradite note scritte in LaTeX, visto che gli autori potrebbero non avere tempo per mettere "in bella" tutto. Questo non vuol dire che non verranno accettate scansioni di manoscritti, ma si preferisce tenere il tutto un po' più in ordine e accessibile.

Anche se non avete le conoscenze tecniche per fare alcune cose, fa niente: del lavoro fatto è meglio che nessun lavoro fatto; e poi ci penseranno altri a migliorare il codice dietro. C'è un file `README.md` che contiene alcune istruzioni per chi vuole contribuire, sperando di rendere più semplice la partecipazione.

Indice

1 Statistica Descrittiva

Esercizio 1.1 (Problema 1, scritto del 12/02/2024)**.** Un'indagine ISTAT su $n = 100$ famiglie analizza la relazione tra il carattere X relativo alla numerosità del nucleo familiare e il carattere Y realtivo al numero di locali della rispettiva abitazione. L'indagine produce la seguente tabella

	$X = 1$	$X = 2$	$X = 3$
$Y = 2$	20	10	7
$Y = 3$	9	13	13
$Y = 4$	2	5	21

Si affrontino i seguenti quesiti.

1. Calcolare la media aritmetica (pesata) e la mediana del carattere Y.
2. Valutare la variabilità di Y tramite l'uso delle differenza media semplice (di ordine $p = 1$).

3. Valutare la concentrazione del carattere Y rispetto alla distribuzione uniforme, esplicitando il grafico della curva di Lorentz-Gini e calcolando un opportuno indice di concentrazione.
4. Calcolare la tabella relativa alla distribuzione di probabilità congiunta corrispondente alla situazione (ideale) di indipendenza dei due caratteri.
5. Scegliere un indice di connessione tra X e Y e calcolarlo numericamente coi dati forniti.
6. Calcolare la tabella corrispondente alla distribuzione di probabilità congiunta corrispondente alla situazione (ideale) di massima concordanza tra i due caratteri.
7. Scegliere un indice di concordanza tra X e Y e calcolarlo numericamente coi dati forniti.

Soluzione. [Da riscrivere. Vedere parte commentata nel sorgente TEX.] □

Esercizio 1.2 (Problema 2, 12/09/2024)**.** Sullo spazio misurabile $(\Omega, \mathcal{F}) := ((1, +\infty), \mathbb{B}(1, +\infty))$ si considera la misura di riferimento

$$Q : \mathcal{F} \to [0, 1], \; Q(B) := \int_B x^{-2} \mathrm{d}x$$

e la misura di prova

$$P : \mathcal{F} \to [0, 1], \; P(B) := \alpha \int_B x^{-1-\alpha} \mathrm{d}x$$

dove $\alpha \in (0, 1)$.

1. Calcolare la mediana di entrambe le misure.
2. Discutere la mutua variabilità rispetto alla distanza $d(x, y) := |x - y|$. In particolare, confrontare le due differenze medie di ordine 1 sfruttando la *formula di de Finetti-Paciello* secondo cui per una funzione di ripartizione F con supporto in $(0, +\infty)$ vale l'identità

$$\int_0^{+\infty} \int_0^{+\infty} |x - y| \, \mathrm{d}F(x)\mathrm{d}F(y) = 2 \int_0^{+\infty} F(x)[1 - F(x)]\mathrm{d}x.$$

3. Dopo aver verificato che P è assolutamente continua rispetto a Q, calcolare esplicitamente la curva di concentrazione di P rispetto a Q.
4. Calcolare l'area di concentrazione e, in particolare, dire se aumenta o diminuisce in funzione di α.

Soluzione. 1. Calcoliamo quindi le funzioni di ripartizione, cioè

$$F_Q : \mathbb{R} \to [0, 1], \; F_Q(t) := Q\{\omega \in \Omega \mid \omega < t\} = \begin{cases} 0 & \text{se } t \leq 1 \\ 1 - \frac{1}{t} & \text{se } t > 1 \end{cases}$$

e

$$F_P : \mathbb{R} \to [0, 1], \; F_P(t) := P\{\omega \in \Omega \mid \omega < t\} = \begin{cases} 0 & \text{se } t \leq 1 \\ 1 - \frac{1}{t^\alpha} & \text{se } t > 1 \end{cases}$$

Le rispettive mediane sono quindi

$$F_Q^{-1}\left(\frac{1}{2}\right) = \inf\left\{s \in \mathbb{R} \,\middle|\, F_Q(s) \geq \frac{1}{2}\right\} = 2$$
$$F_P^{-1}\left(\frac{1}{2}\right) = \inf\left\{s \in \mathbb{R} \,\middle|\, F_P(s) \geq \frac{1}{2}\right\} = 2^{\frac{1}{\alpha}}$$

2. [Da rivedere meglio...]
3. Sia $E \in \mathcal{F}$ qualsiasi tale che $Q(E) = 0$. Ne segue che E sia una insieme di misura nulla secondo la misura unidimensionale di Lebesgue. Ma allora anche $P(E) = 0$. Quindi per il Teorema di Radon-Nikodym esiste unica a meno uguaglianza quasi ovunque una funzione $Z : \Omega \to [0, +\infty]$ misurabile tale che
$$P(E) = \int_E Z \mathrm{d}Q \text{ per ogni } E \in \mathcal{F}.$$
Questa Z è la *derivata di Radon-Nikodym* e si scrive $\frac{\mathrm{d}P}{\mathrm{d}Q}$. Calcoliamola allora. Se indichiamo con λ la misura indotta su $(\Omega, \mathcal{F})$ dalla misura di Lebesgue, osserviamo anche che P e Q sono entrambe assolutamente continue rispetto a λ. Quindi
$$P(E) = \int_E Z \mathrm{d}Q = \int_E Z \frac{\mathrm{d}Q}{\mathrm{d}\lambda} \mathrm{d}\lambda.$$
All'ultimo membro sappiamo quanto vale $\frac{\mathrm{d}Q}{\mathrm{d}\lambda}$ quasi ovunque: grazie al Teorema di Radon-Nikodym infatti possiamo dire
$$\frac{\mathrm{d}Q}{\mathrm{d}\lambda}(x) = x^{-2} \text{ per quasi ogni } x \in \Omega.$$
Ancora nuovamente per il Teorema di Radon-Nikodym, applicato questa volta alla coppia P e λ, si ha che
$$Z(x)\frac{\mathrm{d}Q}{\mathrm{d}\lambda}(x) = \alpha x^{-1-\alpha} \text{ per quasi ogni } x \in \Omega.$$
Da cui possiamo concludere che
$$Z(x) = \alpha x^{1-\alpha} \text{ per quasi ogni } x \in \Omega.$$
In realtà, possiamo scegliere senza perdere nulla che Z di definire in questo modo su *tutto* Ω, visto che l'integrale di Lebesgue ignora insiemi di misura nulla.

   Calcoliamo la funzione di ripartizione di $Z = \frac{\mathrm{d}P}{\mathrm{d}Q}$:
$$\begin{aligned} F(t) &:= Q\{x \in \Omega \mid Z(x) \leq t\} = \\ &= Q\left\{x > 1 \,\middle|\, x \leq \left(\frac{t}{\alpha}\right)^{\frac{1}{1-\alpha}}\right\} \end{aligned}$$
Qui
$$F(t) = \begin{cases} 0 & \text{se } t \leq \alpha \\ 1 - \left(\frac{\alpha}{t}\right)^{\frac{1}{1-\alpha}} & \text{se } t > \alpha \end{cases}.$$
Ecco l'inversa generalizzata di F: per $y \in (0, 1)$
$$F^{-1}(y) = \frac{\alpha}{(1-y)^{1-\alpha}} = \alpha(1-y)^{\alpha-1}.$$
Possiamo finalmente scrivere la *funzione di concentrazione*
$$\phi(s) := \int_0^s F^{-1}(y)\mathrm{d}y = 1 - (1-s)^{\alpha}.$$

4. Abbiamo a questo punto tutto quello che serve per calcolare l'*area di concentrazione*
$$\int_0^1 [s - \phi(s)]\mathrm{d}s = \frac{1}{\alpha+1} - \frac{1}{2}.$$
In particolare, l'area è decrescente rispetto ad α. □

2 Statistica Inferenziale

Esercizio 2.1 (Problema 3, scritto del 12/02/2024)**.** Si consideri il modello statistico a n prove indipendenti avente modello base $\mathbb{X}_1 := \mathbb{R}_{>0}$, $\mathcal{X}_1 := \mathcal{B}\mathbb{R}_{>0}$, $\Theta := \mathbb{R}_{>0}$ con funzione di verosomiglianza

$$\begin{aligned} L_1 &: \Theta \times \mathbb{X}_1 \to [0, +\infty] \\ L_1(\theta, x) &:= \frac{3\theta^3}{x^4} \mathbf{1}_{[\theta,+\infty)}(x) \end{aligned}$$

dominato dalla misura unidimensionale di Lebesgue $\lambda : \mathcal{X}_1 \to [0, +\infty]$. Quindi il modello statistico base ha misure di probabilità

$$\begin{aligned} \pi_\theta^{(1)} &: \mathcal{X}_1 \to [0, 1] \\ \pi_\theta^{(1)}(E) &:= \int_E L_1(\theta, \cdot)\mathrm{d}\lambda. \end{aligned}$$

1. Determinare il modello statistico a n prove indipendenti
$$\mathbb{X}_n := \mathbb{X}_1^n\,,\ \mathcal{X}_n := \mathcal{X}_1^n\,,\ \Theta\,,\ \pi_\theta^{(n)} := \underbrace{\pi_\theta^{(1)} \otimes \cdots \otimes \pi_\theta^{(1)}}_{n \text{ volte}}.$$
e scriverne la funzione di verosomiglianza $L_n : \Theta \times \mathbb{X}_n \to [\theta, +\infty]$.
2. Scrivere lo stimatore di massima verosomiglianza $\widehat{\theta}_n : \mathbb{X}_n \to \Theta$ e calcolarne la sua legge.
3. Dimostrare che $\widehat{\theta}_n$ è uno stimatore distorto ma asintoticamente non distorto. Poi calcolare uno stimatore $\widetilde{\theta}_n : \mathbb{X}_n \to \Theta$ non distorto e proporzionale a $\widehat{\theta}_n$.
4. Dimostrare, usando i teoremi visti a lezione, che $\widetilde{\theta}_n$ è fortemente consistente.
5. Fissato $n \in \mathbb{N}$ e $\alpha \in [0, 1]$, calcolare un intervallo di confidenza $C_{n,\alpha}(x_1, \ldots, x_n)$ per il parametro θ di livello $1 - \alpha$.

Soluzione. 1. La funzione di verosomiglianza si scrive subito

$$\begin{aligned} L_n(\theta; x_1, \ldots, x_n) &= \theta^{3n} \prod_{i=1}^n \frac{\mathbf{1}_{[\theta,+\infty)}(x_i)}{x_i^4} = \\ &= \frac{\theta^{3n}}{\left(\prod_{i=1}^n x_i\right)^4} \mathbf{1}_{[\theta,+\infty)}(\min(x_1, \ldots, x_n)). \end{aligned}$$

2. Fissato $(x_1, \ldots, x_n) \in \mathbb{X}_n$, studiamo la funzione
$$L_n(\cdot, x_1, \ldots, x_n) : \Theta \to [0, +\infty].$$

e cerchiamo un $\widehat{\theta} \in \Theta$ che realizza il valore $\sup_{\theta\in\Theta} L(\theta; x_1, \dots, x_n)$. Osserviamo subito che la funzione non è derivabile: infatti vale

$$L_n(\theta, x_1, \dots, x_n) = \begin{cases} \frac{\theta^{3n}}{(\prod_{i=1}^n x_i)^4} & \text{se } \theta \leq \min(x_1, \dots, x_n) \\ 0 & \text{se } \theta > \min(x_1, \dots, x_n) \end{cases}$$

Stando così le cose, si vede subito che $L(\cdot, x_1, \dots, x_n)$ raggiunge il valore massimo per $\theta = \min(x_1, \dots, x_n)$. Quindi lo stimatore di massima verosomiglianza è

$$\begin{aligned} \widehat{\theta}_n &: \mathbb{X}_n \to \Theta \\ \theta(x_1, \dots, x_n) &:= \min(x_1, \dots, x_n). \end{aligned}$$

Cerchiamo la sua legge ora, e per fare ciò ci serve un po' di contesto: ovvero spazi di probabilità $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$ spazi di probabilità al variare di $\theta \in \Theta$ e una variabile aleatoria $X := (X_1, \dots, X_n) : \Omega \to \mathbb{X}_n$ tali che

$$\pi_\theta^{(n)}(E) = \mathbb{P}_\theta[X \in E] \text{ per ogni } E \in \mathcal{X}, \theta \in \Theta$$

e le variabili aleatorie $X_1, \dots, X_n : \Omega \to \mathbb{X}$ sono i.i.d. con densità $f_\theta := L_1(\theta, \cdot)$. Nello specifico, cercheremo la legge per la variabile aleatoria

$$\widehat{\theta}_n(X_1, \dots, X_n) : \Omega \to \Theta.$$

Con "legge" intendiamo la densità oppure equivalentemente una funzione di ripartizione. Scegliamo di scrivere la funzione di ripartizione:

$$F_\theta(t) = \mathbb{P}_\theta\left[\widehat{\theta}(X_1, \dots, X_n) \leq t\right] = 1 - \left(1 - F_{X_1,\theta}(t)\right)^n$$

dove con $F_{X_i,\theta}$ indichiamo la funzione di ripartizione di X_i. Ricordando la relazione che c'è tra funzione di ripartizione e densità di una stessa variabile aleatoria, possiamo calcolare subito la funzione di ripartizione di ciascuna delle X_i:

$$F_{X_i,\theta}(t) = \int_{-\infty}^t L_1(\theta, x)\mathrm{d}x = \begin{cases} 0 & \text{se } t < \theta \\ 1 - \frac{\theta^3}{t^3} & \text{se } t \geq \theta \end{cases}$$

Possiamo concludere questo punto allora

$$F_{n,\theta}(t) = \begin{cases} 0 & \text{se } t < \theta \\ 1 - \frac{\theta^{3n}}{t^{3n}} & \text{se } t \geq \theta \end{cases} = \left(1 - \frac{\theta^{3n}}{t^{3n}}\right) \mathbf{1}_{[\theta,+\infty)}(t).$$

3. Per verificare se è distorto e se è asintoticamente distorto dobbiamo in ogni caso calcolare il valore attesto $\mathbb{E}_\theta\left[\widehat{\theta}_n(X_1, \dots, X_n)\right]$. Nel punto precedente abbiamo fatto dei conti che possiamo sfruttare. Infatti, la $\widehat{\theta}_n(X_1, \dots, X_n)$ ha densità $f_{n,\theta} = F'_{n,\theta}$ e di conseguenza

$$\mathbb{E}_\theta\left[\widehat{\theta}_n(X_1, \dots, X_n)\right] = \int_{\mathbb{R}} t f_{n,\theta}(t)\mathrm{d}t = \int_\theta^{+\infty} t \left(\frac{3n\theta^{3n}}{t^{3n+1}}\right)\mathrm{d}t = \frac{3n}{3n-1}\theta.$$

Qui di vede subito che non è corretto, ma lo è asintoticamente. Per avere uno stimatore corretto a partire da $\widehat{\theta}_n$ basta considerare

$$\widetilde{\theta}_n := \frac{3n-1}{3n}\widehat{\theta}_n.$$

4. Abbiamo visto nel punto precedente che $\widetilde{\theta}_n$ è corretto e quindi lo è pure asintoticamente. Se riusciamo a dimostrare che è anche
$$\sum_{n=1}^{+\infty} \mathrm{Var}_\theta \left[\widetilde{\theta}_n \left(X_1, \dots, X_n\right)\right] < +\infty.$$
allora abbiamo finito. Dobbiamo calcolare delle varianze, una per ogni $n \in \mathbb{N}$.
$$\begin{aligned}\mathrm{Var}_\theta \left[\widetilde{\theta}_n \left(X_1, \dots, X_n\right)\right] &= \mathbb{E}_\theta \left[\widetilde{\theta}_n \left(X_1, \dots, X_n\right)^2\right] - \theta^2 = \\ &= \left(\frac{3n-1}{3n}\right)^2 \mathbb{E}_\theta \left[\widehat{\theta}_n \left(X_1, \dots, X_n\right)^2\right] - \theta^2.\end{aligned}$$
Possiamo calcolare la media dell'ultimo membro visto che sopra abbiamo calcolato la funzione di ripartizione di $\widehat{\theta}_n(X_1, \dots, X_n)$:
$$\mathbb{E}_\theta \left[\widehat{\theta}_n \left(X_1, \dots, X_n\right)^2\right] = \int_{\mathbb{R}} t^2 f_{n,\theta}(t) \mathrm{d}t = \int_\theta^{+\infty} t^2 \left(\frac{3n\theta^{3n}}{t^{3n+1}}\right) \mathrm{d}t = \frac{3n}{3n-2}\theta^2.$$
Mettendo tutto insieme abbiamo
$$\mathrm{Var}_\theta \left[\widetilde{\theta}_n(X_1, \dots, X_n)\right] = \frac{(3n-1)^2}{3n(3n-2)}\theta^2 - \theta^2 = \frac{\theta^2}{3n(3n-2)}.$$
In conclusione la sommatoria delle varianze si comporta come la serie $\sum_{n=1} \frac{1}{9n^2}$, cioè converge.

5. Essendo $\widetilde{\theta}_n$ uno stimatore corretto, allora un intervallo di confidenza di livello $1-\alpha$, per $\alpha \in [0,1]$, è dato da
$$C_{n,\alpha}(x_1, \dots, x_n) := \left\{ \theta \in \Theta \,\middle|\, \left|\widetilde{\theta}_n(x_1, \dots, x_n) - \theta\right| \le \sqrt{\frac{\mathrm{Var}_\theta \left[\widetilde{\theta}_n(X_1, \dots, X_n)\right]}{\alpha}} \right\}.$$
Conosciamo già la varianza e quindi
$$\gamma_{n,\alpha} := \sqrt{\frac{\mathrm{Var}_\theta \left[\widetilde{\theta}_n(X_1, \dots, X_n)\right]}{\alpha}} = \frac{1}{\sqrt{3n(3n-2)\alpha}}\theta.$$
Rimane soltanto da risolvere una disequazione
$$\left|\theta - \widetilde{\theta}_n\right| \le \gamma_{n,\alpha}\theta \iff \left(1 - \gamma_{n,\alpha}\right)\theta^2 - 2\widetilde{\theta}_n\theta + \widetilde{\theta}_n^2 \le 0$$
Se $|\gamma_{n,\alpha}| < 1$, allora
$$\frac{\widetilde{\theta}_n - \gamma_{n,\alpha}}{1 - \gamma_{n,\alpha}^2} \le \theta \le \frac{\widetilde{\theta}_n + \gamma_{n,\alpha}}{1 - \gamma_{n,\alpha}^2}.$$
mentre se $|\gamma_{n,\alpha}| > 1$, allora
$$\theta \le \frac{\widetilde{\theta}_n - \gamma_{n,\alpha}}{1 - \gamma_{n,\alpha}^2} \text{ oppure } \theta \ge \frac{\widetilde{\theta}_n + \gamma_{n,\alpha}}{1 - \gamma_{n,\alpha}^2}.$$
In ogni caso, comunque fissato $\alpha \in [0,1]$, si ha che gli intervalli di confidenza sono
$$C_{n,\alpha}(x_1, \dots, x_n) = \left[\frac{\widetilde{\theta}_n - \gamma_{n,\alpha}}{1 - \gamma_{n,\alpha}^2}, \frac{\widetilde{\theta}_n + \gamma_{n,\alpha}}{1 - \gamma_{n,\alpha}^2}\right]$$
definitivamente per $n \to +\infty$, cioè da un certo n in poi le regioni di confidenza sono questi intervalli. □

Esercizio 2.2 (Problema 3, 12/09/2024)**.** Si consideri il modello a n prove ripetute avente modello di base $\mathbb{X}_1 := \mathbb{R}_+$, $\mathcal{X}_1 := \mathcal{B}\mathbb{R}_+$, $\Theta := \mathbb{R}_+$ e

$$\frac{\mathrm{d}\pi_\theta^{(1)}}{\mathrm{d}\lambda_1}(x) = \frac{3\theta^3}{(\theta + x)^4}$$

dove λ_1 è la misura unidimensionale di Lebesgue ristretta a $\mathbb{R}_+$.

1. Dimostrare che
$$S_n(x_1, \ldots, x_n) = \frac{2}{n}\sum_{j=1}^n x_j$$
è uno stimatore corretto di θ.
2. Dopo aver calcolato $\mathrm{Var}_\theta\left[S_n\left(X_1, \ldots, X_n\right)\right]$, dimostrare (usando i teoremi visti in classe) che S_n è debolmente consistente.
3. Discutere se vale la disuguaglianza
$$\mathrm{Var}_\theta\left[S_n\left(X_1, \ldots, X_n\right)\right] > \frac{1}{nI_1(\theta)}$$
dove I_1 è l'*informazione di Fisher* del modello di base.
4. Verificare (usando i teoremi di probabilità) che per ogni $\theta \in \Theta$
$$\sqrt{n}\left(S_n\left(X_1, \ldots, X_n\right) - \theta\right) \to N(0, F(\theta)) \text{ in distribuzione}$$
rispetto alla probabilità $\mathbb{P}_\theta$, dove $F : \Theta \to \mathbb{R}$ è una opportuna funzione continua.
5. Usare il risultato precedente per calcolare un intervallo di confidenza asintotico $C_n(X_1, \ldots, X_n) \subset \mathbb{R}_+$ di livello $1 - \alpha$ per il parametro θ.

Soluzione. 1. Come al solito, dobbiamo calcolare
$$\mathbb{E}_\theta\left[S_n\left(X_1, \ldots, X_n\right)\right]$$
dove $X_1, \ldots, X_n$ sono variabili aleatorie a valori in $\mathbb{X}_1$ i.i.d. con densità
$$f_\theta : \mathbb{X}_o \to \mathbb{R}$$
$$f_\theta(x) := \frac{\mathrm{d}\pi_\theta^{(1)}}{\mathrm{d}\lambda_1}(x) = \frac{3\theta^3}{(\theta + x)^4}$$
Eseguiamo il conto del valore atteso quindi
$$\mathbb{E}_\theta\left[S_n\left(X_1, \ldots, X_n\right)\right] = \frac{2}{n}\sum_{j=1}^n \mathbb{E}_\theta X_j = 2\mathbb{E}_\theta X_1$$
Il valore atteso dell'ultimo membro si può calcolare attraverso la densità
$$\mathbb{E}_\theta X_1 = \int_{\mathbb{X}_1} x f_\theta(x)\mathrm{d}x = 3\theta^3 \int_1^{+\infty} \frac{x}{(\theta + x)^4}\mathrm{d}x$$
Rimane da calcolare solo l'integrale all'ultimo membro:
$$\int_1^{+\infty} \frac{x}{(\theta + x)^4} = \int_1^{+\infty} \frac{1}{(\theta + x)^3}\mathrm{d}x - \theta\int_1^{+\infty} \frac{1}{(\theta + x)^4}\mathrm{d}x = \frac{1}{6\theta^2}.$$
Pertanto S_n è uno stimatore corretto di θ perché
$$\mathbb{E}_\theta\left[S_n\left(X_1, \ldots, X_n\right)\right] = \theta$$

2. Calcoliamo le varianze conoscendo la densità delle variabili aleatorie i.i.d. $X_1, \dots, X_n$ del punto precedente:

$$\operatorname{Var}_\theta\left[S_n\left(X_1, \dots, X_n\right)\right] = \frac{4}{n^2}\sum_{j=1}^{n}\operatorname{Var}_\theta X_j = \frac{4}{n}\operatorname{Var}_\theta X_1 = \frac{4}{n}\left(\mathbb{E}_\theta X_1^2 - \left(\mathbb{E}_\theta X_1\right)^2\right)$$

Dal punto precedente sappiamo che $\mathbb{E}_\theta X_1 = \frac{\theta}{2}$. Calcoliamo l'altro valore atteso:

$$\mathbb{E}X_1^2 = \int_{\mathbb{X}_1} x^2 f_\theta(x)\mathrm{d}x = 3\theta^3\int_1^{+\infty}\frac{x^2}{(\theta+x)^4}\mathrm{d}x.$$

L'integrale al secondo membro si può calcolare facilmente, per esempio ricordando che $x^2 = (\theta+x)^2 - \theta^2 - 2\theta x$. Alla fine si arriva a

$$\operatorname{Var}_\theta X_1 = \frac{3}{4}\theta^2$$

e quindi a

$$\operatorname{Var}_\theta\left[S_n\left(X_1, \dots, X_n\right)\right] = \frac{4}{n}\left(\theta^2 - \frac{\theta^2}{4}\right) = \frac{3}{n}\theta^2.$$

3. Il modello è regolare di classe C^2, quindi l'informazione di Fisher del modello di base si scrive come

$$I_1(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right].$$

Calcoliamo quindi il valore atteso al secondo membro:

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] &= \mathbb{E}_\theta\left[-\frac{3}{\theta^2} + \frac{4}{(\theta+X)^2}\right] = \\ &= -\frac{3}{\theta^2} + 4\mathbb{E}_\theta\left[\frac{1}{(\theta+X)^2}\right] = \\ &= -\frac{3}{\theta^2} + 4\int_1^{+\infty}\frac{1}{(\theta+x)^2}f_\theta(x)\mathrm{d}x = \\ &= -\frac{3}{\theta^2} + \frac{12}{5\theta^2} = -\frac{3}{5\theta^2}\end{aligned}$$

È facile verificare la disuguaglianza a questo punto.

4. Usiamo quanto ricavato nei punti precedenti: in particolare, se $\{X_n \mid n \in \mathbb{N}\}$ è una successione di variabili aleatorie i.i.d. con densità $f_\theta = \frac{\mathrm{d}\pi_\theta^{(1)}}{\mathrm{d}\lambda_1}$, allora per ogni $j \in \mathbb{N}$

$$\begin{aligned}\mathbb{E}_\theta X_j &= \frac{1}{2}\theta \\ \operatorname{Var}_\theta X_j &= \frac{3}{4}\theta^2\end{aligned}$$

Quindi per il Teorema Centrale del Limite

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}X_j - \frac{\theta}{2}\right) \to \mathcal{N}\left(0, \frac{3}{4}\theta^2\right).$$

da cui segue che

$$\sqrt{n}\left(\frac{2}{n}\sum_{j=1}^{n}X_j - \theta\right) \to \mathcal{N}\left(0, 3\theta^2\right)$$

che è quello che volevamo dimostrare. In particolare, $F(\theta) = 3\theta^2$.

5. Per il punto precedente, abbiamo visto che

$$Y_{n,\theta} := \frac{1}{\theta}\sqrt{\frac{n}{3}}\left(S_n\left(X_1,\dots,X_n\right)-\theta\right)$$

converge in distribuzione ad una $Z \sim \mathcal{N}(0,1)$. In particolare

$$\lim_{n\to+\infty} \mathbb{P}_\theta\left[-z_{\frac{\alpha}{2}} \le Y_{n,\theta} \le z_{\frac{\alpha}{2}}\right] = \mathbb{P}_\theta\left[-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right].$$

Sappiamo quanto vale l'ultimo membro: se $\Phi : \mathbb{R} \to [0,1]$ è la funzione di ripartizione di Z e z_α è definito tale che $\Phi(z_\alpha) = 1-\alpha$, allora

$$\lim_{n\to+\infty} \mathbb{P}_\theta\left[-z_{\frac{\alpha}{2}} \le Y_{n,\theta} \le z_{\frac{\alpha}{2}}\right] = 1-\alpha.$$

Possiamo quindi scegliere

$$\begin{aligned} C_n\left(x_1,\dots,x_n\right) &:= \left\{\theta\in\Theta \mid -z_{\frac{\alpha}{2}} \le Y_{n,\theta}\left(x_1,\dots,x_n\right) \le z_{\frac{\alpha}{2}}\right\} = \\ &= \left\{\theta\in\Theta \,\middle|\, \theta\left(1 - z_{\frac{\alpha}{2}}\sqrt{\frac{3}{n}}\right) \le S_n\left(x_1,\dots,x_n\right) \le \theta\left(1 - z_{\frac{\alpha}{2}}\sqrt{\frac{3}{n}}\right)\right\} \end{aligned}$$

□

Esercizio 2.3 (Problema 4, 29/01/2024)**.** Considerare il modello a n prove ripetute avente come modello base $\mathbb{X}_1 := [0,+\infty)$, $\mathcal{X}_1 := \mathcal{B}\mathbb{X}_1$, $\Theta := (0,+\infty)$ con funzione di verosomiglianza

$$L_1(\theta,x) := \frac{\mathrm{d}\pi_\theta^{(1)}}{\mathrm{d}\lambda_1}(x) := \frac{1}{4\sqrt{\theta x}}\mathbf{1}_{[0,4\theta]}(x)$$

dove λ_1 è la misura di Lebesgue ristretta a $\mathbb{X}_1$.

1. Scrivere lo stimatore di massima verosomiglianza $\widehat{\theta}_n : \mathbb{X}_n \to \Theta$ per il modello a n prove indipendenti.
2. Dimostrare che $\widehat{\theta}_n$ è distorto ma asintoticamente corretto. Dimostrare che $\widehat{\theta}_n$ è fortemente consistente.
3. Fissare $\theta_1 > 0$ e considerare il test

$$H_1 : \theta\in\Theta_1 := (0,\theta_1] \quad \text{e} \quad H_1 : \theta \in (\theta_1,+\infty).$$

   Calcolare la regione critica $G_{n,\alpha}$ per il test uniformemente più potente per le ipotesi H_1 e H_1.
4. Per n fissato, calcolare la funzione p-value della famiglia di test formulati nel punto precedente.
5. Calcolare la funzione potenza per il test $G_{n,\alpha}$, ovvero

$$\beta_{G_{n,\alpha}}(\theta) := \mathbb{P}_\theta\left[X\in G_{n,\alpha}\right].$$

Soluzione. 1. La funzione di verosomiglianza per il modello statistico a n prove indipendenti è

$$\begin{aligned} &L_n : \Theta\times\mathbb{X}_n \to [0,\infty] \\ &L_n(\theta,x_1,\dots,x_n) := \frac{1}{4^n\theta^{\frac{n}{2}}}\prod_{i=1}^n \frac{\mathbf{1}_{[0,4\theta]}(x_i)}{\sqrt{x_i}} \end{aligned}$$

Osservando che

$$L_n(\theta, x_1, \dots, x_n) = \begin{cases} 0 & \text{se } \max(x_1, \dots, x_n) > 4\theta \\ \frac{1}{4^n \theta^{\frac{n}{2}}} \prod_{i=1}^n \frac{1}{\sqrt{x_i}} & \text{altrimenti} \end{cases}$$

ovvero che

$$L_n(\theta, x_1, \dots, x_n) = \begin{cases} 0 & \text{se } \theta < \frac{\max(x_1, \dots, x_n)}{4} \\ \frac{1}{4^n \theta^{\frac{n}{2}}} \prod_{i=1}^n \frac{1}{\sqrt{x_i}} & \text{altrimenti} \end{cases}$$

si ha che la funzione $L_n(\cdot, x_1, \dots, x_n) : \Theta \to [0, +\infty]$ raggiunge il valore massimo in

$$\widehat{\theta}_n(x_1, \dots, x_n) = \frac{\max(x_1, \dots, x_n)}{4}.$$

2. Dobbiamo quindi calcolare il valore atteso

$$\mathbb{E}_\theta\left[\widehat{\theta}_n(X_1, \dots, X_n)\right]$$

dove $X_1, \dots, X_n$ sono variabili aleatorie a valori in $\mathbb{X}_1$ i.i.d. con densità $f_\theta := L_1(\theta, \cdot)$. Come prima cosa,

$$\mathbb{E}_\theta\left[\widehat{\theta}_n(X_1, \dots, X_n)\right] = \frac{1}{4}\mathbb{E}_\theta\left[\max(X_1, \dots, X_n)\right].$$

Sappiamo come calcolare la funzione di ripartizione di $\max(X_1, \dots, X_n)$:

$$F_{n,\theta}(t) := \left(\int_{-\infty}^t f_\theta(x)\mathrm{d}x\right)^n.$$

L'integrale al secondo membro è

$$\int_{-\infty}^t f_\theta(x)\mathrm{d}x = \frac{1}{4\sqrt{\theta}} \int_{(-\infty,t]\cap[0,4\theta]} \frac{1}{\sqrt{x}}\mathrm{d}x = \begin{cases} 0 & \text{se } t < 0 \\ \frac{\sqrt{\min(4\theta, t))}}{2\sqrt{\theta}} & \text{se } t \geq 0 \end{cases}$$

e quindi la funzione di ripartizione di $\max(X_1, \dots, X_n)$ è

$$F_{n,\theta}(t) := \begin{cases} 0 & \text{se } t < 0 \\ \min\left(1, \frac{t^{\frac{n}{2}}}{2^n \theta^{\frac{n}{2}}}\right) & \text{se } t \geq 0 \end{cases}$$

Possiamo ora facilmente scrivere la densità di $\max(X_1, \dots, X_n)$:

$$f_{n,\theta}(t) = F'_{n,\theta}(t)) = \begin{cases} 0 & \text{se } t < 0 \text{ oppure } t > 4\theta \\ \frac{n}{2^{n+1}\theta^{\frac{n}{2}}} t^{\frac{n}{2}} & \text{se } 0 \leq t \leq 4\theta \end{cases}$$

Il calcolo del valore atteso è si può quindi concludere

$$\mathbb{E}_\theta\left[\widehat{\theta}_n(X_1, \dots, X_n)\right] = \frac{n}{n+2}\theta.$$

Pertanto, lo stimatore di massima verosomiglianza è distorto, ma asintoticamente corretto.

3. Fissati $\theta_0 < \theta_1$, calcoliamo il *rapporto di verosomiglianza monotono*

$$
\begin{aligned}
\mathrm{LR}_n(x_1, \dots, x_n) &:= \frac{L_n(\theta_0, x_1, \dots, x_n)}{L_n(\theta_1, x_1, \dots, x_n)} = \\
&= \left(\frac{\theta_1}{\theta_0}\right)^{\frac{n}{2}} \frac{\mathbf{1}_{[0,4\theta_0]}(\max(x_1, \dots, x_n))}{\mathbf{1}_{[0,4\theta_1]}(\max(x_1, \dots, x_n))} \\
&= \left(\frac{\theta_1}{\theta_0}\right)^{\frac{n}{2}} \frac{\mathbf{1}_{[0,\theta_0]}\left(\widehat{\theta}_n(x_1, \dots, x_n)\right)}{\mathbf{1}_{[0,\theta_1]}\left(\widehat{\theta}_n(x_1, \dots, x_n)\right)}.
\end{aligned}
$$

che si può scrivere molto più semplicemente

$$
\mathrm{LR}_n(x_1, \dots, x_n) = \begin{cases} \left(\frac{\theta_1}{\theta_0}\right)^{\frac{n}{2}} & \text{se } \widehat{\theta}_n(x_1, \dots, x_n) \le \theta_0 \\ 0 & \text{altrimenti} \end{cases}
$$

Una regione critica basata sul rapporto di verosomiglianza monotono è

$$
G_{n,\alpha} := \left\{(x_1, \dots, x_n) \in \mathbb{X}_n \mid \widehat{\theta}_n(x_1, \dots, x_n) \ge c_{n,\alpha}\right\}
$$

dove $c_{n,\alpha}$ è da determinare in modo tale che

$$
\mathbb{P}_{\theta_0}\left[X \in G_{n,\alpha}\right] = \alpha.
$$

Possiamo calcolare questa probabilità usando la funzione di ripartizione di $\max(X_1, \dots, X_n)$, dove $X_1, \dots, X_n$ sono variabili aleatorie i.i.d. con distribuzione f_θ, e che abbiamo già calcolato:

$$
\begin{aligned}
\mathbb{P}_{\theta_0}\left[X \in G_{n,\alpha}\right] &= \mathbb{P}_{\theta_0}\left[\max(X_1, \dots, X_n) \ge 4c_{n,\alpha}\right] = \\
&= 1 - \mathbb{P}_{\theta_0}\left[\max(X_1, \dots, X_n) \le 4c_{n,\alpha}\right] = \\
&= 1 - F_{n,\theta_0}(4c_{n,\alpha}) = \\
&= 1 - \left(\frac{c_{n,\alpha}}{\theta_0}\right)^{\frac{n}{2}}
\end{aligned}
$$

Quindi basta scegliere $c_{n,\alpha} = \theta_0(1-\alpha)^{\frac{2}{n}}$ e una regione critica per il test uniformemente più potente è

$$
\begin{aligned}
G_{n,\alpha} &= \left\{(x_1, \dots, x_n) \in \mathbb{X}_n \mid \widehat{\theta}_n(x_1, \dots, x_n) \ge \theta_0(1-\alpha)^{\frac{2}{n}}\right\} = \\
&= \left\{(x_1, \dots, x_n) \in \mathbb{X}_n \mid \max(x_1, \dots, x_n) \ge 4\theta_0(1-\alpha)^{\frac{2}{n}}\right\}
\end{aligned}
$$

4. Vogliamo ora calcolare il p-value della famiglia di test $\{G_{n,\alpha} \mid \alpha \in [0,1]\}$, vale a dire la funzione $p_n : \mathbb{X}_n \to [0,1]$ definita da

$$
\begin{aligned}
p_n(x) &= \inf\left\{\alpha \in [0,1] \mid \widehat{\theta}_n(x) \ge \theta_0(1-\alpha)^{\frac{2}{n}}\right\} = \\
&= \max\left(0, 1 - \left(\frac{\widehat{\theta}_n(x)}{\theta_0}\right)^{\frac{n}{2}}\right).
\end{aligned}
$$

5. Nei precedenti punti abbiamo calcolato ciò che ci serve:

$$
\begin{aligned}
\beta_{G_{n,\alpha}}(\theta) &= \mathbb{P}_\theta\left[X \in G_{n,\alpha}\right] = \\
&= \mathbb{P}_\theta\left[\max(X_1, \dots, X_n) \ge 4\theta_0(1-\alpha)^{\frac{2}{n}}\right] = \\
&= 1 - \mathbb{P}_\theta\left[\max(X_1, \dots, X_n) \le 4\theta_0(1-\alpha)^{\frac{2}{n}}\right] = \\
&= 1 - F_{n,\theta}\left(4\theta_0(1-\alpha)^{\frac{2}{n}}\right) = \\
&= 1 - \left(\frac{\theta_0}{\theta}\right)^{\frac{n}{2}}(1-\alpha).
\end{aligned}
$$

□