Appunti di Statistica Matematica

Rita Giuliano e Maurizio Pratelli

Indice

1 Impostazione di un'analisi statistica: modelli statistici parametrici

Cominceremo con un esempio, che ci servirà da guida anche in altre situazioni.

Esempio 1.1 *L'analisi di qualità.*
Si vuole analizzare la qualità dei pezzi prodotti da una certa fabbrica. C'è una probabilità p (non nota) che il pezzo generico sia difettoso. Lo scopo dell'analisi è quello di ottenere informazioni sul valore di p. Consideriamo a questo scopo un campione di n pezzi (n è noto) e poniamo

$$X = \textit{numero di pezzi difettosi fra gli } n \textit{ scelti.}$$

Sappiamo che $X \sim \mathcal{B}(n,p)$, cioè

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Questa quantità dipende dal parametro incognito p ($0 < p < 1$).

Più precisamente, la situazione è la seguente. Abbiamo

(i) lo spazio campione Ω (cioè l'insieme di tutti i possibili risultati dell'esperimento) (nel nostro esempio abbiamo $\Omega = \{1, 2, ..., n\}$);

(ii) una $\sigma-$algebra $\mathcal{F} \subseteq \mathcal{P}(\Omega)$(nell'esempio è $\mathcal{F} = \mathcal{P}(\Omega)$);

(iii) una famiglia parametrizzata di leggi di probabilità su Ω $\{P^\theta, \theta \in \Theta\}$ (nell'esempio si tratta delle leggi binomiali $\mathcal{B}(n,p)$. In questo caso il parametro è p; in generale, cioè quando non saremo in situazioni specifiche per le quali si usano tipicamente altre notazioni, il parametro sarà indicato con θ).

Lo scopo di un'analisi statistica parametrica è quello di ottenere informazioni sul parametro θ; parlando genericamente, esse sono di due tipi fondamentali:

(a) si vuole stimare il parametro a partire dai dati;

(b) si vuole effettuare un *test* sul parametro.

Per capire meglio, torniamo all'esempio di partenza.

(a) intuitivamente una buona stima sembra essere la quantità $\frac{X}{n}$ (almeno per valori di n ragionevolmente grandi).

(b) Supponiamo di dover effettuare un test per sapere se $p \leq 0.1$ (ipotesi), e supponiamo che risulti $\frac{X}{n} = 0,07$. In questo caso l'ipotesi verrà accettata con sufficiente tranquillità. Con altrettanta tranquillità l'ipotesi verrebbe respinta se risultasse ad esempio $\frac{X}{n} = 0,35$. Ma come comportarsi se trovassimo $\frac{X}{n} = 0,11$? Vedremo in seguito la risposta.

Definizione 1.2 Si chiama *modello statistico (parametrico)* una terna $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$, dove

- Ω (detto *spazio campione*) è un insieme (da interpretare come l'insieme di tutti i possibili esiti dell'esperimento studiato);
- $\mathcal{F}$ è una $\sigma-$algebra di sottoinsiemi di Ω;
- $\{P^\theta, \theta \in \Theta\}$ è una famiglia di leggi di probabilità su $(\Omega, \mathcal{F})$, dipendente dal parametro $\theta \in \Theta$. Θ è l'*insieme dei parametri.*

L'appellativo di *parametrico* viene utilizzato per questo tipo di modelli tutte le volte che sia necessario distinguerli dai modelli non parametrici, di cui non parleremo in queste note.

Osservazione 1.3 La speranza e la varianza fatte rispetto alla legge P^θ verranno indicate rispettivamente con i simboli E^θ e Var^θ.

Per evitare problemi sulle questioni di trascurabilità (gli eventi trascurabili possono essere diversi a seconda della probabilità P^θ scelta) si dà la seguente

Definizione 1.4 Il modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$ si dice *dominato* se esiste una misura $\sigma-$finita su $(\Omega, \mathcal{F})$ tale che, per ogni $\theta \in \Theta$, risulti $P^\theta \ll \mu$. μ viene detta *misura dominante.*
Se le P^θ sono tutte tra loro equivalenti (e dunque, in particolare, sono tutte dominanti) il modello si dice *regolare.*

ALCUNI ESEMPI DI MODELLI STATISTICI.

Esempio 1.5 *Un modello regolare discreto.* Sia n un intero fissato e poniamo $\Omega = \{0, 1, \dots, n\}$. Su $\big(\Omega, \mathcal{P}(\Omega)\big)$ consideriamo le leggi binomiali $\mathcal{B}(n, \theta)$, dove $\theta \in (0, 1) =: \Theta$. Si ha cioè

$$P^\theta(\{k\}) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \qquad k = 0, 1, \dots, n.$$

Questo è un modello regolare, perché tutte le P^θ "caricano "i numeri $0, 1, \dots, n$.

Esempio 1.6 *Un modello discreto dominato ma non regolare.* Su $\big(\mathbb{N}, \mathcal{P}(\mathbb{N})\big)$ consideriamo le stesse leggi $\mathcal{B}(n, \theta)$, con $(n, \theta) \in (\mathbb{N}, (0, 1) =: \Theta$. Si tratta di un modello dominato: una misura dominante è una qualsiasi misura che "carica"tutti gli interi, per esempio la misura che "conta"i punti (cioè assegna massa unitaria ad ogni punto di $\mathbb{N}$). Tuttavia, evidentemente, questo modello non è regolare.

Esempio 1.7 *Un modello continuo dominato ma non regolare.* Su $\big(\mathbb{R}, \mathcal{P}(\mathbb{R})\big)$ le leggi uniformi su $[a, b]$, con $(a, b) \in \big\{(a, b) \in \mathbb{R}^2, a < b\big\} =: \Theta$ costituiscono un modello dominato ma non regolare, per gli stessi motivi dell'esempio precedente. Una misura dominante è per esempio $\lambda =$ misura di Lebesgue su $\mathbb{R}$.

Esempio 1.8 *Un modello continuo regolare.* Su $\big(\mathbb{R}, \mathcal{P}(\mathbb{R})\big)$ le leggi $\mathcal{N}(m, \sigma^2)$ (dove $\theta := (m, \sigma) \in \mathbb{R} \times \mathbb{R}^+ =: \Theta$) hanno tutte densità rispetto a $\lambda =$ misura di Lebesgue su $\mathbb{R}$; una densità è data dalla formula

$$f^{(m,\sigma)}(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(x-m)^2}{2\sigma^2}\Big), \quad x \in \mathbb{R}.$$

Pertanto le leggi normali su $\big(\mathbb{R}, \mathcal{P}(\mathbb{R})\big)$ sono un modello dominato (per esempio da λ). Poiché $f^{(m,\sigma)}$ è strettamente positiva per ogni $\theta := (m, \sigma)$, ciascuna legge $\mathcal{N}(m, \sigma^2)$ è equivalente a λ. Dunque questo modello è addirittura regolare.

Esempio 1.9 *Un modello non dominato.* Su $\big(\mathbb{R}, \mathcal{P}(\mathbb{R})\big)$ la famiglia di misure $\big\{\delta_x, x \in \mathbb{R}\big\}$ (dove δ_x= misura di Dirac in x) non sono un modello dominato. Questo fatto seguirà dal Teorema 1.14 (si veda l'Osservazione 1.17). Questo esempio mostra che la nozione di modello dominato non è banale.

Vale il seguente risultato (per la dimostrazione si veda ad es. [1], p.244 e ss.)

Teorema 1.10 *(di* **Radon–Nikodym***). Siano μ e ν due misure positive e $\sigma-$finite su $(\Omega, \mathcal{F})$, con $\nu \ll \mu$. Allora esiste una e una sola $f \in L^1(\mu)$ tale che*

$$\nu(E) = \int_E f \, \mathrm{d}\mu, \qquad \forall\, E \in \mathcal{F}.$$

Definizione 1.11 Nelle ipotesi del Teorema di Radon–Nikodym, ogni rappresentante di f si chiama *(versione della) derivata di Radon–Nikodym* di ν rispetto a μ, e si indica spesso con il simbolo

$$f = \frac{d\nu}{\mathrm{d}\mu}.$$

Si usa anche la scrittura $\nu = f.\mu$ (che si legge dicendo che ν *ha base* f *rispetto a* μ).

Osservazione 1.12 Siano μ_1, μ_2 e μ_3 tre misure $\sigma-$finite su $(\Omega, \mathcal{F})$ tali che $\mu_1 \ll \mu_2$ e $\mu_2 \ll \mu_3$; è facile vedere che

(i) $\mu_1 \ll \mu_3$;

(ii) una versione della derivata $\frac{d\mu_1}{\mathrm{d}\mu_3}$ è data μ_3-q.o. da

$$\frac{d\mu_1}{\mathrm{d}\mu_3} = \frac{d\mu_1}{\mathrm{d}\mu_2} \cdot \frac{d\mu_2}{\mathrm{d}\mu_3}.$$

Si noti che nella formula precedente l'espressione $d\mu_2$ si può formalmente "semplificare", trattandola come se fosse un numero. Questo artificio di scrittura sarà usato spesso nel seguito per velocizzare le formule.

Osservazione 1.13 Sia $\nu \ll \mu$, e poniamo

$$f = \frac{d\nu}{\mathrm{d}\mu}.$$

Allora $\nu \sim \mu$ se e solo se $f > 0$ μ-q.o. ed inoltre, ν-q.o.

$$\frac{d\mu}{\mathrm{d}\nu} = \frac{1}{f} = \frac{1}{\frac{d\nu}{\mathrm{d}\mu}},$$

Come si vede, anche in questo caso i termini $\mathrm{d}\mu$ e $\mathrm{d}\nu$ che compaiono nella formula possono essere trattati formalmente come numeri.

In un modello dominato, la misura dominante non è unica; tra tutte le misure dominanti, alcune hanno importanza particolare. Vale infatti il

Teorema 1.14 *(di Halmos–Savage) Sia* $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$ *un modello statistico dominato. Allora esistono una successione* $(\theta_n)_n$ *di elementi di* Θ *e una successione* $(a_n)_n$ *di numeri con*

$$a_n \geq 0, \qquad \sum_n a_n = 1$$

tali che la probabilità

$$P^\bullet := \sum_n a_n P^{\theta_n}$$

sia una misura dominante del modello. Una tale $P^\bullet$ *(combinazione convessa di una successione* $\big(P^{\theta_n}\big)_n$*) è detta "dominante privilegiata".*

Dimostrazione. (a) Ogni misura σ-finita è equivalente ad una misura di probabilità. Infatti esiste una successione di eventi $(\Omega_n)_{n\geq 1}$, con

$$\bigcup_{n=1}^{\infty} \Omega_n = \Omega, \qquad 0 < \mu(\Omega_n) < \infty, \quad \forall\, n.$$

Posto allora

$$f(x) = \sum_{n=1}^{\infty} \frac{1_{\Omega_n}(x)}{2^n \mu(\Omega_n)}$$

risulta $\int f \mathrm{d}\mu = 1$, e quindi la misura $\nu = f.\mu$ è una probabilità equivalente a μ.

Dunque nel seguito supporremo $\mu(\Omega) = 1$.

(b) Poniamo ora

$$p^\theta = \frac{dP^\theta}{\mathrm{d}\mu}, \qquad \mathcal{C} = \big\{C \in \mathcal{F}\big\} : \mu(C) > 0 \text{ e } \exists\, \overline{\theta} \in \Theta,\, p^{\overline{\theta}} > 0 \text{ su } C\big\}$$

$\mathcal{C}$ non è vuoto: infatti, fissato $\overline{\theta}$ e posto $\Omega^{\overline{\theta}} = \big\{p^{\overline{\theta}} > 0\big\}$, si vede subito che $\Omega^{\overline{\theta}} \in \mathcal{C}$ perché

$$P^{\overline{\theta}}(\Omega^{\overline{\theta}}) = \int_{\Omega^{\overline{\theta}}} p^{\overline{\theta}}\, \mathrm{d}\mu = \int_{\Omega^{\overline{\theta}}} p^{\overline{\theta}}\, \mathrm{d}\mu + \underbrace{\int_{(\Omega^{\overline{\theta}})^c} p^{\overline{\theta}}\, \mathrm{d}\mu}_{=0} = \int_\Omega p^{\overline{\theta}}\, \mathrm{d}\mu = \int_\Omega \frac{dP^{\overline{\theta}}}{\mathrm{d}\mu}\, \mathrm{d}\mu = \int_\Omega \mathrm{d}P^{\overline{\theta}} = 1,$$

e dunque $\mu(\Omega^{\overline{\theta}}) > 0$ perché $P^{\overline{\theta}} \ll \mu$.

(c) Consideriamo ora la famiglia

$$\mathcal{D} := \{\text{unioni finite di elementi di } \mathcal{C}\}$$

e poniamo

$$M = \sup_{D \in \mathcal{D}} \mu(D).$$

Dato che μ è una misura di probabilità, si ha $M \leq 1$.
Sia poi (D_k) una successione crescente di elementi di $\mathcal{D}$, tale che

$$\sup_k \mu(D_k) = M$$

e poniamo

$$D^\bullet = \bigcup_k D_k.$$

(d) $D^\bullet$ è unione numerabile di elementi di $\mathcal{C}$ (in quanto unione numerabile di unioni finite di elementi di $\mathcal{C}$), cioè esiste una successione (C_n) di elementi di $\mathcal{C}$ tale che

$$D^\bullet = \bigcup_n C_n.$$

Per definizione della classe $\mathcal{C}$, per ogni $n \geq 1$ esiste θ_n tale che $p^{\theta_n} > 0$ su C_n. Posto allora

$$p^\bullet = \sum_{n=1}^{\infty} \frac{1}{2^n} p^{\theta_n}; \qquad P^\bullet = p^\bullet.\mu,$$

proveremo che $P^\bullet$ è una probabilità dominante.
(e) Intanto, $P^\bullet \sim \mu$ su $D^\bullet$ poiché dal punto (d) segue che, per ogni $x \in D^\bullet$, esiste $C_{n_0} \in \mathcal{C}$ tale che $x \in C_{n_0}$, e dunque

$$p^\bullet(x) \geq \frac{1}{2^{n_0}} p^{\theta_{n_0}}(x) > 0.$$

(f) Vediamo ora che, $\forall\, \theta$, si ha $p^\theta = 0$ su $(D^\bullet)^c$, μ-q.o. Infatti, supponiamo per assurdo che esistano un $\theta \in \Theta$ ed un $A \subset (D^\bullet)^c$ tali che $\mu(A) > 0$ e $p^\theta > 0$ su A; dunque che $A \in \mathcal{C}$ per definizione di $\mathcal{C}$. Per ogni k si ha $A \cap D_k = \emptyset$ (perché $A \subset (D^\bullet)^c$) e, posto $E_k = A \cup D_k$, E_k appartiene a $\mathcal{D}$ (unioni finite di elementi di $\mathcal{C}$). Pertanto

$$\mu(A) + \mu(D_k) = \mu(E_k) \leq \sup_{D \in \mathcal{D}} \mu(D) = M$$

e, passando al sup rispetto a k, si ottiene

$$\mu(A) + M \leq M$$

relazione assurda in quanto $\mu(A) > 0$ e M è un numero finito (ricordare quanto detto nel punto (c)).

(g) I punti (e) e (f) implicano che ogni P^θ è assolutamente continua rispetto a $P^\bullet$. Sia infatti $E \in \mathcal{F}$ tale che $P^\bullet(E) = 0$. Allora, per ogni P^θ:

(i) Per il punto (f) si ha

$$P^\theta(E \cap (D^\bullet)^c) = \int_{E \cap (D^\bullet)^c} \mathrm{d}P^\theta = \int_{E \cap (D^\bullet)^c} p^\theta \mathrm{d}\mu = 0;$$

(ii) Adesso vediamo che è anche $P^\theta(E \cap D^\bullet) = 0$. Dato che P^θè assolutamente continua rispetto a μ, basta verificare che $\mu(E \cap D^\bullet) = 0$. Questo è vero in quanto

$$P^\bullet(E \cap D^\bullet) \leq P^\bullet(E) = 0$$

e dunque anche $\mu(E \cap D^\bullet) = 0$ perché $P^\bullet$ e μ sono equivalenti su $D^\bullet$ per il punto (e).

Si conclude pertanto che

$$P^\theta(E) = P^\theta(E \cap (D^\bullet)^c) + P^\theta(E \cap D^\bullet) = 0 + 0 = 0,$$

e cioè che $P^\bullet$ domina ogni P^θ. La dimostrazione è completa.

□

Esempio 1.15 Consideriamo il modello di un campione bernoulliano di taglia n, che è dato da

$$\Big(\{0,1\}^n, \mathcal{B}(\{0,1\}^n), \{P^\theta, \theta \in (0,1)\}\Big)$$

dove

$$P^\theta(\omega) = \theta^{\sum_i \omega_i}(1-\theta)^{n-\sum_i \omega_i}, \qquad \omega = (\omega_1, \dots, \omega_n.)$$

Sia $\theta_0 \in (0,1)$ fissato; P^{θ_0} è una dominante privilegiata, dato che l'unico elemento di $\mathcal{B}(\{0,1\}^n)$ che sia trascurabile rispetto a P^{θ_0} è l'evento vuoto, che è trascurabile anche rispetto a ogni altra P^θ. Una densità di P_θ rispetto a P^{θ_0} è data da

$$f^{\theta:\theta_0}(\omega) = \frac{(\theta)^{\sum_i \omega_i}(1-\theta)^{n-\sum_i \omega_i}}{(\theta_0)^{\sum_i \omega_i}(1-\theta_0)^{n-\sum_i \omega_i}} = \Big(\frac{\theta}{\theta_0}\Big)^{\sum_i \omega_i} \cdot \Big(\frac{1-\theta}{1-\theta_0}\Big)^{n-\sum_i \omega_i}.$$

Esercizio 1.16 *Trovare una dominante privilegiata*

(i) per un campione gaussiano $\mathcal{N}(m, \sigma^2)$;

(ii) per un campione di legge $\mathcal{U}(0,1)$ (uniforme sull'intervallo $(0,1)$).

Osservazione 1.17 (i) L'esempio e l'esercizio precedenti mostrano che, se il modello è regolare, ogni P^θ è del tipo della Proposizione 1.14: basta prendere $\theta_1 = \theta$, $a_1 = 1$.

(ii) *(Séguito dell'esempio 1.9).* Su $\big(\mathbb{R}, \mathcal{P}(\mathbb{R})\big)$ la famiglia di misure $\big\{\delta_x, x \in \mathbb{R}\big\}$ non sono un modello dominato. Infatti, se per assurdo lo fossero, esisterebbe una $P^\bullet$ dominante del tipo $P^\bullet = \sum_n a_n \delta_{x_n}$; tuttavia non è possibile che $P^\bullet$ domini <u>tutte</u> le δ_x: sia $x \notin \{x_1, x_2, x_3 \dots, \}$. Allora x è trascurabile per $P^\bullet$, ma non per δ_x.

(iii) Se $P^\bullet = \sum_n a_n P^{\theta_n}$, allora, per ogni $A \in \mathcal{F}$ si ha

$$P^\bullet(A) = \sum_n a_n P^{\theta_n}(A),$$

e quindi, per ogni g limitata,

$$\int g \, \mathrm{d}P^\bullet = \sum_n a_n \int g \, \mathrm{d}P^{\theta_n}.$$

Definizione 1.18 Sia $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$ un modello statistico dominato da una misura (σ-finita) μ. Si chiama *verosimiglianza del modello* ogni funzione $L : \Theta \times \Omega \to \mathbb{R}$ tale che, per ogni θ fissato, la funzione $L^\theta := L(\theta, \cdot) : \omega \mapsto L(\theta, \omega)$ sia una versione della densità di P^θ rispetto a μ.

Osservazione 1.19 Per ogni $\theta \in \Theta$ la funzione $L^\theta : \omega \mapsto L(\theta, \omega)$ è misurabile; inoltre, per ogni $A \in \mathcal{F}$ si ha

$$P^\theta(A) = \int_A \mathrm{d}P^\theta = \int_A \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu} \, \mathrm{d}\mu = \int_A L^\theta \, \mathrm{d}\mu \Big(= \int_A L(\theta, \omega) \, \mathrm{d}\mu(\omega) \Big).$$

Esempio 1.20 *Il modello del campione.*

Definizione 1.21 (a) Sia μ una misura di probabilità su $\mathbb{R}$. Si chiama *campione di taglia (o numerosità) n e legge μ* una famiglia $(X_1, X_2, \dots, X_n)$ di v.a. reali indipendenti, aventi tutte legge μ.

Una realizzazione si ottiene nel modo seguente. Prendiamo

$$\Omega = \mathbb{R}^n, \quad \mathcal{F} = \mathcal{B}(\mathbb{R}^n), \quad P = \underbrace{\mu \otimes \mu \otimes \cdots \otimes \mu}_{n \text{ volte}} = \mu^{\otimes n}.$$

Poniamo poi per ogni $i = 1, \ldots, n$ $X_i : \Omega \to \mathbb{R}$= proiezione i-esima, definita da

$$X_i : \omega = (x_1, \ldots, x_n) \mapsto x_i.$$

Allora, come si verifica facilmente, la legge di ogni X_i è μ e le X_i sono tra loro indipendenti.

(b) Assegnata una famiglia $\{\mu^\theta, \theta \in \Theta\}$ di leggi di probabilità su $\mathbb{R}$, chiamiamo *campione di taglia (o numerosità) n e legge μ^θ* una famiglia $(X_1, X_2, \ldots, X_n)$ di v.a. reali indipendenti, aventi tutte legge μ^θ. Con leggero abuso di linguaggio, chiameremo nello stesso modo il modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$ definito da

$$\Omega = \mathbb{R}^n, \quad \mathcal{F} = \mathcal{B}(\mathbb{R}^n), \quad P^\theta = \underbrace{\mu^\theta \otimes \mu^\theta \otimes \cdots \otimes \mu^\theta}_{n \text{ volte}} = \big(\mu^\theta\big)^{\otimes n}.$$

Se le leggi $\{\mu^\theta, \theta \in \Theta\}$ sono discrete, il modello precedente può essere semplificato in modo ovvio. Si lasciano i dettagli per esercizio.

Osservazione 1.22 Sia $(X_1, X_2, \ldots, X_n)$ un campione di taglia n e legge μ^θ, $\theta \in \Theta$. Dunque si tratta del modello precedente,

$$\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \big(\mu^\theta\big)^{\otimes n}\big).$$

Sia μ è una misura su $\mathbb{R}$ tale che $\mu^\theta \ll \mu$ per ogni $\theta \in \Theta$, e poniamo

$$f^\theta(x) := \frac{\mathrm{d}\mu^\theta}{\mathrm{d}\mu}(x).$$

Allora $P^\theta = \big(\mu^\theta\big)^{\otimes n} \ll \mu^{\otimes n}$ (cioè anche il modello del campione è dominato) ed inoltre

$$L(\theta, \omega) = \frac{\mathrm{d}P^\theta}{\mathrm{d}\big(\mu\big)^{\otimes n}}(\omega) = \prod_{i=1}^n f^\theta(\omega_i).$$

Ovvero la verosimiglianza del campione è il prodotto tensoriale delle verosimiglianze dei singoli elementi del campione. Si lascia la verifica per esercizio.

Esempio 1.23 *Il modello per una catena di Markov.*

Supponiamo di avere una catena di Markov sullo spazio misurabile $(E, \mathcal{E})$, di legge iniziale ρ assegnata e operatore di transizione

$$\Pi^\theta(x, A) = \int_A \ell^\theta(x, y) \Pi(x, \mathrm{d}y), \qquad x \in E,\ A \in \mathcal{E},$$

dove Π è un operatore di transizione fissato di $(E, \mathcal{E})$ in $(E, \mathcal{E})$ e $\{\ell^\theta(\cdot, \cdot)\}_{\theta \in \Theta}$ è una famiglia di verosimiglianze.

Una notazione. Dato un operatore di transizione $(x, A) \mapsto \Pi(x, A)$ di $(E, \mathcal{E})$ in $(E, \mathcal{E})$, poniamo $\Pi^{\otimes 1} = \Pi$ e per induzione definiamo

$$\Pi^{\otimes(n+1)}(x, B) = \int \Pi(x, \mathrm{d}y) \int \Pi^{\otimes n}(y, \mathrm{d}z) 1_B(y, z), \qquad x \in E,\ B \in \mathcal{E}^{\otimes(n+1)}.$$

Allora $\Pi^{\otimes n}$ è un operatore di transizione di $(E, \mathcal{E})$ in $(E^n, \mathcal{E}^{\otimes n})$, e la misura di probabilità (definita su $\mathcal{E}^{\otimes n}$)

$$B \mapsto \Pi^{\otimes n}(x, B)$$

rappresenta la legge del vettore $(X_1, \dots, X_n)$ condizionata a $X_0 = x$; in altre parole si ha

$$P\big((X_1, \dots, X_n) \in B | X_0 = x\big) = \Pi^{\otimes n}(x, B).$$

Pertanto, se la legge di X_0 è ρ, la legge congiunta del vettore $(X_0, \dots, X_n)$ è data da

$$C \mapsto \int \rho(\mathrm{d}x)\Pi^{\otimes n}(x, \mathrm{d}y)1_C(x, y) =: \big(\rho \otimes \Pi^{\otimes n}\big)(C), \qquad C \in \mathcal{E}^{\otimes(n+1)}.$$

Torniamo alla situazione iniziale. Se si osservano i primi $n+1$ passi, $X_0, X_1, \dots, X_n$ possiamo prendere come modello statistico

$$\big(E^{n+1}, \mathcal{E}^{\otimes(n+1)}, \big\{\big(\rho \otimes (\Pi^\theta)^{\otimes n}\big), \theta \in \Theta\big\}\big).$$

Si vede facilmente che si tratta di un modello dominato: una misura dominante è $\big(\rho \otimes \Pi^{\otimes n}\big)$, e una versione della verosimiglianza è $(x = (x_0, \dots, x_n))$

$$L(\theta, x) = \ell^\theta(x_0, x_1) \cdot \ell^\theta(x_1, x_2) \cdot \dots \cdot \ell^\theta(x_{n-1}, x_n).$$

Infatti (caso $n = 2$ per semplicità) per $C \in \mathcal{E}^{\otimes 3}$ si ha $(y = (y_1, y_2))$

$$\begin{aligned}
\big(\rho \otimes (\Pi^\theta)^{\otimes 2}\big)(C) &= \int \rho(\mathrm{d}x)(\Pi^\theta)^{\otimes 2}(x, \mathrm{d}y)1_C(x, y) \\
&= \int \rho(\mathrm{d}x) \int \Pi^\theta(x, \mathrm{d}y_1) \int \Pi^\theta(y_1, \mathrm{d}y_2)1_C(x, y_1, y_2) \\
&= \int \rho(\mathrm{d}x) \int \ell^\theta(x, y_1)\Pi(x, \mathrm{d}y_1) \int \ell^\theta(y_1, y_2)\Pi(y_1, \mathrm{d}y_2)1_C(x, y_1, y_2).
\end{aligned}$$

2 La nozione di riassunto esaustivo

Sia $(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\})$ un modello statistico.

Definizione 2.1 Si chiama *statistica* (definita sul modello statistico) ogni applicazione misurabile T di $(\Omega, \mathcal{F})$ in uno spazio misurabile $(E, \mathcal{E})$ che non dipenda da θ.

Ad esempio, nel modello di un campione è una statistica la v.a. $\overline{X} := \frac{X_1 + \dots + X_n}{n}$ (detta *media campionaria*), mentre non è una statistica ad esempio la v.a. $X_1 + \theta$.

Diamo ora una idea intuitiva della nozione di *riassunto esaustivo*, che definiremo tra poco, utilizzando ancora l'esempio del controllo di qualità; si estraggono in modo indipendente l'uno dall'altro n pezzi prodotti, e poniamo, per $i = 1, \dots, n$

$$X_i = \begin{cases} 1 & \text{se l'}i\text{-esimo pezzo è difettoso} \\ 0 & \text{se no.} \end{cases}$$

Sia p (non nota) la probabilità che il generico pezzo sia difettoso. $(X_1, \dots, X_n)$ è il nostro campione, e ha legge $\mathcal{B}(1, p)$. Poniamo poi

$$X_1 + \dots + X_n = \text{numero di pezzi difettosi.}$$

Se vogliamo stimare p, è chiaro che basta sapere qual è il valore di $X_1 + \cdots + X_n$, non è necessario conoscere il vettore $(X_1, \ldots, X_n)$ (che significherebbe sapere se il primo pezzo era difettoso o no, lo stesso per il secondo, e così via). In altri termini la statistica

$$T = T(X_1, \ldots, X_n) := X_1 + \cdots + X_n$$

contiene tutta l'informazione contenuta nel modello: conoscere separatamente i valori di $X_1, \ldots, X_n$ non ci direbbe di più.

Cerchiamo di arrivare ad una buona definizione matematica. Sia assegnato il modello statistico $(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\})$. Osservare $T(X_1, \ldots, X_n)$ equivale ad osservare la σ-algebra $\mathcal{B}$ generata da T, $\mathcal{B} = \sigma(T)$. In generale risulta $\mathcal{B} \subset \mathcal{F}$; il nostro scopo è (o meglio sarebbe) individuare P^θ su tutta la σ-algebra $\mathcal{F}$ conoscendo solo i valori che essa assume su $\mathcal{B}$ (o, in modo equivalente, per esprimersi in termini di T, conoscendo $P^\theta(T \in I)$, al variare di I sottoinsieme misurabile di $\mathbb{R}$, e cioè conoscendo la legge di T secondo P^θ).

Ora, in generale, assegnata una sotto-σ-algebra $\mathcal{B}$ di $\mathcal{F}$, P^θ è individuata dalla sua traccia su $\mathcal{B}$ (cioè dai valori che essa assume sugli elementi $B \in \mathcal{B}$, $P^\theta(B)$) e dalle speranze, condizionate a $\mathcal{B}$, delle v.a. limitate: infatti, per ogni $A \in \mathcal{F}$, risulta

$$P^\theta(A) = E^\theta[1_A] = E^\theta\big[E^\theta[1_A|\mathcal{B}]\big] = \int E^\theta[1_A|\mathcal{B}]\mathrm{d}P^\theta = \int E^\theta[1_A|\mathcal{B}]\mathrm{d}P^\theta|_{\mathcal{B}}. \tag{1}$$

Dunque, se riusciamo a trovare una versione della speranza condizionale $E^\theta[1_A|\mathcal{B}]$ che non dipende da θ, per individuare P^θ è sufficiente conoscere $P^\theta|_{\mathcal{B}}$.

Esempio 2.2 Di nuovo il controllo di qualità. Prendiamo $n = 2$ per semplicità. Vogliamo mostrare che la statistica $T = X_1 + X_2$ è esaustiva. In questo caso si ha $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$ e $\mu^\theta = \mathcal{B}(1, \theta)$. Faremo i conti nel solo caso particolare

$$A = \big\{(X_1, X_2) \in [1/2, 3/2] \times [-1/2, 1/2]\big\} = \big\{X_1 = 1, X_2 = 0\big\}.$$

La σ-algebra generata da $X_1 + X_2$, cioè

$$\mathcal{B} = \big\{B : B = \{X_1 + X_2 \in I\}, I \text{ misurabile } \subseteq \mathbb{R}\big\}$$

è generata dai tre eventi $B_0 = \{X_1 + X_2 = 0\}$, $B_1 = \{X_1 + X_2 = 1\}$,,$B_2 = \{X_1 + X_2 = 2\}$.

Poniamo $(\omega = (\omega_1, \omega_2))$

$$Y(\omega) = \begin{cases} \frac{1}{2} & \text{per } \omega \in B_1 \\ 0 & \text{altrimenti.} \end{cases}$$

Vogliamo far vedere che $Y = E^\theta[1_A|\mathcal{B}]$. Y è ovviamente misurabile, dunque sarà la speranza condizionale cercata se, $\forall\, B \in \mathbb{B}$, risulta

$$\int_B 1_A \mathrm{d}P^\theta = \int_B Y \mathrm{d}P^\theta.$$

Basterà verificare la relazione precedente per i tre generatori B_0, B_1 e B_2.

(i) Per B_0. Si ha

$$\int_{B_0} 1_A \mathrm{d}P^\theta = P^\theta(A \cap B_0) = P^\theta(X_1 = 1, X_2 = 0, X_1 + X_2 = 0) = 0 = \int_B Y \mathrm{d}P^\theta.$$

(ii) Per B_2 la verifica è identica.

(iii) Per B_1. In questo caso risulta

$$\int_{B_1} 1_A \mathrm{d}P^\theta = P^\theta(A \cap B_1)$$
$$= P^\theta(X_1 = 1, X_2 = 0, X_1 + X_2 = 1) = P^\theta(X_1 = 1, X_2 = 0) = \theta(1-\theta);$$

$$\int_{B_1} Y \mathrm{d}P^\theta = \frac{1}{2} P^\theta(B_1) = \frac{1}{2} \cdot 2\theta(1-\theta) = \theta(1-\theta).$$

Concludiamo questo esempio con la verifica che la speranza condizionale, cioè la v.a. Y, verifica la relazione (1). Si trova infatti

$$\int Y \mathrm{d}P^\theta|_{\mathcal{B}} = \frac{1}{2} P^\theta(B_1) = \theta(1-\theta) = P^\theta(A).$$

Siamo pronti per dare la definizione formale di *riassunto esaustivo*:

Definizione 2.3 È assegnato il modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$, dominato da una misura μ. La statistica $T : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ si dice *riassunto esaustivo* (o *statistica esaustiva* o anche *sufficiente*, *sufficient statistic* in inglese) se, per ogni v.a. Y definita su $(\Omega, \mathcal{F})$, a valori reali e limitata, esiste, definita μ-q.o., una versione della speranza condizionale $E^\theta[Y|T]$ che non dipenda da θ. Se T è cosiffatta, scriveremo $E^\diamond[Y|T]$ invece che $E^\theta[Y|T]$.

Osservazione 2.4 Vediamo un caso particolare importante. Sia $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$del tipo $\Omega = \mathbb{R}^n$; $X_i = i$-esima proiezione, $\mathcal{F} = \sigma$-algebra generata da $X = (X_1, \ldots, X_n)$, $P^\theta = (\mu^\theta)^{\otimes n}$. In altre parole X è un campione di taglia n e legge μ^θ. Per il criterio di misurabilità di Doob, ogni v.a. Y che sia $\mathcal{F}$-misurabile è del tipo $Y = \phi(X_1, \ldots, X_n)$, con $\phi : \mathbb{R}^n \to \mathbb{R}$ misurabile. Dunque, dire che T è esaustiva equivale a dire che, qualunque sia ϕ, la speranza condizionale $E^\theta[\phi(X_1, \ldots, X_n)|T]$ non dipende da θ o, ciò che è lo stesso, che la legge condizionale di $(X_1, \ldots, X_n)$, data T, non dipende da θ. Questa seconda definizione è quella preferita nei libri di tipo applicativo.

Esempio 2.5 (Calcolo di una statistica sufficiente nel caso particolare dell'Osservazione precedente). Sia $(X_1, \ldots, X_n)$ un campione di legge Π_θ, $\theta > 0$. Mostrare che la statistica $T = T(X_1, \ldots, X_n) = X_1 + \cdots + X_n$ è esaustiva.
Soluzione. Calcolare la legge condizionale del vettore $(X_1, \ldots, X_n)$, data T, significa calcolare $P(X_1 = x_1, \ldots, X_n = x_n | T = t)$ al variare di $(x_1, \ldots, x_n) \in \mathbb{N}^n$ e $t \in \mathbb{N}$. Si ha facilmente

$$P(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)}$$
$$= \begin{cases} 0 & \text{se } t \neq x_1 + \cdots + x_n \\ \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)} & \text{se } t = x_1 + \cdots + x_n; \end{cases}$$

continuiamo il calcolo nel secondo caso: poiché le X_i sono tra loro indipendenti e di legge Π_θ, si ha $T \sim \Pi_{n\theta}$; di conseguenza

$$\frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)} = \frac{\left(\frac{\theta^{x_1}}{x_1!}\mathrm{e}^{-\theta}\right) \cdots \left(\frac{\theta^{x_n}}{x_n!}\mathrm{e}^{-\theta}\right)}{\frac{n^t \theta^t}{t!}\mathrm{e}^{-n\theta}} = \frac{\frac{\theta^{x_1 + \cdots + x_n}}{x_1! \cdots x_n!}\mathrm{e}^{-n\theta}}{\frac{n^t \theta^t}{t!}\mathrm{e}^{-n\theta}} = \frac{t!}{x_1! \cdots x_n! n^t},$$

dato che $t = x_1 + \cdots + x_n$. Come si vede, il risultato non dipende da θ, e quindi si conclude che T è un riassunto esaustivo.

È chiaro che la definizione non è comoda per decidere se una statistica è esaustiva oppure no: il tipo di calcolo effettuato nell'esempio 2.5 necessita di conoscere in anticipo la statistica candidata per essere esaustiva. Il teorema che segue costituisce un criterio operativo, che in particolare fornisce anche la statistica candidata.

Teorema 2.6 (DI NEYMAN–FISHER, O DI FATTORIZZAZIONE) *Sia* $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$ *un modello statistico dominato; siano* μ *una misura dominante,* $T : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ *una statistica. Allora sono fatti equivalenti*

(a) T *è un riassunto esaustivo;*

(b) esistono una funzione misurabile h *(non dipendente da* θ*) e, per ogni* θ*, una funzione misurabile* g^θ *tali che*

$$L(\theta, \omega) := \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}(\omega) = g^\theta\big(T(\omega)\big) \cdot h(\omega), \qquad \mu - q.o.;$$

(c) se P *è una dominate privilegiata, allora , per ogni* θ*,esiste una funzione misurabile* g^θ *tale che*

$$L(\theta, \omega) := \frac{\mathrm{d}P^\theta}{\mathrm{d}P}(\omega) = g^\theta\big(T(\omega)\big), \qquad P - q.o.,$$

(cioè in questo caso si può prendere $h \equiv 1$*, ovvero* $\frac{\mathrm{d}P^\theta}{\mathrm{d}P}$ *coincide* P*-q.o. con una variabile della forma* $g^\theta \circ T$*.*

DIMOSTRAZIONE. Mostreremo che (a)⇒(c) ⇒(b) ⇒(a).

(a)⇒(c). Osserviamo prima di tutto che, per ogni v.a. limitata X, se $E^\diamond[X|T]$ (non dipendente da θ!) è una versione della speranza condizionale di X rispetto a T relativamente ad ogni P^θ, allora $E^\diamond[X|T]$ è una versione della speranza condizionale di X rispetto a T anche relativamente a P (che è una dominante privilegiata). Infatti, per ogni $B \in \sigma(T)$($\sigma(T) = \sigma$-algebra generata da T), si ha

$$\int_B X \,\mathrm{d}P = \sum_n a_n \int_B X \,\mathrm{d}P^{\theta_n} = \sum_n a_n \int_B E^{\theta_n}[X|T]\mathrm{d}P^{\theta_n}$$
$$= \sum_n a_n \int_B E^\diamond[X|T]\mathrm{d}P^{\theta_n} = \int_B E^\diamond[X|T]\mathrm{d}\Big(\sum_n a_n P^{\theta_n}\Big) = \int_B E^\diamond[X|T]\mathrm{d}P,$$

per l'Osservazione 1.17 (iii).

Sia ora $\frac{\mathrm{d}P^\theta}{\mathrm{d}P} = L(\theta)$ una versione della verosimiglianza (rispetto a P), e mostriamo che la v.a.

$$g^\theta(T) := E^P[L(\theta)|T]$$

è anch'essa una versione della verosimiglianza. Infatti, per ogni v.a. X limitata, si ha

$$E^\theta[X] = E^\theta\big[E^\theta[X|T]\big] = \int E^\theta[X|T]\mathrm{d}P^\theta = \int E^\theta[X|T]\frac{\mathrm{d}P^\theta}{\mathrm{d}P}\mathrm{d}P = E^P\big[E^\theta[X|T]L(\theta)\big]$$
$$= E^P\big[E^\diamond[X|T]L(\theta)\big] = E^P\Big[E^P\big[E^\diamond[X|T]L(\theta)|T\big]\Big] = E^P\Big[E^\diamond[X|T]\underbrace{E^P\big[L(\theta)|T\big]}_{=g^\theta(T)}\Big]$$
$$= E^P\Big[E^\diamond[X|T]g^\theta(T)\Big] \underbrace{=}_{g^\theta(T)misur.} E^P\Big[E^\diamond[g^\theta(T)X|T]\Big] \underbrace{=}_{oss.prec} E^P\Big[E^P[g^\theta(T)X|T]\Big]$$
$$= E^P[g^\theta(T)X].$$

In altri termini, per ogni v.a. X si ha

$$\int X \mathrm{d}P^\theta = \int X g^\theta(T) \mathrm{d}P,$$

il che significa proprio che $g^\theta(T)$ è una versione di $\frac{\mathrm{d}P^\theta}{\mathrm{d}P}$, e il punto (c) è dimostrato.

(c) $\Rightarrow$ (b). Dato che μ domina tutte le P^θ, e di conseguenza anche P, per l'ipotesi si ha

$$\frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}(\omega) = \frac{\mathrm{d}P^\theta}{\mathrm{d}P}(\omega) \cdot \frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega) = g^\theta\big(T(\omega)\big) \cdot h(\omega),$$

dove si è posto

$$h := \frac{\mathrm{d}P}{\mathrm{d}\mu}.$$

(b) $\Rightarrow$ (a). Premettiamo un

Lemma 2.7 (Formula di Bayes). *Su $(\Omega, \mathcal{F})$ siano Q_1 e Q_2 due probabilità , con $Q_1 \ll Q_2$. Sia $\mathcal{B}$ una sotto-σ-algebra di $\mathcal{F}$. Posto $Z = \frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}$, per ogni v.a. positiva X si ha*

$$E^{Q_1}[X|\mathcal{B}] = \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]}, \qquad Q_1 - q.o.$$

Continuiamo con l'implicazione (b) $\Rightarrow$ (a). La dimostrazione del Lemma è rimandata al termine della dimostrazione del Teorema. Prima di tutto vediamo che non è restrittivo supporre che la misura dominante μ sia una probabilità . Infatti, come nella Proposizione 1.14 costruiamo la funzione f strettamente positiva tale che $Q = f.\mu$ sia una probabilità equivalente a μ. Si tratta allora di far vedere che anche la Q verifica l'ipotesi (b). Questo è vero, in quanto, poiché l'ipotesi (b) è valida per μ, si ha

$$\frac{\mathrm{d}P^\theta}{\mathrm{d}Q} = \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu} \cdot \frac{\mathrm{d}\mu}{\mathrm{d}Q} = \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu} \cdot \frac{1}{f} = g^\theta\big(T\big) \cdot \frac{h}{f} =: g^\theta\big(T\big) \cdot \kappa.$$

Poiché dunque Q è una probabilità, possiamo applicare il Lemma 2.7 alla coppia di probabilità $Q_1 = P^\theta$ e $Q_2 = Q$ (con $\mathcal{B} = \sigma(T)$). Si trova

$$E^\theta[X|T] = \frac{E^Q[X g^\theta(T)\kappa|T]}{E^Q[g^\theta(T)\kappa|T]} = \frac{g^\theta(T) \cdot E^Q[X\kappa|T]}{g^\theta(T) \cdot E^Q[\kappa|T]} = \frac{E^Q[X\kappa|T]}{E^Q[\kappa|T]};$$

dato che quest'ultima variabile aleatoria non dipende da θ, si conclude che T è un riassunto esaustivo.

□

Osservazione 2.8 Conseguenza del Teorema di fattorizzazione è che, nel caso di un modello dominato, se T è esaustiva e $T = \phi(S)$, allora anche S è esaustiva. Questa proprietà non vale se il modello non è dominato.

Dimostrazione del Lemma 2.7. Prima di tutto osserviamo che l'evento

$$A = \{E^{Q_2}[Z|\mathcal{B}] = 0\}$$

è trascurabile rispetto a Q_1 (e dunque il secondo membro della relazione precedente è ben definito). Infatti, A è ovviamente $\mathcal{B}$- misurabile e quindi, per la definizione di speranza condizionale $E^{Q_2}[Z|\mathcal{B}]$ si ha

$$Q_1(A) = \int_A \mathrm{d}Q_1 = \int_A \frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}\,\mathrm{d}Q_2 = \int_A Z\,\mathrm{d}Q_2 = \int_A E^{Q_2}[Z|\mathcal{B}]\,\mathrm{d}Q_2 = 0.$$

Poniamo ora per semplicità

$$U = \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]}.$$

U è chiaramente $\mathcal{B}$- misurabile; per terminare la dimostrazione basta allora far vedere che, per ogni $B \in \mathcal{B}$, si ha

$$\int_B U\,\mathrm{d}Q_1 = \int_B X\,\mathrm{d}Q_1.$$

Infatti

$$\begin{aligned}
&\int_B U\,\mathrm{d}Q_1 = \int_B \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]}\,\mathrm{d}Q_1 = \int 1_B \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]}\,\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}\,\mathrm{d}Q_2 = \int 1_B \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]}\,Z\,\mathrm{d}Q_2 \\
&= E^{Q_2}\Big[1_B \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]}\,Z\Big] = E^{Q_2}\Big[E^{Q_2}\Big[1_B \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]}\,Z\Big|\mathcal{B}\Big]\Big] = E^{Q_2}\Big[1_B \frac{E^{Q_2}[XZ|\mathcal{B}]}{E^{Q_2}[Z|\mathcal{B}]} E^{Q_2}[Z|\mathcal{B}]\Big] \\
&= E^{Q_2}\Big[1_B E^{Q_2}[XZ|\mathcal{B}]\Big] = E^{Q_2}\Big[E^{Q_2}[1_B XZ|\mathcal{B}]\Big] = E^{Q_2}[1_B XZ] = \int_B XZ\,\mathrm{d}Q_2 = \int_B X\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}\,\mathrm{d}Q_2 \\
&= \int_B X\,\mathrm{d}Q_1.
\end{aligned}$$

□

Osservazione 2.9 Il Lemma 2.7 è una generalizzazione della formula di Bayes: sia infatti $\mathcal{B} = \{\emptyset, \Omega\}$ la σ- algebra banale (la speranza condizionale rispetto a $\mathcal{B}$ non è altro che la speranza), $B \in \mathcal{F}$ con $P(B) > 0$ e siano $A_1, \ldots, A_n \in \mathcal{F}$, con $P(A_k) > 0$ per ogni k, una partizione di Ω; poniamo

$$Q_1(\cdot) = P(\cdot|B), \qquad Q_2(\cdot) = P(\cdot) = \sum_{k=1}^n P(\cdot|A_k)P(A_k).$$

Si ha evidentemente $Q_1 \ll Q_2$ ed inoltre una versione della densità $\frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}$ è la funzione

$$Z = \frac{1_B}{P(B)} = \frac{1_B}{\sum_{k=1}^n P(B|A_k)P(A_k)}.$$

Infatti, per ogni $C \in \mathcal{F}$

$$\begin{aligned}
&Q_1(C) = P(C|B) = \frac{P(C\cap B)}{P(B)} = \frac{\int 1_{C\cap B}\mathrm{d}\,P}{P(B)} = \frac{\int 1_C 1_B \mathrm{d}P}{P(B)} = \frac{\int_C 1_B\,\mathrm{d}P}{P(B)} \\
&= \int_C Z\,\mathrm{d}P = \int_C Z\,\mathrm{d}Q_2.
\end{aligned}$$

Si ha poi

$$E^{Q_1}[1_{A_j}] = P(A_j|B); \qquad E^{Q_2}[Z] = E^P\Big[\frac{1_B}{P(B)}\Big] = 1$$

e, per ogni j,

$$E^{Q_2}[1_{A_j}Z] = E^P\Big[\frac{1_{A_j}1_B}{\sum_{k=1}^n P(B|A_k)P(A_k)}\Big] = \frac{P(A_j\cap B)}{\sum_{k=1}^n P(B|A_k)P(A_k)} = \frac{P(B|A_j)P(A_j)}{\sum_{k=1}^n P(B|A_k)P(A_k)}.$$

Usando la formula del Lemma con $X = 1_{A_j}$, si trova dunque

$$P(A_j|B) = E^{Q_1}[1_{A_j}] = \frac{E^{Q_2}[1_{A_j} Z]}{E^{Q_2}[Z]} = \frac{P(B|A_j)P(A_j)}{\sum_{k=1}^n P(B|A_k)P(A_k)},$$

ovvero la ben nota formula di Bayes.

ALCUNI ESEMPI DI STATISTICHE ESAUSTIVE.

(a) Controllo di qualità. Poniamo $\Omega = \{0,1\}^n$, $\mathcal{F} = \mathcal{P}(\Omega)$ e, per ogni $i = 1, \dots, n$

$$X_i = \begin{cases} 1 & \text{se il pezzo } i\text{-esimo è difettoso} \\ 0 & \text{se no.} \end{cases}$$

In altre parole, se $\omega \in \Omega$, ($\omega = (\omega_1, \dots, \omega_n)$ con $\omega_i \in \{0,1\}$), si pone $X_i(\omega) = \omega_i$, cioè $X_i : \Omega \to \{0,1\}$ non è altro che l'i-esima proiezione. Poniamo poi

$$P^\theta(\omega) = \theta^{\sum_{i=1}^n \omega_i}(1-\theta)^{n-\sum_{i=1}^n \omega_i} m(\omega), \quad \text{(prodotto tensoriale di } n \text{ leggi } \mathcal{B}(1,\theta)),$$

dove m è la misura che assegna massa 1 ad ogni punto di Ω (dunque m è una misura dominante). Allora, se $T(\omega) = \sum_{i=1}^n \omega_i = \sum_{i=1}^n X_i(\omega)$, si può scrivere

$$\frac{\mathrm{d}P^\theta}{\mathrm{d}m}(\omega) = \theta^{T(\omega)}(1-\theta)^{n-T(\omega)} =: g^\theta(T(\omega)).$$

Quindi T è un riassunto esaustivo (per la condizione (b) del Teorema di Neymann–Fisher).

(b) Sia $X = (X_1, \dots, X_n)$ un campione di taglia n avente legge di Poisson di parametro $\theta > 0$ ($X_i \sim \Pi_\theta$ indipendenti); questo significa che la legge di ogni X_i è data dalla formula

$$\mu_\theta(k) = \frac{\theta^k}{k!}\mathrm{e}^{-\theta} m(k), \qquad k \in \mathbb{N},$$

dove m è la misura che "conta"i punti di $\mathbb{N}$. Secondo la definizione di campione, il modello statistico è

$$\Omega = \mathbb{R}^n, \qquad \mathcal{F} = \mathbb{R}^n, \qquad P^\theta = (\mu^\theta)^{\otimes n},$$

$X_i = i$-esima proiezione, $T = X_1 + \cdots + X_n$. La misura dominante in questo modello è $m^{\otimes n}$, cioè la misura che "conta"i punti di $\mathbb{N}^n$. Allora, indicando con $\omega = (k_1, \dots, k_n)$ ($k_i \in \mathbb{N}$) il generico elemento di Ω, per l'Osservazione 1.22 si ha

$$L(\theta; k_1, \dots, k_n) = \frac{\mathrm{d}P^\theta}{\mathrm{d}(m^{\otimes n})}(k_1, \dots, k_n) = \prod_{i=1}^n \frac{\theta^{k_i}}{k_i!}\mathrm{e}^{-\theta} = \mathrm{e}^{-n\theta}\frac{\theta^{T(\omega)}}{\prod_{i=1}^n k_i!} = g^\theta\big(T(\omega)\big) \cdot h(\omega),$$

dove

$$g^\theta(k) = \mathrm{e}^{-n\theta}\theta^k, \qquad h(\omega) = h(\omega_1, \dots, \omega_n) = \frac{1}{\prod_{i=1}^n k_i!}.$$

Dunque si conclude che $T = X_1 + \cdots + X_n$ è un riassunto esaustivo.

(c) Sia μ^θ la legge concentrata su $(0,1)$ e assolutamente continua rispetto alla misura di Lebesgue con densità

$$f^\theta(x) = (\theta+1)x^\theta 1_{(0,1)}(x), \qquad \theta > -1.$$

Sia $X = (X_1, \dots, X_n)$ un campione di taglia n e legge μ^θ (costruito con il solito metodo). Una misura dominante è la misura di Lebesgue n-dimensionale (su $(0,1)^n$). Indichiamo con $(x_1, \dots, x_n)$ il generico elemento $\omega \in \Omega = (0,1)^n$. Ancora per l'Osservazione 1.22, la verosimiglianza ha la forma

$$L(\theta; x_1, \dots, x_n) = (\theta+1)^n \Big(\prod_{i=1}^n x_i \Big)^\theta =: g^\theta(T(\omega)),$$

dove

$$g^\theta(x) = (\theta+1)^n x^\theta, \quad T(\omega) = T(x_1, \dots, x_n) = \prod_{i=1}^n x_i.$$

Detto in termini di statistiche, significa che la statistica

$$T(X_1, \dots, X_n) = \prod_{i=1}^n X_i$$

è un riassunto esaustivo.

(d) Campione di taglia n e legge $\mathcal{N}(m, \sigma^2)$. Distinguiamo tre casi:

(1) m e σ^2 sono entrambi sconosciuti, e dunque in questo caso il parametro θ è vettoriale; $\theta = (m, \sigma^2)$. La legge $\mathcal{N}(m, \sigma^2)$ è assolutamente continua rispetto alla misura di Lebesgue su $\mathbb{R}$, e la verosimiglianza è

$$\begin{aligned} L(m, \sigma^2; x_1, \dots, x_n) &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \prod_{i=1}^n \exp\Big(- \frac{(x_i - m)^2}{2\sigma^2} \Big) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(- \frac{\sum_{i=1}^n (x_i - m)^2}{2\sigma^2} - n \log \sigma \Big) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(- \frac{1}{2\sigma^2} \big(\sum_{i=1}^n x_i^2 + nm^2 - 2m \sum_{i=1}^n x_i \big) - n \log \sigma \Big) \\ &= g^\theta\Big(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \Big), \end{aligned}$$

dove

$$g^\theta(x, y) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(- \frac{1}{2\sigma^2} \big(y + nm^2 - 2mx \big) - n \log \sigma \Big).$$

Si conclude che il vettore aleatorio

$$\Big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \Big)$$

è un riassunto esaustivo (bidimensionale questa volta). Si dice anche che le due statistiche

$$\sum_{i=1}^n X_i, \qquad \sum_{i=1}^n X_i^2$$

sono *congiuntamente sufficienti.*

(2) Se m non è nota e $\sigma^2 = {\sigma_0}^2$ è nota quindi $\theta = m$), si può scrivere la verosimiglianza nella forma

$$L(m; x_1, \dots, x_n) = \underbrace{\frac{1}{(2\pi)^{\frac{n}{2}} {\sigma_0}^n} \exp\Big(- \frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2} \Big)}_{=h(\omega)} \underbrace{\exp\Big(\frac{2m \sum_{i=1}^n x_i - nm^2}{2{\sigma_0}^2} \Big)}_{=g^\theta(T(\omega))},$$

e si riconosce che la statistica $T = \sum_{i=1}^n X_i$ è sufficiente per m.

(3) Se $m = m_0$ è nota e σ^2 non è nota, scrivendo la verosimiglianza nella forma

$$L(\sigma^2; x_1, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(- \frac{\sum_{i=1}^n (x_i - m_0)^2}{2\sigma^2} - n \log \sigma \Big) = g^\theta\Big(\sum_{i=1}^n (x_i - m_0)^2 \Big),$$

si riconosce che la statistica $T = \sum_{i=1}^n (X_i - m_0)^2$ è sufficiente per σ^2.

(e) Sia $(X_1, \dots, X_n)$ un campione di taglia n di legge uniforme sull'intervallo (θ_1, θ_2), con $\theta_1 < \theta_2$ entrambi sconosciuti. Anche in questo caso il parametro è bidimensionale, $\theta = (\theta_1, \theta_2)$ con $\theta \in \mathbb{R}^2 \cap \{y > x\}$. μ^θ è la legge su $\mathbb{R}$ avente densità (rispetto alla misura di Lebesgue)

$$f^\theta(x) = \frac{1}{\theta_2 - \theta_1} 1_{(\theta_1, \theta_2)}(x).$$

La misura dominante è al solito la misura di Lebesque n-dimensionale, e la verosimiglianza ha la forma

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\theta_2 - \theta_1} 1_{(\theta_1, \theta_2)}(x_i) = \frac{1}{(\theta_2 - \theta_1)^n} \prod_{i=1}^n 1_{(\theta_1, \theta_2)}(x_i).$$

Come si vede facilemente, si ha

$$\begin{aligned}
\prod_{i=1}^n 1_{(\theta_1, \theta_2)}(x_i) &= \begin{cases} 1 & \text{se } x_i \in (\theta_1, \theta_2) \quad \forall i = 1, \dots, n \\ 0 & \text{altrimenti} \end{cases} \\
&= \begin{cases} 1 & \text{se } \begin{cases} u = \min(x_1, \dots, x_n) > \theta_1 \\ v = \max(x_1, \dots, x_n) < \theta_2 \end{cases} \\ 0 & \text{altrimenti} \end{cases} \quad = 1_{(\theta_1, \theta_2) \times (\theta_1, \theta_2)}(u, v) \\
&= g^\theta(u, v),
\end{aligned}$$

dove evidentemente si pone

$$g^\theta(x, y) = 1_{(\theta_1, \theta_2) \times (\theta_1, \theta_2)}(x, y).$$

Tutto questo significa che le statistiche

$$\begin{cases} U = \min(X_1, \dots, X_n) \\ V = \max(X_1, \dots, X_n) \end{cases}$$

sono congiuntamente sufficienti (ved. esempio precedente).

Osservazione 2.10 (i) Nell'esempio (d)(1) il risultato ottenuto è ragionevole, se si pensa che, per la Legge dei Grandi Numeri, le statistiche $\frac{\sum_{i=1}^n X_i}{n}$ e $\frac{\sum_{i=1}^n X_i^2}{n}$ sono buone approssimazioni di m e di $\sigma^2 + m^2$ rispettivamente. Considerazioni simili valgono per i punti (d)(2) e (d)(3).

(ii) Nell'esempio (e), il risultato è interpretabile intuitivamente pensando che una v.a. uniforme sull'intervallo (θ_1, θ_2) modellizza la scelta di un punto a caso nell'intervallo dato; è ragionevole individuare i due estremi dell'intervallo mediante il più piccolo e il più grande dei punti ottenuti nelle n scelte.

Osservazione 2.11 Gli esempi (d) e (e) fanno pensare che la statistica necessaria per individuare un parametro s-dimensionale sia di dimensione s. In realtà le cose non vanno sempre così, come mostra l'esempio seguente (nel quale per un parametro unidimensionale è necessario avere una statistica bidimensionale).

ESERCIZIO. Sia $X = (X_1, \dots, X_n)$ un campione di legge μ^θ, $\theta \in \mathbb{N} \setminus \{0\}$, avente densità rispetto alla misura che conta i punti di $\mathbb{N}$

$$f^\theta(x) = \begin{cases} C(\theta) 2^{-\frac{x}{\theta}}, & x = \theta, \theta+1, \theta+2, \dots \\ 0 & \text{altrimenti} \end{cases}$$

($C(\theta)$ è l'opportuna costante normalizzatrice). Trovare una statistica sufficiente per θ.

SOLUZIONE. La verosimiglianza (rispetto alla misura n-dimensionale che conta i punti) è

$$\begin{aligned} L(\theta; x_1, \dots, x_n) &= \begin{cases} \prod_{i=1}^n C(\theta) 2^{-\frac{x_i}{\theta}} & x_i \text{ intero } \geq \theta \ \forall\, i = 1, \dots, n \\ 0 & \text{altrimenti} \end{cases} \\ &= \begin{cases} C(\theta)^n 2^{-\frac{\sum_{i=1}^n x_i}{\theta}} & x_i \text{ intero } \geq \theta \ \forall\, i = 1, \dots, n \\ 0 & \text{altrimenti} \end{cases} \\ &= C(\theta)^n 2^{-\frac{\sum_{i=1}^n x_i}{\theta}} \phi^\theta\big(\min(x_1, \dots, x_n)\big) \\ & g^\theta\Big(\sum_{i=1}^n x_i, \min(x_1, \dots, x_n)\Big) \end{aligned}$$

dove

$$\phi^\theta(t) = \begin{cases} 1 & t \geq \theta \\ 0 & \text{altrimenti} \end{cases}$$

e

$$g^\theta(x, y) = C(\theta)^n 2^{-\frac{x}{\theta}} \phi^\theta(y).$$

Dunque in questo caso la statistica è bidimensionale:

$$T = \Big(\sum_{i=1}^n X_i, \min(X_1, \dots, X_n)\Big).$$

Situazioni come questa si incontrano quando l'insieme $\{x : f^\theta(x) \neq 0\}$ dipende effettivamente da θ (nei casi degli esempi (c) e (d) tale insieme è costante in θ).

3 Teoria della stima. La nozione di stimatore

Sia $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$un modello statistico. Supponiamo dapprima che $\Theta \subset \mathbb{R}$. Stimare il parametro θ significa assegnare una funzione $U : \Omega \to \mathbb{R}$ che supporremo misurabile (e dunque U è una v.a.); intuitivamente ciò significa che, se il risultato del nostro esperimento è un $\omega \in \Omega$, in base a tale risultato assegniamo un numero, che indichiamo con $U(\omega)$ perché dipende da ω. Più in generale, senza restringerci cioè al caso di $\Theta \subset \mathbb{R}$, assegnata $g : \Theta \to \mathbb{R}$, potremmo aver bisogno di stimare $g(\theta)$; ad esempio, in un campione di legge gaussiana nel quale il parametro è la coppia $\theta = (m, \sigma^2)$, possiamo essere interessati a stimare solo la media m o solo la varianza σ^2.

Definizione 3.1 Assegnati il modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$e una funzione $g : \Theta \to D$ (D aperto di $\mathbb{R}$) si chiama *stimatore* di $g(\theta)$ una v.a. $U : \Omega \to D$ che non dipenda da θ (si tratta dunque di una statistica nel senso della definizione 2.1). Assegnato $\omega \in \Omega$, il numero $U(\omega)$ si chiama *stima* di $g(\theta)$.

Intuitivamente uno stimatore è "buono" quando si avvicina (in qualche senso da definire) al vero valore del parametro θ (o della sua funzione $g(\theta)$ nel caso generale). Più precisamente, è chiaro che la sostituzione del valore vero con il suo stimatore U comporta un "costo" (o "perdita"), ed uno stimatore sarà tanto migliore quanto più piccolo è questo costo. Dunque si tratta di chiarire cosa intendiamo per costo. In generale, il "costo"conseguenza della sostituzione di $g(\theta)$ con un numero reale a è una funzione non negativa $(\theta, a) \mapsto C(\theta, a)$. Poiché noi sostituiamo $g(\theta)$ con $U(\omega)$, per ogni θ otteniamo la funzione $\omega \mapsto C(\theta, U(\omega))$, che supporremo misurabile (dunque sarà una v.a.). Di questa v.a. faremo poi la media rispetto a P^θ.

Definizione 3.2 Si chiama *rischio* dello stimatore U il costo medio rispetto a P^θ o, più precisamente, la funzione, definita su Θ,

$$\theta \mapsto R_U(\theta) := E^\theta[C(\theta, U)] = \int_\Omega C(\theta, U(\omega))\, \mathrm{d}P^\theta(\omega).$$

Generalmente come funzione "costo" si prende $(\theta, a) \mapsto |g(\theta) - a|^2$, e si parla in questo caso di *costo quadratico*. In corrispondenza, il rischio $R_U(\theta) = E^\theta\left[|g(\theta) - U|^2\right]$ si chiama *rischio quadratico*. Tuttavia altre scelte sono possibili per il costo, e di conseguenza per il rischio.

Avvertenza. D'ora in avanti, salvo diverso avviso, utilizzeremo sempre la funzione "costo quadratico"e, di conseguenza, parlando di rischio, intenderemo il "rischio quadratico".

Avendo a disposizione il rischio, siamo ora in grado di stabilire una "gerarchia" tra stimatori.

Definizione 3.3 (a) Uno stimatore U si dice *preferibile* ad un altro stimatore V se, per ogni $\theta \in \Theta$, si ha

$$R_U(\theta) \le R_V(\theta).$$

(b) U si dice *strettamente preferibile* a V se è preferibile ed esiste almeno un $\theta_0 \in \Theta$ tale che

$$R_U(\theta_0) < R_V(\theta_0)$$

(cioè con la disuguaglianza stretta).

(c) Sia $\mathcal{D}$ una famiglia di stimatori, e sia $U \in \mathcal{D}$. U si dice ammissibile (relativamente a $\mathcal{D}$) se in $\mathcal{D}$ non esistono altri stimatori strettamente preferibili a U.

(d) $U \in \mathcal{D}$ si dice *ottimale* (relativamente a $\mathcal{D}$) se è preferibile ad ogni altro stimatore della famiglia $\mathcal{D}$.

Osservazione 3.4 La relazione di preferibilità è un preordinamento nella classe degli stimatori: due stimatori possono essere non confrontabili, cioè può accadere che per alcuni valori di θ si abbia $R_U(\theta) \le R_V(\theta)$ e per altri si abbia invece $R_U(\theta) \ge R_V(\theta)$, come mostra l'esempio seguente.

Esempio 3.5 Nell'esempio del controllo di qualità, consideriamo uno stimatore del parametro θ della forma

$$U = h(X_1 + \cdots + X_n), \qquad \text{dove} \quad h(t) = \frac{t + a}{n + b}, \quad a \ge 0,\, b \ge 0.$$

Calcoliamo $R_U(\theta)$. Poniamo $S_n = X_1 + \cdots + X_n$ e ricordiamo che

$$E^\theta[S_n] = n\theta, \qquad Var^\theta(S_n) = n\theta(1 - \theta).$$

Si ha allora

$$R_U(\theta) = E^\theta\big[(h(S_n)-\theta)^2\big] = \frac{1}{(n+b)^2}E^\theta\big[\{(S_n-n\theta)+(a-b\theta)\}^2\big]$$
$$= \frac{1}{(n+b)^2}\Big(E^\theta\big[(S_n-n\theta)^2\big] + (a-b\theta)^2 + 2(a-b\theta)E^\theta\big[S_n-n\theta\big]\Big)$$
$$= \frac{1}{(n+b)^2}\Big(\underbrace{E^\theta\big[(S_n-E^\theta[S_n])^2\big]}_{=Var^\theta(S_n)} + (a-b\theta)^2 + 2(a-b\theta)\underbrace{E^\theta\big[S_n-E^\theta[S_n]\big]}_{=0}\Big)$$
$$= \frac{1}{(n+b)^2}\Big(n\theta(1-\theta)+(a-b\theta)^2\Big) = \frac{(b^2-n)\theta^2+(n-2ab)\theta+a^2}{(n+b)^2}.$$

A seconda dei valori che si danno ai parametri a e b, si hanno diversi andamenti della funzione di rischio. In particolare si vede che due di questi stimatori possono essere non confrontabili tra loro. Per esempio, ponendo $a = \frac{\sqrt{n}}{2}$ e $b = \sqrt{n}$ si ottiene lo stimatore U_1 di rischio costante $R_{U_1}(\theta) = \frac{1}{4(\sqrt{n}+1)^2}$ (si tratta dello stimatore bayesiano, v....).
Invce, per $a = b = 0$ (che corrisponde al caso in cui si prende come stimatore U_2 la *media campionaria*, di cui parleremo al....) si ottiene la funzione di rischio $R_{U_2}(\theta) = \frac{\theta(1-\theta)}{n}$; si vede facilmente che $R_{U_2}(\frac{1}{2}) \geq R_{U_1}(\frac{1}{2})$, mentre $R_{U_2}(\theta) \leq R_{U_1}(\theta)$ per θ abbastanza vicino a 0 oppure a 1.

Osservazione 3.6 Tutte le definizione date sopra restano valide nel caso multidimensionale, cioè se si considera

(i) una funzione $g : \theta \to D$, con D aperto di $\mathbb{R}^k$;

(ii) una famiglia di stimatori a valori in $\mathbb{R}^k$;

(iii) il costo quadratico definito da $C(\theta, a) = \|g(\theta) - a\|^2$.

Sarebbe bello che la stima $U(\omega)$ coincidesse proprio con la quantità da stimare $g(\theta)$, ma ovviamente questa è una richiesta impossibile a soddisfare. Possiamo però chiedere che ciò accada "in media", cioè si dà la seguente

Definizione 3.7 (a) Uno stimatore U di $g(\theta)$ si dice *corretto* (o *non distorto*, *unbiased* in inglese) se è integrabile (il che significa integrabile per ogni P^θ) ed inoltre

$$E^\theta[U] = g(\theta), \qquad \forall\, \theta \in \Theta.$$

Osservazione 3.8 Sia U uno stimatore di $g(\theta)$. Si ha allora

$$R_U(\theta) = E^\theta\big[(U-g(\theta))^2\big] = E^\theta\big[\{(U-E^\theta[U])+(E^\theta[U]-g(\theta)\}^2\big]$$
$$= E^\theta\big[(U-E^\theta[U])^2\big] + (E^\theta[U]-g(\theta))^2 + 2(E^\theta[U]-g(\theta))\underbrace{E^\theta\big[U-E^\theta[U]\big]}_{=0}$$
$$= Var^\theta(U) + (E^\theta[U]-g(\theta))^2 \geq Var^\theta(U)$$

Evidentemente, se U è corretto, allora $R_U(\theta) = Var^\theta(U)$.

Esercizio 3.9 Si consideri un campione di legge su $\mathbb{R}$ $\{\mu^\theta, \theta \in \Theta\}$ con momento secondo finito, cioè tale che per ogni $\theta \in \Theta$, si abbia

$$\int x^2 \,\mathrm{d}\mu^\theta(x) < +\infty.$$

(a) Mostrare che, tra tutti gli stimatori corretti della media $m(\theta) = \int x \, \mathrm{d}\mu^\theta(x)$ (supposta non nulla) della forma $\sum_{i=1}^n a_i X_i$, con $a_i \in \mathbb{R}$, la *media campionaria* (o *media empirica*)

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

è ottimale.

Soluzione. Intanto osserviamo che, se $U = \sum_{i=1}^n a_i X_i$ è uno stimatore corretto della media, allora

$$m(\theta) = E^\theta\Big[\sum_{i=1}^n a_i X_i\Big] = \sum_{i=1}^n a_i E^\theta[X_i] = \sum_{i=1}^n a_i m(\theta)$$

da cui segue

$$\sum_{i=1}^n a_i = 1.$$

Il rischio di U è uguale alla sua varianza, e cioè

$$R_U(\theta) = Var^\theta(U) = Var^\theta\Big(\sum_{i=1}^n a_i X_i\Big) = \sum_{i=1}^n Var^\theta\Big(a_i X_i\Big) = \sum_{i=1}^n a_i^2 Var^\theta(X_i) = \sum_{i=1}^n a_i^2 \sigma(\theta)^2,$$

dove $\sigma^2(\theta) = \int (x - m(\theta))^2 \, \mathrm{d}\mu^\theta(x)$ è la varianza della legge μ^θ. Dunque il problema si riduce al calcolo del minimo della quantità $\sum_{i=1}^n a_i^2$ sotto il vincolo $\sum_{i=1}^n a_i = 1$, ed è ben noto che tale minimo si raggiunge per

$$a_i = \frac{1}{n}, \qquad \forall, i = 1, \ldots, n.$$

Inoltre esso vale

$$\sum_{i=1}^n \frac{1}{n^2}\sigma^2(\theta) = \frac{\sigma^2(\theta)}{n}, \quad \text{(varianza della media campionaria)}.$$

(b) Si chiama *varianza campionaria* la statistica

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

Mostare che S^2 è uno stimatore corretto di $\sigma^2(\theta)$.

Soluzione

$$\begin{aligned} &E^\theta\Big[\sum_{i=1}^n (X_i - \overline{X})^2\Big] = E^\theta\Big[\sum_{i=1}^n \{(X_i - m(\theta)) - (\overline{X} - m(\theta))\}^2\Big] \\ &= \sum_{i=1}^n E^\theta\big[(X_i - m(\theta))^2\big] + \sum_{i=1}^n E^\theta\big[(\overline{X} - m(\theta))^2\big] - 2E^\theta\Big[(\overline{X} - m(\theta)) \sum_{i=1}^n (X_i - m(\theta))\Big]. \end{aligned}$$

Osserviamo ora che $E^\theta\big[(X_i - m(\theta))^2\big] = Var^\theta(X_i) = \sigma^2(\theta)$; inoltre, ricordando che $E^\theta[\overline{X}] = m(\theta)$ abbiamo

$$E^\theta\big[(\overline{X} - m(\theta))^2\big] = Var^\theta(\overline{X}) = Var^\theta\Big(\frac{\sum_{i=1}^n X_i}{n}\Big) = \frac{1}{n^2}\sum_{i=1}^n Var^\theta(X_i) = \frac{\sigma^2(\theta)}{n}$$

(qui è stata usata l'indipendenza delle X_i); infine

$$E^\theta\Big[(\overline{X}-m(\theta))\sum_{i=1}^n(X_i-m(\theta))\Big]=E^\theta\Big[(\overline{X}-m(\theta))\Big(\sum_{i=1}^n X_i-nm(\theta)\Big)\Big]$$
$$=nE^\theta\Big[(\overline{X}-m(\theta))\Big(\frac{\sum_{i=1}^n X_i}{n}-m(\theta)\Big)\Big]=nE^\theta\big[(\overline{X}-m(\theta))^2\big]=nVar^\theta(\overline{X})=\sigma^2(\theta).$$

Usando le precedenti relazioni si ottiene

$$E^\theta\Big[\sum_{i=1}^n(X_i-\overline{X})^2\Big]=n\sigma^2(\theta)+n\cdot\frac{\sigma^2(\theta)}{n}-2\sigma^2(\theta)=(n-1)\sigma^2(\theta).$$

4 Stimatori ed esaustività

In tutto il paragrafo sono assegnati un modello statistico $\big(\Omega,\mathcal{F},\{P^\theta,\theta\in\Theta\}\big)$ed una funzione $g:\Theta\to\mathbb{R}$. Il Teorema che segue indica in quali casi esistono "buoni" stimatori.

Teorema 4.1 (di Blackwell-Rao). *Sia T un riassunto esaustivo a valori in $(E,\mathcal{E})$, e sia U uno stimatore di $g(\theta)$ di quadrato integrabile (rispetto ad ogni P^θ). Allora*

(i) lo stimatore $V:=E^\diamond[U|T]$ (versione della speranza condizionale non dipendente da θ) è preferibile a U; inoltre V è strettamente preferibile a U a meno che U non sia $\sigma(T)$-misurabile.

(ii) Se U è corretto, anche V è corretto.

Dimostrazione. (i) Si ha

$$R_U(\theta)=E^\theta\big[(U-g(\theta))^2\big]=\big\|U-g(\theta)\big\|^2_{L^2(\Omega,\mathcal{F},P^\theta)}.$$

Inoltre

$$V-g(\theta)=E^\diamond[U|T]-g(\theta)=E^\diamond[U-g(\theta)|T]=E^\theta[U-g(\theta)|T].$$

Ricordiamo ora che $E^\theta[U-g(\theta)|T]$ coincide con la proiezione ortogonale di $U-g(\theta)$ sul sottospazio chiuso $L^2(\Omega,\sigma(T),P^\theta)$, e quindi la sua norma L^2 è più piccola di quella di $U-g(\theta)$; inoltre essa è strettamente più piccola a meno che $U-g(\theta)$ non appartenga già a $L^2(\Omega,\sigma(T),P^\theta)$, e cioè non coincida P^θ-q.o. con una funzione $\sigma(T)$ misurabile.
(ii) Si ha

$$E^\theta[V]=E^\theta\big[E^\diamond[U|T]\big]=E^\theta\big[E^\theta[U|T]\big]=E^\theta[U]=g(\theta).$$

□

Osservazione 4.2 Il significato del Teorema 4.1 è che, in presenza di un riassunto esaustivo, è necessario cercare stimatori che siano T-misurabili, cioè della forma $h(T)$, con $h:E\to\mathbb{R}$ funzione misurabile opportuna. In altre parole, i "buoni" stimatori dipendono da ω solo attraverso $T(\omega)$.

Esercizio 4.3 Sia $(X_1,\dots,X_n)$ un campione avente legge Π_θ, $\theta>0$; si vuole stimare la quantità $g(\theta)=\mathrm{e}^{-\theta}$ con i due stimatori

$$T_1=\begin{cases}1 & \text{su} \quad \{X_1=0\}\\ 0 & \text{su} \quad \{X_1\geq 1\}\end{cases};\qquad T_2=\Big(\frac{n-1}{n}\Big)^{\sum_{k=1}^n X_k}.$$

(i) Mostrare che T_1 e T_2 sono stimatori corretti di $g(\theta)$.

(ii) Calcolare R_{T_1} e R_{T_2}.

SOLUZIONE (i) Si ha subito

$$E^\theta[T_1] = P^\theta(X_1 = 0) = \mathrm{e}^{-\theta}.$$

Ricordando che $\sum_{k=1}^n X_k \sim \Pi_{n\theta}$ si trova

$$E^\theta[T_2] = \sum_{j=0}^{\infty} \Big(\frac{n-1}{n}\Big)^j \frac{(n\theta)^j}{j!}\mathrm{e}^{-n\theta} = \sum_{j=0}^{\infty} \frac{\big((n-1)\theta\big)^j}{j!}\mathrm{e}^{-n\theta} = \mathrm{e}^{-n\theta}\mathrm{e}^{(n-1)\theta} = \mathrm{e}^{-\theta}.$$

(ii) Poiché i due stimatori sono corretti, i loro rischi sono uguali alle loro varianze. Quindi

$$R_{T_1}(\theta) = Var^\theta(T_1) = \mathrm{e}^{-\theta}\big(1-\mathrm{e}^{-\theta}\big) = \mathrm{e}^{-2\theta}\big(\mathrm{e}^{\theta}-1\big);$$
$$R_{T_2}(\theta) = Var^\theta(T_2) = E^\theta[T_2^2] - \big(E^\theta[T_2]\big)^2 = E^\theta[T_2^2] - \mathrm{e}^{-2\theta}.$$

Calcoliamo $E^\theta[T_2^2]$.

$$E^\theta[T_2^2] = \sum_{j=0}^{\infty} \Big(\frac{n-1}{n}\Big)^{2j} \frac{(n\theta)^j}{j!}\mathrm{e}^{-n\theta} = \sum_{j=0}^{\infty} \Big(\frac{\theta(n-1)^2}{n}\Big)^j \frac{1}{j!}\mathrm{e}^{-n\theta} = \mathrm{e}^{-n\theta+\frac{\theta(n-1)^2}{n}} = \mathrm{e}^{\frac{\theta(1-2n)}{n}}.$$

Dunque

$$R_{T_2}(\theta) = E^\theta[T_2^2] - \mathrm{e}^{-2\theta} = \mathrm{e}^{\frac{\theta(1-2n)}{n}} - \mathrm{e}^{-2\theta} = \mathrm{e}^{-2\theta}\big(\mathrm{e}^{\frac{\theta}{n}}-1\big).$$

Si vede dunque che

$$R_{T_2}(\theta) < R_{T_1}(\theta),$$

come prevede il Teorema 4.1, dato che, come sappiamo, T_2 è una statistica esaustiva.

Definizione 4.4 Una statistica esaustiva T si dice *completa* se ogni v.a. che sia T-misurabile (cioè della forma $h(T)$), integrabile e tale che $E^\theta[Y] = 0$ per ogni $\theta \in \Theta$ è nulla (P^θ-q.c. per ogni $\theta \in \Theta$).

Il seguente teorema giustifica l'importanza della definizione appena data.

Teorema 4.5 Supponiamo che esista una statistica esaustiva completa e sia U uno stimatore corretto e di quadrato integrabile di $g(\theta)$. Allora $E^\diamond[U|T]$ è preferibile ad ogni altro stimatore corretto e di quadrato integrabile.

DIMOSTRAZIONE. Per il Teorema di 4.1 sappiamo che $E^\diamond[U|T]$ è preferibile a U. Quindi basterà dimostare che, per ogni coppia U e V di stimatori corretti e di quadrato integrabile, risulta $E^\diamond[U|T] = E^\diamond[V|T]$. Questo è vero: infatti, intanto $E^\diamond[U-V|T]$ è T-misurabile, ed inoltre

$$E^\theta\big[E^\diamond[U-V|T]\big] = E^\theta\big[E^\theta[U-V|T]\big] = E^\theta[U-V] = E^\theta[U] - E^\theta[V] = g(\theta) - g(\theta) = 0,$$

perché U e V sono stimatori corretti di $g(\theta)$. Dunque

$$E^\diamond[U|T] - E^\diamond[V|T] = E^\diamond[U-V|T] = 0,$$

perché la statistica T è completa.

□

Osservazione 4.6 Il Teorema precedente dice in altre parole che, nel caso esista una statistica esaustiva completa, o non esiste alcuno stimatore corretto di quadrato integrabile, oppure nella classe di tali stimatori esiste un elemento ottimale.

Le statistiche esaustive complete possono essenzialmente essere trovate in due modi: o per calcolo diretto oppure mediante l'uso dei modelli esponenziali, che vedremo nel prossimo paragrafo. Qui facciamo un calcolo diretto, nel caso di un modello che non è esponenziale.

Esercizio 4.7 Consideriamo un campione $(X_1, \dots, X_n)$ avente legge $\mathcal{U}\big([0,\theta]\big)$, con $\theta > 0$. La verosimiglianza rispetto alla misura di Lebesgue n-dimensionale sullo spazio $\big(\mathbb{R}^{+n}, \mathcal{B}(\mathbb{R}^{+n})\big)$ ha la forma

$$L(\theta; x_1, \dots, x_n) = \frac{1}{\theta^n} \prod_{i=1}^{n} 1_{(0,\theta)}(x_i) = \frac{1}{\theta^n} 1_{(0,\theta)}\big(\max_{1\le i\le n} x_i\big).$$

Dunque $T = \max_{1\le i\le n} X_i$ è una statistica esaustiva. Vogliamo vedere che T è completa. Poiché, come vedremo, il modello non è esponenziale, non ci resta che fare un calcolo diretto. Sia dunque $Y = h(T)$ una v.a. T-misurabile (h boreliana integrabile) tale che $E^\theta[h(T)] = 0$ per ogni $\theta \in \Theta$. Dobbiamo vedere che $Y = 0$, P^θ-q.c. Indicando con f^θ la densità della legge di T sotto la probabilità P^θ, si ha

$$E^\theta[h(T)] = \int_{-\infty}^{+\infty} h(t) f^\theta(t)\, \mathrm{d}t,$$

e quindi tutto è ricondotto al calcolo di f^θ. Calcoliamo dapprima la funzione di ripartizione di T. Si ha evidentemente

$$F^\theta(t) = P^\theta\big(\max_{1\le i\le n} X_i \le t\big) = \begin{cases} 0 & \text{per } t < 0 \\ 1 & \text{per } t > \theta. \end{cases}$$

Per $0 \le t \le \theta$ si ha invece

$$P^\theta\big(\max_{1\le i\le n} X_i \le t\big) = P^\theta\big(\bigcap_{i=1}^{n} \{X_i \le t\}\big) = \prod_{i=1}^{n} P^\theta(X_i \le t) = \big\{P^\theta(X_i \le t)\big\}^n.$$

Dato che

$$P^\theta(X_i \le t) = \int_0^t \frac{1}{\theta}\, \mathrm{d}u = \frac{t}{\theta},$$

si conclude che, per $0 \le t \le \theta$ risulta

$$P^\theta\big(\max_{1\le i\le n} X_i \le t\big) = \frac{t^n}{\theta^n}.$$

Una densità di T si ottiene allora "per derivazione", e vale

$$f^\theta(t) = \begin{cases} \frac{nt^{n-1}}{\theta^n} & \text{per } 0 < t < \theta \\ 0 & \text{altrimenti} \end{cases} = \frac{nt^{n-1}}{\theta^n} 1_{(0,\theta)}(t).$$

Si conclude quindi che

$$E^\theta[h(t)] = \int_{-\infty}^{+\infty} h(t) f^\theta(t)\, \mathrm{d}t = \frac{n}{\theta^n} \int_0^\theta h(t) t^{n-1}\, \mathrm{d}t = 0, \qquad \forall\, \theta > 0,$$

e non è difficile vedere che questa relazione vale se e solo se $h(t)t^{n-1} = 0$, e quindi anche $h(t) = 0$, $\forall\, t \in (0,\theta)$, q.c. rispetto alla misura di Lebesgue. In altre parole abbiamo dimostrato che $Y = h(T) = 0$, P^θ-q.c. per ogni θ, e cioè che T è completa.

Cerchiamo uno stimatore corretto di θ. Non ci sono regole generali; in questo caso possiamo procedere così : calcoliamo

$$E^\theta[T] = \frac{n}{\theta^n}\int_0^\theta t\cdot t^{n-1}\,\mathrm{d}t = \frac{n}{\theta^n}\int_0^\theta t^n\,\mathrm{d}t = \frac{n}{n+1}\cdot\frac{\theta^{n+1}}{\theta^n} = \frac{n\theta}{n+1}.$$

Da questo calcolo si deduce che

$$\frac{n+1}{n}\big(\max_{1\le i\le n} X_i\big)$$

è uno stimatore corretto (e dunque ottimale per il Teorema 4.5) di θ.

Diamo ora la nozione di statistica *libera.* Si tratta dell'opposto della nozione di esaustività, nel senso che, mentre una statistica esaustiva conserva tutta l'"informazione"sul parametro θ fornita dal modello statistico, una statistica libera non dà su θ alcuna informazione (questa affermazione sarà più chiara una volta che avremo il concetto di *informazione di Fisher*). La formalizzazione di questa idea è data dalla seguente

Definizione 4.8 Assegnato il modello statistico $\big(\Omega,\mathcal{F},\{P^\theta,\theta\in\Theta\}\big)$, una statistica S, definita su di esso e a valori in $(E,\mathcal{E})$, si dice *libera* se la sua legge $S(P^\theta)$ non dipende da θ.

Esempio 4.9 Sia $(X_1,\dots,X_n)$ un campione di legge $\mathcal{N}(\theta,1)$, e poniamo

$$\overline{X} = \frac{X_1+\cdots+X_n}{n}$$

(media campionaria).

(a) La statistica

$$S_1 = \sum_{i=1}^n (X_i-\overline{X})^2$$

è libera in quanto, posto $Y_i = X_i-\theta$, si ha $\overline{Y} = \overline{X}-\theta$, e quindi

$$S_1 = \sum_{i=1}^n (Y_i-\overline{Y}^2).$$

Il vettore $(Y_1,\dots,Y_n)$ ha per legge il prodotto tensoriale di n leggi $\mathcal{N}(0,1)$, che non dipende da θ; di conseguenza anche S_1 ha una legge non dipendente da θ (vedremo in seguito che si tratta della $\chi^2(n-1)$, ved. richiamo dopo la Proposizione 9.1).

(b) Poniamo $S_2 = \max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i$. Si vede subito che $S_2 = \max_{1\le i\le n} Y_i - \min_{1\le i\le n} Y_i$, dove il vettore $(Y_1,\dots,Y_n)$ è quello introdotto sopra; dunque anche S_2 è una statistica libera.

Vediamo ora quali relazioni sussistono tra i concetti di statistica esaustiva e statistica libera.

Teorema 4.10 *Sul modello statistico* $\big(\Omega,\mathcal{F},\{P^\theta,\theta\in\Theta\}\big)$*siano S e T due statistiche. Supponiamo che T sia esaustiva completa e S sia libera. Allora S e T sono indipendenti (come statistiche, cioè relativamente ad ogni P^θ).*

DIMOSTRAZIONE. La tesi equivale al fatto che, per ogni funzione boreliana limitata h, risulta $E^\theta\big[h(S)|T\big] = E^\theta[h(S)]$, ovvero

$$E^\theta\big[h(S)|T\big] - E^\theta[h(S)] = 0. \tag{2}$$

La v.a. al primo membro della (2) è T-misurabile ($E^\theta\big[h(S)|T\big]$ è una speranza condizionata a T e $E^\theta[h(S)]$ è costante); dunque, grazie al criterio di Doob, per ogni θ, esiste una funzione misurabile ϕ^θ tale che

$$E^\theta\big[h(S)|T\big] - E^\theta[h(S)] = \phi^\theta(T), \qquad \forall\, \theta \in \Theta.$$

D'altra parte la stessa v.a. non dipende da θ ($E^\theta\big[h(S)|T\big]$ non dipende da θ perché T è un riassunto esaustivo e $E^\theta[h(S)]$ non dipende da θ perché S è libera), dunque $\theta \mapsto \phi^\theta$ è costante in θ; in altre parole si ha

$$E^\theta\big[h(S)|T\big] - E^\theta[h(S)] = \phi(T) \qquad \forall\, \theta \in \Theta, \tag{3}$$

per un'opportuna funzione misurabile ϕ.

D'altra parte, per la (3), si ha

$$E^\theta[\phi(T)] = E^\theta\big[E^\theta[h(S)|T] - E^\theta[h(S)]\big] = E^\theta\big[E^\theta[h(S)|T]\big] - E^\theta\big[E^\theta[h(S)]\big] = E^\theta[h(S)] - E^\theta[h(S)] = 0,$$

e si conclude che $\phi(T) = 0$ poiché T è completa. La (2) segue allora dalla (3).

□

Osservazione 4.11 Torniamo alla situazione dell'esempio 4.9; vedremo tra poco (nella sezione sui modelli esponenziali) che la statistica $\overline{X}$ è esaustiva completa. Si deduce allora dal Teorema precedente che $\overline{X}$ è indipendente da S_1 e da S_2. Ritroveremo l'indipendenza di S_1 e $\overline{X}$ come conseguenza del Teorema di Cochran (Teorema 9.3).

Del Teorema precedente vale un viceversa, in ipotesi particolari. Precisamente

Teorema 4.12 *Sul modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$, dominato dalla misura μ, siano S e T due statistiche. Supponiamo che T sia esaustiva e T e S siano indipendenti (al solito in quanto statistiche). Supponiamo inoltre che, per ogni coppia (θ_1, θ_2), P^{θ_1} e P^{θ_2} non siano tra loro singolari. Allora S è libera.*

DIMOSTRAZIONE. Dimostrare che S è libera significa dimostrare che, per ogni h limitata boreliana, si ha

$$E^{\theta_1}[h(S)] = E^{\theta_2}[h(S)].$$

Intanto, poiché T è esaustiva, esiste $E^\diamond[h(S)|T]$ (versione non dipendente da θ). Allora, per ogni coppia (θ_1, θ_2) fissata, esiste $B \in \mathcal{F}$ con $\mu(B) = 1$, tale che

$$E^\diamond[h(S)|T](\omega) = E^{\theta_1}[h(S)|T](\omega) = E^{\theta_2}[h(S)|T](\omega), \qquad \text{per ogni } \omega \in B.$$

D'altra parte, dato che S e T sono indipendenti, esistono $A_1 \in \mathcal{F}$ e $A_1 \in \mathcal{F}$, con $P^{\theta_1}(A_1) = P^{\theta_2}(A_2) = 1$ e

$$E^{\theta_1}[h(S)|T](\omega) = E^{\theta_1}[h(S)], \quad \forall \omega \in A_1; \quad E^{\theta_2}[h(S)|T](\omega) = E^{\theta_2}[h(S)], \quad \forall \omega \in A_2.$$

Possiamo supporre che $A_1 \subseteq B$ e $A_2 \subseteq B$: infatti, dato che $\mu(B^c) = 0$, si ha anche $P^{\theta_1}(B^c) = 0$ (poiché μ domina P^{θ_1}), e dunque anche $P^{\theta_1}(A_1 \cap B^c) = 0$. In modo analogo si vede che anche $P^{\theta_2}(A_2 \cap B^c) = 0$. Si può allora sostituire A_1 con $A_1 \cap B$ e A_2 con $A_2 \cap B$.

A_1 e A_2 non sono disgiunti, in quanto il primo porta la probabilità P^{θ_1} e il secondo la probabilità P^{θ_2}, che sono tra loro non singolari. Dunque esiste almeno un $\omega_0 \in A_1 \cap A_2 \subseteq B$; per tale ω_0 abbiamo allora la catena di uguaglianze

$$E^{\theta_1}[h(S)] \underbrace{=}_{\omega_0 \in A_1} E^{\theta_1}[h(S)|T](\omega_0) \underbrace{=}_{\omega_0 \in B} E^{\theta_2}[h(S)|T](\omega_0) \underbrace{=}_{\omega_0 \in A_2} E^{\theta_2}[h(S)],$$

come volevamo. □

Applicazione. Sia $X = (X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$; sia assegnata una matrice $A = (a_{i,j})_{i,j=1,\dots,n}$ simmetrica e semidefinita positiva. Condizione necessaria e sufficiente affinché $\overline{X}$ e ${}^tXAX$ siano indipendenti è che $\sum_{i=1}^n a_{i,j} = 0$, per ogni $j = 1, \dots, n$ ($\sum_{i=1}^n a_{i,j}$ è la somma degli elementi della matrice A che si trovano sulla colonna j-esima, ma per simmetria anche la somma degli elementi di A che si trovano sulla riga j-esima).

Dimostrazione. Supponiamo dapprima che sia $\sigma^2 = 1$ e indichiamo con $\mathbf{0}$ ed $\mathbf{e}$ i vettori n-dimensionali ${}^t(0, \dots, 0)$ e ${}^t(1, \dots, 1)$ (vettori colonna con tutte le componenti uguali a 0 e 1 rispettivamente). Poniamo infine $Y = X - m\mathbf{e}$.
Il vettore aleatorio X ha legge $\mathcal{N}_n(m\mathbf{e}, I)$ (si tratta della legge gaussiana n-dimensionale di media $m\mathbf{e}$ e matrice di covarianza I). Quindi Y ha legge $\mathcal{N}_n(0, I)$ (e di conseguenza ${}^tYAY$ è libera). Si può scrivere

$$\sum_{i=1}^n a_{i,j} = \big(A\mathbf{e}\big)_j = ({}^t\mathbf{e}A)_j,$$

dato che A è simmetrica. In altri termini il vettore $\big(\sum_{i=1}^n a_{i,j}\big)_j$ non è altro che $A\mathbf{e} =^t \mathbf{e}A$.

$$\begin{aligned} {}^tXAX &=^t \big(Y + m\mathbf{e}\big)A\big(Y + m\mathbf{e}\big) =^t YAY +^t (m\mathbf{e})AY +^t YA(m\mathbf{e}) +^t (m\mathbf{e})A(m\mathbf{e}) \\ &=^t YAY + m\big({}^t\mathbf{e}AY\big) + m\big({}^tYA\mathbf{e}\big) + m^2\big({}^t\mathbf{e}A\mathbf{e}\big) =^t YAY + 2m\big({}^t\mathbf{e}AY\big) + m^2\big({}^t\mathbf{e}A\mathbf{e}\big). \end{aligned}$$

Dunque, se $\sum_{i=1}^n a_{i,j} = 0$ per ogni $j = 1, \dots, n$ (cioè se $A\mathbf{e} =^t \mathbf{e}A = 0$), si ha ${}^tXAX =^t YAY$, e quindi ${}^tXAX$ è libera. Poiché $\overline{X}$ è esaustiva (a varianza fissata, ved. esempio (d) (2) sulle statistiche esaustive) e completa, ${}^tXAX$ e $\overline{X}$ sono indipendenti per il Teorema 4.10.

Viceversa, se ${}^tXAX$ e $\overline{X}$ sono indipendenti, allora ${}^tXAX$ è libera per il Teorema 4.12. Dobbiamo far vedere che questo implica che $A\mathbf{e} = \mathbf{0}$.
Ora, ${}^tXAX$ è libera se la sua legge non dipende dal parametro m; pertanto non dipende da m neppure

$$E^m[{}^tXAX] = E^m[{}^tYAY] + 2m{}^t\mathbf{e}AE^m[Y] + m^2\big({}^t\mathbf{e}A\mathbf{e}\big)$$

Ora, dato che $E^m[{}^tYAY]$ non dipende da m e $E^m[Y] = 0$, $E^m[{}^tXAX]$ non dipende da m se e solo se ${}^t\mathbf{e}A\mathbf{e} = 0$, e da qui è semplice dedurre che $A\mathbf{e} = \mathbf{0}$. Infatti, essendo A semidefinita positiva, i suoi autovalori $\lambda_1, \dots, \lambda_n$ sono tutti non negativi; inoltre, indicata con K la matrice diagonale $K = (\lambda_k \delta_{k,r})_{k,r}$, cioè

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

esiste una matrice simmetrica ortogonale $O = (o_{i,j})_{i,j}$ tale che $A =^t OKO$. L'elemento di posto (i, j) della matrice A si scrive allora nella forma

$$a_{i,j} = \sum_{k,r} o_{i,k}\lambda_k \delta_{k,r} o_{r,j} = \sum_r o_{i,r}\lambda_r o_{r,j}.$$

Se si pone $\mathbf{v} = O\mathbf{e}$, il vettore (colonna) $\mathbf{v} =^t (v_1, \dots, v_n)$ ha la forma

$$v_r = \sum_{j=1}^{n} o_{j,r} = \sum_{j=1}^{n} o_{r,j},$$

e la relazione

$${}^t\mathbf{v}K\mathbf{v}0 =^t (O\mathbf{e})K(O\mathbf{e}) =^t \mathbf{e}({}^tOKO)\mathbf{e} =^t \mathbf{e}A\mathbf{e} = 0$$

può essere scritta nella forma

$$\sum_r \lambda_r v_r^2 = 0,$$

e, dato che tutti gli addendi di questa somma sono non negativi, si deduce che, per ogni $r = 1, \dots, n$, si ha $\lambda_r v_r^2 = 0$, e quindi anche $\lambda_r v_r = 0$. La componente j-esima del vettore $A\mathbf{e}$ è allora data da

$$\sum_{i=1}^{n} a_{i,j} = \sum_{i,r} o_{i,r}\lambda_r o_{r,j} = \sum_r o_{i,r}\lambda_r \sum_j o_{r,j} = \sum_r o_{i,r}\big(\lambda_r v_r\big) = 0.$$

Nel caso di varianza σ^2 generica, si considera il vettore $\tilde{X} = \frac{X}{\sigma}$, e si dimostra quindi che $\frac{{}^tXAX}{\sigma^2}$ e $\frac{\overline{X}}{\sigma}$ sono indipendenti se e solo se $\sum_{i=1}^n a_{i,j} = 0$, per ogni $j = 1, \dots, n$. Ovviamente $\frac{{}^tXAX}{\sigma^2}$ e $\frac{\overline{X}}{\sigma}$ sono indipendenti se e solo se altrettanto sono ${}^tXAX$ e $\overline{X}$.

5 I modelli esponenziali

Prima di introdurre i modelli esponenziali, occorre richiamare qualche proprietà della trasformata di Laplace di una misura.

Sia μ una misura σ-finita su $\big(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k)\big)$, e sia

$$D_\mu = \big\{\theta \in \mathbb{R}^k : \int \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x) < +\infty\big\}.$$

Per $\theta \in D_\mu$ definiamo

$$\mathcal{L}_\mu(\theta) = \int \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x).$$

Definizione 5.1 La funzione $\mathcal{L}_\mu : D_\mu \to \mathbb{R}^+$ definita sopra si chiama *trasformata di Laplace* della misura μ.

Elenchiamo alcune delle proprietà essenziali che ci serviranno in seguito.

(i) D_μ è un convesso di $\mathbb{R}^k$ (eventualmente vuoto);

(ii) all'interno di D_μ, l'applicazione $\theta \mapsto D_\mu(\theta)$ è di classe C^∞ ed inoltre la derivata "passa sotto il segno" di integrale, cioè

$$\frac{\partial}{\partial\theta_i}\mathcal{L}_\mu(\theta) = \frac{\partial}{\partial\theta_i}\int \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x) = \int \frac{\partial}{\partial\theta_i}\exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x) = \int x_i \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x).$$

Più in generale, se $\alpha = (\alpha_1, \dots, \alpha_k)$ è un multiindice, si ha

$$D^\alpha\mathcal{L}_\mu(\theta) = D^\alpha\int \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x) = \int D^\alpha \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x) = \int x^\alpha \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu(x),$$

dove come al solito si pone

$$D^\alpha = D_1^{\alpha_1}\cdots D_k^{\alpha_k}; \qquad x^\alpha = x_1^{\alpha_1}\cdots x_k^{\alpha_k}.$$

(iii) Siano μ e ν due misure; se esiste un aperto non vuoto $A \subseteq D_\mu \cap D_\nu$ tale che $\mathcal{L}_\mu(\theta) = \mathcal{L}_\nu(\theta)$ per ogni $\theta \in A$, allora $\mu = \nu$.

CENNO DI DIMOSTRAZIONE. (i) poiché $u \mapsto \mathrm{e}^u$ è convessa, si ha

$$\mathrm{e}^{t\langle\theta_1,x\rangle+(1-t)\langle\theta_2,x\rangle} \leq t\mathrm{e}^{\langle\theta_1,x\rangle} + (1-t)\mathrm{e}^{\langle\theta_2,x\rangle},$$

e dunque, integrando

$$\mathcal{L}_\mu(t\theta_1 + (1-t)\theta_2) \leq t\mathcal{L}_\mu(\theta_1) + (1-t)\mathcal{L}_\mu(\theta_2).$$

(ii) è immediata, a condizione di provare che, se $\theta \in D_\mu^\circ$, allora $x \mapsto x_i \mathrm{e}^{\langle\theta,x\rangle}$ è μ-integrabile.
(iii) si può dedurre dall'analogo risultato riguadante le funzioni caratteristiche (o *trasformate di Fourier*), nel modo seguente (cenno): nelle ipotesi fatte, per alcuni risultati sulle funzioni analitiche (di più variabili) le funzioni $\theta \mapsto \mathcal{L}_\mu(\theta)$ e $\theta \mapsto \mathcal{L}_\nu(\theta)$ (definite su un sottoinsieme aperto di $\mathbb{R}^k$) possono essere estese a (un sottoinsieme aperto di) $\mathbb{C}^k$. Prendendo allora $\theta = (\mathrm{i}t_1, \ldots, \mathrm{i}t_k)$, si vede che μ e ν hanno la stessa funzione caratteristica, e pertanto coincidono.

Definizione 5.2 Il modello statistico dominato $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$si dice esponenziale se esistono una misura dominante μ e una statistica T a valori in $\mathbb{R}^k$ tali che si abbia

$$L(\theta,\omega) = \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}(\omega) = C(\theta)\exp\big(\langle\theta, T(\omega)\rangle\big), \quad \forall(\theta,\omega) \in \Theta \times \Omega,$$

dove

(i) $C(\theta)$ è una costante (rispetto a ω);

(ii) Θ è un aperto convesso di $\mathbb{R}^k$, contenuto in

$$D := \Big\{\theta \in \mathbb{R}^k : \int_\Omega \exp\big(\langle\theta, T(\omega)\rangle\big)\,\mathrm{d}\mu(\omega) < +\infty\Big\}.$$

T viene detta *statistica del modello.*

Poiché si deve avere $\int L(\theta)\,\mathrm{d}\mu = 1$, si ha necessariamente

$$C(\theta) = \Big(\int_\Omega \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu\Big)^{-1}.$$

Dunque, se si pone

$$\psi(\theta) = \log\Big(\int_\Omega \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu\Big),$$

si ottiene la *forma canonica della verosimiglianza*

$$L(\theta) = \exp\Big(\langle\theta, T\rangle - \psi(\theta)\Big).$$

Osservazione 5.3 (i) T è una statistica esaustiva (ovvio per il Teorema di fattorizzazione 2.6).
(ii) Ogni modello statistico esponenziale è regolare; infatti $L(\theta)$ è strettamente positivo ovunque (se fosse $C(\theta) = 0$ per qualche valore di θ, la relativa densità $\frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}(\omega)$ sarebbe nulla per ogni $\omega \in \Omega$, e questo non è possibile).

Come abbiamo anticipato, i modelli esponenziali sono comodi per trovare statistiche complete. Vale infatti il

Teorema 5.4 *In un modello esponenziale per il quale* $L(\theta) = \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu} = C(\theta)\exp\big(\langle\theta, T\rangle\big)$, T *è una statistica completa (se* $T(\mu) = \mu_T$, *misura immagine di* T *secondo* μ, *è* σ*-finita).*

Dimostrazione. Sia $Y = h(T)$ tale che, per ogni θ, si abbia

$$0 = E^\theta[h(T)] = C(\theta)\int_\Omega h(T(\omega))\exp\big(\langle\theta, T(\omega)\rangle\big)\,\mathrm{d}\mu(\omega) = \int_{\mathbb{R}^k} h(x)\exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu_T(x)$$

(per il teorema di integrazione rispetto alla legge immagine). Scrivendo h nella forma $h^+ - h^-$ si ottiene allora

$$\int_{\mathbb{R}^k} h^+(x)\exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu_T(x) = \int_{\mathbb{R}^k} h^-(x)\exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu_T(x);$$

dunque, per la proprietà (iii) della trasformata di Laplace, le due misure $h^+.\mu_T$ e $h^-.\mu_T$ coincidono, cioè $h^+ = h^-$, μ_T-q. o. Ciò significa che $h = 0$, μ_T-q. o., ossia che $h(T) = 0$, μ-q. o.

□

Esempio 5.5 Riprendiamo l'Esempio (c) sulle statistiche esaustive. Avevamo trovato che, rispetto alla misura di Lebesgue su $[0,1]^n$, la verosimiglianza del campione è

$$L(\theta; x_1, \dots, x_n) = (\theta+1)^n\Big(\prod_{i=1}^n x_i\Big)^\theta = \exp\Big(\theta\sum_{i=1}^n \log x_i + n\log(\theta+1)\Big) = \exp\big(\theta T - \psi(\theta)\big),$$

dove

$$T(X_1, \dots, X_n) = \sum_{i=1}^n \log X_i, \qquad \psi(\theta) = -n\log(\theta+1).$$

Dunque T è una statistica esaustiva completa. Possiamo sfruttare questo fatto per calcolare uno stimatore di un'opportuna quantità $g(\theta)$, che spunterà fuori però a posteriori. Cominciamo calcolando la speranza di T rispetto a P^θ. Si ha

$$E^\theta[T] = E^\theta\Big[\sum_{i=1}^n \log X_i\Big] = \sum_{i=1}^n E^\theta\big[\log X_i\big] = nE^\theta[\log X_1] = n(\theta+1)\int_0^1 x^\theta \log x\,\mathrm{d}x$$
$$= n(\theta+1)\Big\{\Big[\frac{x^{\theta+1}}{\theta+1}\log x\Big]_0^1 - \int_0^1 \frac{x^{\theta+1}}{\theta+1}\cdot\frac{1}{x}\,\mathrm{d}x\Big\} = -n\Big[\frac{x^{\theta+1}}{\theta+1}\Big]_0^1 = -\frac{n}{\theta+1}.$$

Si deduce da questo calcolo che

$$S = -\frac{T}{n} = -\frac{\sum_{i=1}^n \log X_i}{n}$$

è uno stimatore corretto di $\frac{1}{\theta+1}$. Poichè S è funzione di T (statistica esaustiva completa) si conclude per il Teorema 4.5 che S è uno stimatore ottimale di $\frac{1}{\theta+1}$.

Osservazione 5.6 Per le stesse considerazioni, nell'Esercizio 4.3 T_2 è stimatore ottimale di $g(\theta) = \mathrm{e}^{-\theta}$.

Osservazione 5.7 Attenzione: In generale non è vero che se S è uno stimatore ottimale di $g(\theta)$ ed è assegnata una funzione ϕ, allora $\phi(S)$ è uno stimatore ottimale di $\phi(g(\theta))$. Per esempio,in riferimento all'Esempio 5.5, non è vero che $\frac{1}{S} = -\frac{n}{\sum_{i=1}^n \log X_i}$ è uno stimatore ottimale di $\theta + 1$.

Esercizio 5.8 (a) Mostrare che, in un campione di taglia n e di legge Π_θ, la media campionaria $\overline{X}$ è uno stimatore ottimale di θ.

SOLUZIONE. Abbiamo visto nell'Esempio (b) sulle statistiche esaustive che, rispetto alla misura μ che conta i punti di $\mathbb{N}^n$, la verosimiglianza è data da

$$L(\theta; k_1, \dots, k_n) = \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu} = \frac{\mathrm{e}^{-n\theta}\theta^{k_1+\cdots+k_n}}{k_1! \cdots k_n!}.$$

In questa forma il modello non è esponenziale, a causa del denominatore della frazione qui sopra. Tuttavia, per "scaricare" i fattoriali, è sufficiente cambiare la misura dominante, prendendo

$$m(k_1, \dots, k_n) = \frac{1}{k_1! \cdots k_n!},$$

rispetto alla quale la verosimiglianza diventa

$$\frac{\mathrm{d}P^\theta}{\mathrm{d}m} = \mathrm{e}^{-n\theta}\theta^{k_1+\cdots+k_n} = \exp\big((k_1 + \cdots + k_n)\log\theta - n\theta\big).$$

Posto allora $T(k_1, \dots, k_n) = k_1 + \cdots + k_n = X_1 + \cdots + X_n$, e cambiando il parametro , $t = \log\theta$, $\psi(t) = n\mathrm{e}^t$, possiamo scrivere la nuova verosimiglianza nella forma

$$L(t) = \exp\Big(t(k_1 + \cdots + k_n) - \psi(t)\Big),$$

e si riconosce quindi un modello esponenziale. Dunque la statistica $T = X_1 + \cdots + X_n$ è esaustiva completa, e, dato che

$$E^\theta[T] = E^\theta\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n E^\theta\big[X_i\big] = nE^\theta\big[X_1\big] = n\theta,$$

si ottiene, ragionando come sopra (Esempio 5.5), che $\overline{X}$ è uno stimatore ottimale di θ.

(a) Mostrare che, in un campione di taglia n e di legge $\mathcal{E}(\theta)$ (esponenziale di parametro θ), la media campionaria $\overline{X}$ è uno stimatore ottimale di $\frac{1}{\theta}$.

SOLUZIONE (identica a quella del punto (a)). In questo caso la verosimiglianza rispetto alla misura di Lebesgue su $\mathbb{R}^{+n}$ è

$$L(\theta; x_1 + \cdots + x_n) = \theta^n \mathrm{e}^{-\theta(x_1+\cdots+x_n)} = \exp\big(-\theta(x_1 + \cdots + x_n) + n\log\theta\big)$$

che è esponenziale prendendo

$$T(x_1, \dots, x_n) = -(x_1 + \cdots + x_n),$$

e dunque $T = X_1 + \cdots + X_n$ è statistica esaustiva completa. Si ha poi subito $E^\theta[T] = \frac{n}{\theta}$, e dunque $\overline{X}$ è uno stimatore ottimale di $\frac{1}{\theta}$.

Osservazione 5.9 Come abbiamo appena visto nell'Esercizio 5.8 (a), talvolta il modello non si presenta in forma esponenziale, ma lo diventa se si effettua un opportuno cambio di parametro $t = g(\theta)$.

Osservazione 5.10 Sia (μ^θ) una famiglia esponenziale di probabilità su $\mathbb{R}$, cioè tale che per un'opportuna misura dominante μ, si abbia

$$\frac{\mathrm{d}\mu^\theta}{\mathrm{d}\mu}(x) = \exp\big(\langle\theta, T(x)\rangle - \psi(\theta)\big).$$

Sia ora $X = (X_1, \dots, X_n)$ un campione di legge μ^θ; allora la verosimiglianza rispetto alla misura dominante $\mu^{\otimes n}$ è data da

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n \exp\big(\langle \theta, T(x_i)\rangle - \psi(\theta)\big) = \exp\Big(\langle \theta, \sum_{i=1}^n T(x_i)\rangle - n\psi(\theta)\Big);$$

quindi anche il campione è un modello esponenziale e, se si pone come al solito $X_i(x_1, \dots, x_n) = X_i$, la statistica del modello è

$$T = \sum_{i=1}^n T(X_i),$$

e dunque è una statistica esaustiva completa.

ESEMPI DI MODELLI ESPONENZIALI SU $\mathbb{R}$

Vediamo allora alcuni modelli esponenziali su $\mathbb{R}$. Come detto nell'Osservazione 5.10, i relativi campioni sono ancora modelli esponenziali.

(a) Legge $\mathcal{B}(1, \theta)$. La densità (rispetto alla misura che assegna massa unitaria ai punti 0 e 1) è data da

$$L(\theta; k) = \theta^k (1-\theta)^{1-k}) = \exp\Big(k \log \frac{\theta}{1-\theta} + \log(1-\theta)\Big),$$

che diventa esponenziale passando al parametro

$$t = \log \frac{\theta}{1-\theta}$$

(Osservazione 5.9).

(b) Legge di Poisson di parametro θ. Abbiamo già trattato questo caso nell'Esercizio 5.8 (a).

(c) Legge $\Gamma(\alpha, \lambda)$ ($\alpha > 0$, $\lambda > 0$). La verosimiglianza rispetto alla misura di Lebesgue su $\mathbb{R}^+$ è

$$L(\alpha, \lambda; x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\lambda x} = \exp\big(-\lambda x + (\alpha - 1)\log x + \alpha \log \lambda - \log \Gamma(\alpha)\big),$$

che, nel caso di parametri entrambi sconosciuti, diventa esponenziale se si pone

$$\theta = (-\lambda, \alpha - 1), \qquad T(x) = (x, \log x).$$

Se solo λ non è noto, si ha evidentemente

$$\theta = -\lambda, \qquad T(x) = x,$$

mentre, se non è noto solo α, si pone

$$\theta = \alpha - 1, \qquad T(x) = \log x.$$

Vediamo ora qual è il legame tra la statistica T del modello e la funzione ψ. Si ha il seguente risultato:

Teorema 5.11 *Valgono le seguenti equazioni*

(a) $$\frac{\partial\psi(\theta)}{\partial\theta_i} = E^\theta[T_i], \qquad i = 1, \dots, k;$$

(b) $$\frac{\partial^2\psi(\theta)}{\partial\theta_i\partial\theta_j} = Cov^\theta(T_i, T_j), \qquad i, j = 1, \dots, k.$$

Dimostrazione. (a) Si ha

$$\frac{\partial\psi(\theta)}{\partial\theta_i} = \frac{\partial}{\partial\theta_i}\log\Big(\int_\Omega \exp\big(\langle\theta, T\rangle\big)\Big) = \Big(\int_\Omega \exp\big(\langle\theta, T\rangle\big)\Big)^{-1}\cdot\frac{\partial}{\partial\theta_i}\int_\Omega \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu$$
$$= \mathrm{e}^{-\psi(\theta)}\cdot\frac{\partial}{\partial\theta_i}\int_\Omega \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu. \tag{4}$$

D'altra parte

$$\int_\Omega \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu = \int_{\mathbb{R}^k} \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu_T(x),$$

dove μ_T è l'immagine di μ secondo T (μ_T è supposta come al solito σ-finita). Dunque, per la prorietà (ii) della trasformata di Laplace sopra ricordata, si ha

$$\frac{\partial}{\partial\theta_i}\int_\Omega \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu = \frac{\partial}{\partial\theta_i}\int_{\mathbb{R}^k} \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu_T(x) = \int_{\mathbb{R}^k} x_i \exp\big(\langle\theta, x\rangle\big)\,\mathrm{d}\mu_T(x)$$
$$= \int_\Omega T_i \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu.$$

Inserendo questa uguaglianza nella (4) troviamo allora

$$\frac{\partial\psi(\theta)}{\partial\theta_i} = \mathrm{e}^{-\psi(\theta)}\cdot\frac{\partial}{\partial\theta_i}\int_\Omega \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu = \mathrm{e}^{-\psi(\theta)}\int_\Omega T_i \exp\big(\langle\theta, T\rangle\big)\,\mathrm{d}\mu$$
$$= \int_\Omega T_i \exp\big(\langle\theta, T\rangle - \psi(\theta)\big)\,\mathrm{d}\mu = \int_\Omega T_i L(\theta)\,\mathrm{d}\mu = \int_\Omega T_i\,\mathrm{d}P^\theta = E^\theta[T_i].$$

(b) Continuando a derivare, si trova

$$\frac{\partial^2\psi(\theta)}{\partial\theta_i\partial\theta_j} = \frac{\partial}{\partial\theta_j}\Big\{\int_\Omega T_i \exp\big(\langle\theta, T\rangle - \psi(\theta)\big)\,\mathrm{d}\mu\Big\} = \int_\Omega T_i\frac{\partial}{\partial\theta_j}\exp\big(\langle\theta, T\rangle - \psi(\theta)\big)\,\mathrm{d}\mu$$
$$= \int_\Omega T_i \exp\big(\langle\theta, T\rangle - \psi(\theta)\big)\frac{\partial}{\partial\theta_j}\big(\langle\theta, T\rangle - \psi(\theta)\big)\mathrm{d}\mu = \int_\Omega T_i \exp\big(\langle\theta, T\rangle - \psi(\theta)\big)\Big(T_j - \frac{\partial\psi(\theta)}{\partial\theta_j}\Big)\,\mathrm{d}\mu$$
$$= \int_\Omega T_i \exp\big(\langle\theta, T\rangle - \psi(\theta)\big)\Big(T_j - E^\theta[T_j]\Big)\,\mathrm{d}\mu = \int_\Omega T_i\Big(T_j - E^\theta[T_j]\Big)\,\mathrm{d}P^\theta = E^\theta[T_iT_j] - E^\theta[T_i]E^\theta[T_j]$$
$$= Cov^\theta(T_i, T_j).$$

6 L'informazione secondo Fisher e la disuguaglianza di Cramer-Rao

Quando si deve stimare il parametro $\theta \in \Theta$ (con Θ aperto di $\mathbb{R}^k$), l'importante è il tipo di variabilità delle leggi P^θ intorno al vero valore del parametro θ_0. Di qui l'idea di avere un'"informazione locale". A questa esigenza risponde il concetto di *informazione* introdotto da *Fisher*. Questa

affermazione sarà chiara quando, dopo l'informazione di Fisher, avremo introdotto anche il concetto di *informazione di Kullback* (nel paragrafo seguente).

In tutta questa sezione supporremo assegnato un modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$, dominato dalla misura μ, e sia L una scelta della verosimiglianza rispetto a μ:

$$L(\theta, \omega) = \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}.$$

Indicheremo con E l'operatore di speranza fatta rispetto a μ. Si suppone che Θ sia un aperto di $\mathbb{R}^k$, e che la funzione $\theta \mapsto L(\theta, \omega)$ sia strettamente positiva per ogni $\omega \in \Omega$.

Osservazione 6.1 A PROPOSITO DELL'IPOTESI DI STRETTA POSITIVITÀ DI $\theta \mapsto L(\theta, \omega)$. Un esempio è quello di un campione di legge gaussiana: la densità gaussiana è strettamente positiva su tutto $\mathbb{R}$, dunque la verosimiglianza del campione è strettamente positiva su tutto $\mathbb{R}^n$. Un altro esempio è quello di un campione di legge avente densità $(\theta+1)x^\theta 1_{[0,1]}(x)$: in questo caso $L(\theta, \omega) > 0$ per ogni $\omega \in [0,1]^n$ soltanto (e non su tutto $\mathbb{R}^n$); tuttavia, in questo caso il modello del campione può essere definito prendendo $\Omega = [0,1]^n$ anziché $\mathbb{R}^n$ e in questo modo l'ipotesi richiesta è verificata. Per questo motivo, pensando il modello definito comunque con $\Omega = \mathbb{R}^n$, alcuni la formulano nella forma seguente: *l'insieme* $A_\theta = \{\omega \in \Omega : L(\theta, \omega) > 0\}$ *non dipende da* θ.

Osserviamo che l'ipotesi non è valida per esempio nel caso di un campione di legge $\mathcal{U}[(0, \theta)]$, poiché in questo caso il modello deve obbligatoriamente essere definito con $\Omega = \mathbb{R}^n$.

Osserviamo anche che con questa ipotesi il modello $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$risulta essere un modello regolare.

Supporremo inoltre che si possa scambiare l'operazione di integrazione con con quella di derivazione nel derivare le funzioni del tipo $\theta \mapsto E[L(\theta)Y]$, con Y v.a. in $L^2(\mu)$ (questa ipotesi è vera ad esempio se $\nabla L(\theta)$ è di quadrato integrabile secondo μ). In particolare, poiché si ha

$$E[L(\theta)] = \int_\Omega \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}\,\mathrm{d}\mu = \int_\Omega \mathrm{d}P^\theta = 1,$$

derivando e sfruttando l'ipotesi appena fatta si trova, per ogni $i = 1, \ldots, k$

$$E\Big[\frac{\partial}{\partial\theta_i} L(\theta)\Big] = \frac{\partial}{\partial\theta_i} E[L(\theta)] = \frac{\partial}{\partial\theta_i} 1 = 0,$$

che, in forma vettoriale, significa

$$E[\nabla L(\theta)] = \nabla E[L(\theta)] = \nabla \mathbf{1} = \mathbf{0},$$

dove $\mathbf{1}$ (risp. $\mathbf{0}$ è il vettore con componenti tutte uguali a 1 (risp. tutte nulle).
Da quest'ultima relazione segue che

$$E^\theta[\nabla \log L(\theta)] = E^\theta\Big[\frac{1}{L(\theta)}\nabla L(\theta)\Big] = \int_\Omega \frac{1}{L(\theta)}\nabla L(\theta)\,\mathrm{d}P^\theta = \int_\Omega \frac{1}{L(\theta)}\nabla L(\theta)\,\underbrace{\frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}}_{=L(\theta)}\,\mathrm{d}\mu$$

$$= \int_\Omega \nabla L(\theta)\,\mathrm{d}\mu = E[\nabla L(\theta)] = \mathbf{0}.$$

In altre parole, questa relazione dice che , per ogni $i = 1, \ldots, k$, le v.a.

$$\omega \mapsto \frac{\partial}{\partial\theta_i} \log L(\theta, \omega)$$

sono centrate rispetto a P^θ, per ogni θ.

Come ultima ipotesi, supporremo che il vettore aleatorio $\nabla \log L(\theta)$ sia di quadrato integrabile secondo P^θ, per ogni θ.

A questo punto possiamo dare la

Definizione 6.2 Si chiama *matrice di informazione secondo Fisher* la matrice $I(\theta) = \big(I(\theta)_{i,j}\big)$ dove

$$I(\theta)_{i,j} = E^\theta\Big[\frac{\partial}{\partial\theta_i}\log L(\theta)\cdot\frac{\partial}{\partial\theta_j}\log L(\theta)\Big], \qquad i,j=1,\dots,k.$$

Osservazione 6.3 Dato che, per ogni $i=1,\dots,k$

$$\frac{\partial}{\partial\theta_i}\log L(\theta) = \frac{1}{L(\theta)}\cdot\frac{\partial L(\theta)}{\partial\theta_i},$$

si ha la formula alternativa

$$I(\theta)_{i,j} = E^\theta\Big[\frac{1}{L^2(\theta)}\cdot\frac{\partial L(\theta)}{\partial\theta_i}\cdot\frac{\partial L(\theta)}{\partial\theta_j}\Big].$$

Osservazione 6.4 $I(\theta)$ è una matrice simmetrica e semi-definita positiva. Infatti essa è la matrice di covarianza del vettore $\nabla \log L(\theta)$ (che come abbiamo visto è centrato). D'altra parte, la matrice di covarianza C di un vettore aleatorio $(X_1,\dots,X_k)$ (con X_i definita su $(\Omega,\mathcal{F},P)$) è sempre simmetrica (ovvio) e semi-definita positiva. Infatti

$$\begin{aligned}
\langle Cx,x\rangle &= \sum_{i,j} Cov(X_i,X_j)x_ix_j = \sum_{i,j} E\big[x_i(X_i-E[X_i])x_j(X_j-E[X_j])\big]\\
&= E\Big[\sum_{i,j} x_i(X_i-E[X_i])x_j(X_j-E[X_j])\Big]\\
&= E\Big[\Big(\sum_i x_i(X_i-E[X_i])\Big)\Big(\sum_j x_j(X_j-E[X_j])\Big)\Big]\\
&= E\Big[\Big(\sum_i x_i(X_i-E[X_i])\Big)^2\Big] = E\big[\langle x, X-E[X]\rangle^2\big] \geq 0.
\end{aligned}$$

Né la definizione né la forma alternativa sono comode per calcolare l'informazione di Fisher. Vediamo dunque una formula più pratica. Premettiamo il

Lemma 6.5 *Per ogni $i,j=1,\dots,k$ si ha*

$$E^\theta\Big[\frac{1}{L(\theta)}\cdot\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}\Big] = 0.$$

Dimostrazione.

$$\begin{aligned}
E^\theta\Big[\frac{1}{L(\theta)}\cdot\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}\Big] &= \int \frac{1}{L(\theta)}\cdot\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}\,\mathrm{d}P^\theta = \int \frac{1}{L(\theta)}\cdot\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}\,\underbrace{\frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}}_{=L(\theta)}\,\mathrm{d}\mu\\
&= \int \frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}\,\mathrm{d}\mu = E\Big[\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}\Big] = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\underbrace{E\Big[L(\theta)\Big]}_{=1} = 0.
\end{aligned}$$

□

Proposition 6.6 *Vale la formula*

$$I(\theta)_{i,j} = -E^\theta\Big[\frac{\partial^2 \log L(\theta)}{\partial\theta_i\partial\theta_j}\Big].$$

Dimostrazione. Si ha

$$\frac{\partial^2 \log L(\theta)}{\partial\theta_i\partial\theta_j} = \frac{\partial}{\partial\theta_i}\Big(\frac{\partial \log L(\theta)}{\partial\theta_j}\Big) = \frac{\partial}{\partial\theta_i}\Big(\frac{1}{L(\theta)}\cdot\frac{\partial L(\theta)}{\partial\theta_j}\Big) = \frac{1}{L^2(\theta)}\Big(L(\theta)\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j} - \frac{\partial L(\theta)}{\partial\theta_i}\cdot\frac{\partial L(\theta)}{\partial\theta_j}\Big)$$
$$= \frac{1}{L(\theta)}\cdot\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j} - \frac{1}{L^2(\theta)}\cdot\frac{\partial L(\theta)}{\partial\theta_i}\cdot\frac{\partial L(\theta)}{\partial\theta_j}.$$

Passando alle speranze e utilizzando il Lemma 6.5 si trova

$$E^\theta\Big[\frac{\partial^2 \log L(\theta)}{\partial\theta_i\partial\theta_j}\Big] = -E^\theta\Big[\frac{1}{L^2(\theta)}\cdot\frac{\partial L(\theta)}{\partial\theta_i}\cdot\frac{\partial L(\theta)}{\partial\theta_j}\Big] = -I(\theta)_{i,j},$$

dove l'ultima uguaglianza segue dall'Osservazione 6.4.

□

Vogliamo ora vedere come si comporta l'informazione quando il numero di esperienze fatte aumenta.

Il caso di prove indipendenti. Per prima cosa, vediamo cosa accade quando le varie esperienze sono tra loro indipendenti, formalizzando la situazione così: si hanno due modelli statistici dominati $\big(\Omega_1, \mathcal{F}_1, \{P_1^\theta, \theta\in\Theta\}\big)$, con misura dominante μ_1 e verosimiglianza $L_1(\theta)$ (primo esperimento) e $\big(\Omega_2, \mathcal{F}_2, \{P_2^\theta, \theta\in\Theta\}\big)$, con misura dominante μ_2 e verosimiglianza $L_2(\theta)$ (secondo esperimento). Per modellizzare l'esperimento composto, poniamo $\Omega = \Omega_1\times\Omega_2$; $\mathcal{F} = \mathcal{F}_1\otimes\mathcal{F}_2$; $P^\theta = P_1^\theta\otimes P_2^\theta$ ($\omega\in\Omega$ è del tipo $\omega = (\omega_1,\omega_2)$, con $\omega_1\in\Omega_1$ e $\omega_2\in\Omega_2$); è immediato vedere che il modello $\big(\Omega, \mathcal{F}, \{P^\theta, \theta\in\Theta\}\big)$ora costruito è dominato da $\mu_1\otimes\mu_2$ e che una versione della verosimiglianza è data da $L(\theta,\omega) = L_1(\theta,\omega_1)\cdot L_2(\theta,\omega_2)$. Denotiamo (come al solito) con X_1e X_2 le proiezioni ($X_1(\omega) = \omega_1$ e ($X_1(\omega) = \omega_2$), che risultano indipendenti per costruzione. Indichiamo infine con I, I_1 e I_2 le informazioni di Fisher nei relativi modelli (con notazioni ovvie).

Lemma 6.7 *Su $(\Omega,\mathcal{F},P)$ siano U e V due vettori aleatori a valori in $\big(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k)\big)$ tra loro indipendenti. Indichiamo con $\mathbf{Cov}U$, $\mathbf{Cov}V$ e $\mathbf{Cov}(U+V)$ le matrici di covarianza di U, V e $U+V$ rispettivamente. Si ha allora*

$$\mathbf{Cov}(U+V) = \mathbf{Cov}U + \mathbf{Cov}V.$$

La dimostrazione è conseguenza immediata del caso $k=1$, ben noto.

Teorema 6.8 (di additività). *Nelle ipotesi fatte sopra, per ogni $\theta\in\Theta$ si ha*

$$I(\theta) = I_1(\theta) + I_2(\theta).$$

Dimostrazione. Indicheremo con $X(\omega)=\omega$ l'applicazione identica di Ω in se stesso. Si ha prima di tutto

$$\log L(\theta, X) = \log\big(L_1(\theta, X_1)\cdot L_2(\theta, X_2)\big) = \log L_1(\theta, X_1) + \log L_2(\theta, X_2).$$

Dunque, ricordando che $I(\theta)$ è la matrice di covarianza del vettore $\nabla\log(\theta, X)$ (Osservazione 6.4), si ha

$$I(\theta) = \mathbf{Cov}\big(\nabla\log L(\theta, X)\big) = \mathbf{Cov}\big(\nabla\{\log L_1(\theta, X_1) + \log L_2(\theta, X_2)\}\big)$$
$$= \mathbf{Cov}\big(\nabla\log L_1(\theta, X_1) + \nabla\log L_2(\theta, X_2)\big)$$
$$= \mathbf{Cov}\big(\nabla\log L(\theta, X_1)\big) + \mathbf{Cov}\big(\nabla\log L(\theta, X_2)\big) = I_1(\theta) + I_2(\theta),$$

per l'indipendenza di X_1 e X_2 (Lemma 6.7).

□

Osservazione 6.9 (a) Naturalmente il Teorema precedente può essere formulato (in modo ovvio) anche quando i modelli di partenza sono più di due.

Il significato del Teorema è che, in caso di indipendenza, le informazioni si sommano; in particolare l'informazione $I_n(\Theta)$ fornita da un campione $(X_1, \dots, X_n)$ (di legge μ^θ), e cioè l'informazione del modello

$$\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{(\mu^\theta)^{\otimes n}, \theta \in \Theta\}\big)$$

è uguale a n volte l'informazione I fornita dal modello

$$\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mu^\theta, \theta \in \Theta\}\big).$$

Cioè l'informazione fornita da n prove i.i.d. è n volte l'informazione fornita da una singola prova. ,

Il caso di una catena di Markov. Un'altra situazione interessante in cui studiare come varia l'informazione all'aumentare delle osservazioni è quella del modello statistico dell'Esempio 1.23 (catena di Markov). Supponiamo di avere una catena di Markov sullo spazio misurabile $(E, \mathcal{E})$ di legge iniziale ρ assegnata e operatore di transizione

$$\Pi(\theta, x; A) = \int_A \ell(\theta, x, y)\Pi(x, \mathrm{d}y),$$

dove $\Theta \subseteq \mathbb{R}$, Π è un operatore di transizion fissato di $(E, \mathcal{E})$ in $(E, \mathcal{E})$ e $\{\ell(\theta, \cdot, \cdot)\}_{\theta\in\Theta}$ è una famiglia di verosimiglianze. Sappiamo che il modello statistico (definito in 1.23)

$$\big(E^{n+1}, \mathcal{E}^{\otimes n+1}, \{P^\theta, \theta \in \Theta\}\big), \qquad \text{dove} \qquad P^\theta = \rho \otimes (\Pi^\theta)^{\otimes n}$$

è un modello dominato; una misura dominante è $\rho \otimes \Pi^{\otimes n}$ e una versione della verosimiglianza è $(x = (x_0, \dots, x_n))$

$$L(\theta, x) = \ell(\theta, x_0, x_1) \cdot \ell(\theta, x_1, x_2) \cdot \dots \cdot \ell(\theta, x_n, x_{n+1}).$$

Calcoliamo $I_n(\theta)$ con la formula alternativa dell'Osservazione 6.3, cioè

$$I_2(\theta) = E\Big[\Big\{\frac{L'(\theta)}{L(\theta)}\Big\}^2\Big].$$

Per semplicità consideriamo solo il caso $n = 2$. Poniamo

$$\dot\ell = \frac{\partial d}{\partial\theta}\ell(\theta, x, y).$$

Allora $L(\theta, X_0, X_1, X_2) = \ell(\theta, X_0, X_1)\ell(\theta, X_1, X_2)$ e

$$L'(\theta) = \dot\ell(\theta, X_0, X_1)\ell(\theta, X_1, X_2) + \ell(\theta, X_0, X_1)\dot\ell(\theta, X_1, X_2)$$

e quindi

$$\frac{L'(\theta)}{L(\theta)} = \frac{\dot\ell(\theta, X_0, X_1)}{\ell(\theta, X_0, X_1)} + \frac{\dot\ell(\theta, X_1, X_2)}{\ell(\theta, X_1, X_2)}.$$

Dunque

$$\Big\{\frac{L'(\theta)}{L(\theta)}\Big\}^2 = \Big\{\frac{\dot\ell(\theta, X_0, X_1)}{\ell(\theta, X_0, X_1)}\Big\}^2 + \Big\{\frac{\dot\ell(\theta, X_1, X_2)}{\ell(\theta, X_1, X_2)}\Big\}^2 + 2\frac{\dot\ell(\theta, X_0, X_1)}{\ell(\theta, X_0, X_1)} \cdot \frac{\dot\ell(\theta, X_1, X_2)}{\ell(\theta, X_1, X_2)},$$

e passando alle speranze si trova

$$I_2(\theta) = E^\theta\Big[\Big\{\frac{\dot\ell(\theta, X_0, X_1)}{\ell(\theta, X_0, X_1)}\Big\}^2\Big] + E^\theta\Big[\Big\{\frac{\dot\ell(\theta, X_1, X_2)}{\ell(\theta, X_1, X_2)}\Big\}^2\Big] + 2E^\theta\Big[\frac{\dot\ell(\theta, X_0, X_1)}{\ell(\theta, X_0, X_1)} \cdot \frac{\dot\ell(\theta, X_1, X_2)}{\ell(\theta, X_1, X_2)}\Big].$$

D'altra parte

$$E^\theta\Big[\frac{\dot\ell(\theta,X_0,X_1)}{\ell(\theta,X_0,X_1)}\cdot\frac{\dot\ell(\theta,X_1,X_2)}{\ell(\theta,X_1,X_2)}\Big]=E^\theta\bigg[E^\theta\Big[\frac{\dot\ell(\theta,X_0,X_1)}{\ell(\theta,X_0,X_1)}\cdot\frac{\dot\ell(\theta,X_1,X_2)}{\ell(\theta,X_1,X_2)}\Big|X_0,X_1\Big]\bigg]$$
$$=E^\theta\bigg[\frac{\dot\ell(\theta,X_0,X_1)}{\ell(\theta,X_0,X_1)}E^\theta\Big[\frac{\dot\ell(\theta,X_1,X_2)}{\ell(\theta,X_1,X_2)}\Big|X_0,X_1\Big]\bigg]=0,$$

perché

$$E^\theta\Big[\frac{\dot\ell(\theta,X_1,X_2)}{\ell(\theta,X_1,X_2)}\Big|X_0,X_1\Big]=\int\frac{\dot\ell(\theta,X_1,y)}{\ell(\theta,X_1,y)}\Pi^\theta(X_1,\mathrm{d}y)=\int\frac{\dot\ell(\theta,X_1,y)}{\ell(\theta,X_1,y)}\ell(\theta,X_1,y)\Pi(X_1,\mathrm{d}y)$$
$$=\int\dot\ell(\theta,X_1,y)\Pi(X_1,\mathrm{d}y)=\frac{\partial}{\partial\theta}\int\ell(\theta,X_1,y)\Pi(X_1,\mathrm{d}y)=\frac{\partial}{\partial\theta}\underbrace{\int\Pi^\theta(X_1,\mathrm{d}y)}_{=1}=0.$$

Si ottiene dunque che

$$I_2(\theta)=E^\theta\Big[\Big\{\frac{\dot\ell(\theta,X_0,X_1)}{\ell(\theta,X_0,X_1)}\Big\}^2\Big]+E^\theta\Big[\Big\{\frac{\dot\ell(\theta,X_1,X_2)}{\ell(\theta,X_1,X_2)}\Big\}^2\Big].$$

In generale, se ϕ è una funzione di due variabili limitata, si ha

$$E^\theta\big[\phi(X_0,X_1)\big]=\int\rho(\mathrm{d}x_0)\int\Pi^\theta(x_0,\mathrm{d}x_1)\phi(x_0,x_1)=\int\phi(x_0,x_1)\mathrm{d}\big(\rho\otimes\Pi^\theta\big)(x_0,x_1)$$

e

$$E^\theta\big[\phi(X_1,X_2)\big]=\int\rho(\mathrm{d}x_0)\int\Pi^\theta(x_0,\mathrm{d}x_1)\int\Pi^\theta(x_1,\mathrm{d}x_2)\phi(x_1,x_2)$$
$$=\int\phi(x_1,x_2)\mathrm{d}\big(\rho\otimes(\Pi^\theta)^{\otimes2}\big)(x_0,x_1,x_2).$$

Inoltre, se ρ è una misura invariante per la catena, la legge di X_1, e cioè l'applicazione

$$A\mapsto\int\rho(\mathrm{d}x_0)\Pi^\theta(x_0,A)$$

coincide con ρ. Pertanto

$$E^\theta\big[\phi(X_1,X_2)\big]=\int\rho(\mathrm{d}x_0)\int\Pi^\theta(x_0,\mathrm{d}x_1)\int\Pi^\theta(x_1,\mathrm{d}x_2)\phi(x_1,x_2)$$
$$=\int\Big\{\int\rho(\mathrm{d}x_0)\Pi^\theta(x_0,\mathrm{d}x_1)\Big\}\int\Pi^\theta(x_1,\mathrm{d}x_2)\phi(x_1,x_2)=\int\rho(\mathrm{d}x_1)\int\Pi^\theta(x_1,\mathrm{d}x_2)\phi(x_1,x_2)$$
$$=E^\theta\big[\phi(X_0,X_1)\big].$$

Si può dunque scrivere

$$E^\theta\Big[\Big\{\frac{\dot\ell(\theta,X_0,X_1)}{\ell(\theta,X_0,X_1)}\Big\}^2\Big]=\int\Big\{\frac{\dot\ell(\theta,x_0,x_1)}{\ell(\theta,x_0,x_1)}\Big\}^2\mathrm{d}\big(\rho\otimes\Pi^\theta\big)(x_0,x_1)$$

e

$$E^\theta\Big[\Big\{\frac{\dot\ell(\theta,X_1,X_2)}{\ell(\theta,X_1,X_2)}\Big\}^2\Big]=\int\Big\{\frac{\dot\ell(\theta,x_1,x_2)}{\ell(\theta,x_1,x_2)}\Big\}^2\mathrm{d}\big(\rho\otimes(\Pi^\theta)^{\otimes2}\big)(x_0,x_1,x_2)$$

; quindi

$$I_2(\theta) = \int \Big\{\frac{\dot\ell(\theta, x_0, x_1)}{\ell(\theta, x_0, x_1)}\Big\}^2 \mathrm{d}\big(\rho \otimes \Pi^\theta\big)(x_0, x_1) + \int \Big\{\frac{\dot\ell(\theta, x_1, x_2)}{\ell(\theta, x_1, x_2)}\Big\}^2 \mathrm{d}\big(\rho \otimes (\Pi^\theta)^{\otimes 2}\big)(x_0, x_1, x_2)$$

e, se ρ è invariante per la catena, si trova

$$I_2(\theta) = 2I_1(\theta).$$

In generale ($n \geq 1$) otteniamo

$$I_n(\theta) = \sum_{i=1}^n E^\theta\Big[\Big\{\frac{\dot\ell(\theta, X_{i-1}, X_i)}{\ell(\theta, X_{i-1}, X_i)}\Big\}^2\Big] = \sum_{i=1}^n \int\int \Big\{\frac{\dot\ell(\theta, x_{i-1}, x_i)}{\ell(\theta, x_{i-1}, x_i)}\Big\}^2 \mathrm{d}\big(\rho \otimes (\Pi^\theta)^{\otimes i}\big)(x_0, x_1, \ldots, x_{i-1,x_i});$$

inotre, se ρ è invariante per a catena, si trova

$$I_n(\theta) = nE^\theta\Big[\Big\{\frac{\dot\ell(\theta, X_0, X_1)}{\ell(\theta, X_0, X_1)}\Big\}^2\Big] = nI_1(\theta).$$

Si nota dunque che l'informazione cresce al crescere del numero n di passi osservati; se per di più ρ è invariante per la catena, si ritrova la stessa situazione che avevamo visto nel caso di un campione: l'informazione ottenuta dopo n passi della catena è n volte quella ottenuta dopo un passo.
In ogni caso, il concetto di informazione di Fisher è in accordo con l'intuizione per quanto riguarda l'aumento delle osservazioni: più osservazioni, più informazione.

Esempio 6.10 (a) Sia $(X_1, \ldots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$, dove σ^2 è noto. Calcoliamo $I_n(m)$ (la notazione $I_n(m)$, così come la successiva $L(m)$, sta a rammentare che che in questo momento siamo interessati a considerare queste quantità come funzioni di m, che è il parametro che ci interessa). Per il calcolo, useremo la Proposizione 6.6.
La verosimiglianza è data da

$$L(m; x_1, \ldots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}}\sigma^{-n} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - m)^2\Big),$$

dunque

$$\log L(m) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - m)^2;$$

quindi

$$\frac{\mathrm{d}^2}{\mathrm{d}m^2}\log L(m) = \frac{\mathrm{d}^2}{\mathrm{d}m^2}\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - m)^2\Big\} = \frac{\mathrm{d}}{\mathrm{d}m}\Big\{\frac{1}{\sigma^2}\sum_{i=1}^n (x_i - m)\Big\} = -\frac{n}{\sigma^2}.$$

Per la Proposizione 6.6 si ha allora

$$I_n(m) = -E^m\Big[\frac{\mathrm{d}^2}{\mathrm{d}m^2}\log L(m)\Big] = \frac{n}{\sigma^2}.$$

Osserviamo comunque che, per il Teorema di additività 6.8, si ha $I_n(m) = nI_1(m)$, e quindi sarebbe bastato calcolare $I_1(m) = \frac{1}{\sigma^2}$.

(b) Come in (a), sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$, dove però questa volta consideriamo noto m. Calcoliamo $I_n(\sigma^2)$. La verosimiglianza è la stessa che in (a), ma questa volta va guardata come funzione di σ^2, quindi scriviamo

$$\log L(\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2;$$

da cui

$$\frac{\mathrm{d}^2}{\mathrm{d}(\sigma^2)^2}\log L(\sigma^2) = \frac{\mathrm{d}^2}{\mathrm{d}(\sigma^2)^2}\Big\{ -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\Big\}$$
$$= \frac{\mathrm{d}}{\mathrm{d}(\sigma^2)}\Big\{ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - m)^2\Big\} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - m)^2.$$

Quindi, sempre per la Proposizione 6.6,

$$I_n(\sigma^2) = -E^{\sigma^2}\Big[\frac{\mathrm{d}^2}{\mathrm{d}(\sigma^2)^2}\log L(\sigma^2)\Big] = -E^{\sigma^2}\Big[\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i - m)^2\Big]$$
$$= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^4}\sum_{i=1}^{n}\underbrace{E^{\sigma^2}\Big[\Big(\frac{X_i - m}{\sigma}\Big)^2\Big]}_{=Var\left(\frac{X_i - m}{\sigma}\right)=1} = -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} = \frac{n}{2\sigma^4}.$$

(c) Ancora come in (a), sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$. desso calcoliamo $I_n(\sigma)$ (il significato della notazione dovrebbe essere ormai chiaro). Ora scriviamo

$$\log L(\sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2;$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\log L(\sigma) = \frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\Big\{ -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\Big\}$$
$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\Big\{ -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - m)^2\Big\} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i - m)^2;$$

$$I_n(\sigma) = -E^{\sigma}\Big[\frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\log L(\sigma)\Big] = -E^{\sigma}\Big[\frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i - m)^2\Big]$$
$$= -\frac{n}{\sigma^2} + \frac{3}{\sigma^2}\sum_{i=1}^{n}\underbrace{E^{\sigma}\Big[\Big(\frac{X_i - m}{\sigma}\Big)^2\Big]}_{=Var\left(\frac{X_i - m}{\sigma}\right)=1} = -\frac{n}{\sigma^2} + \frac{3n}{\sigma^2} = \frac{2n}{\sigma^2}.$$

In tutti gli esempi precedenti sono valide le ipotesi per poter sviluppare la teoria dell'informazione di Fisher; dunque, fra l'altro, come abbiamo potuto verificare con i calcoli diretti, vale il Teorema di additività. Vediamo cosa accade nell'esempio che segue, in cui, come abbiamo osservato in 6.1, le ipotesi non valgono

Esempio 6.11 Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{U}[(0,\theta)]$ $(\theta > 0)$.

(i) Calcoliamo $I_1(\theta)$ (informazione relativa a X_i). La verosimiglianza è

$$L(\theta, x) = \frac{1}{\theta} 1_{(0,\theta)}(x).$$

Dunque $A_\theta = \{x : L(\theta, x) > 0\} = (0, \theta)$ e, per $x \in A_\theta$, abbiamo

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta, x) = -\frac{1}{\theta};$$

in altre parole

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta, x) = -\frac{1}{\theta} 1_{(0,\theta)}(x)$$

e quindi

$$I_1(\theta) = E^\theta\Big[\Big\{\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta, X_i)\Big\}^2\Big] = E^\theta\Big[\Big\{-\frac{1}{\theta} 1_{(0,\theta)}(X_i)\Big\}^2\Big] = \frac{1}{\theta^2} E^\theta\Big[1_{(0,\theta)}(X_i)\Big] = \frac{1}{\theta^2}\int_0^\theta \frac{1}{\theta}\,\mathrm{d}x = \frac{1}{\theta^2}.$$

(ii) Proviamo ora ad usare la formula della Proposizione 6.6:

$$I_1(\theta) = -E^\theta\Big[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log L(\theta)\Big] = -E^\theta\Big[\frac{1}{\theta^2} 1_{(0,\theta)}(X_i)\Big] = -\frac{1}{\theta^2}.$$

Si osserva che in questo modo non otteniamo il valore corretto (quello del punto (i)); fra l'altro si tratta di un numero negativo (viceversa, sappiamo che l'informazione è sempre non negativa).

(iii)Verifichiamo che non vale il Teorema di additività 6.8, calcolando $I_n(\theta)$ (informazione relativa al campione). La verosimiglianza del campione è

$$L(\theta; x, \dots, x_n) = \frac{1}{\theta^n} 1_{(0,\theta)}(x_1) \cdot \dots \cdot 1_{(0,\theta)}(x_n).$$

Dunque

$$A_\theta = \{x_1, \dots, x_n : L(\theta; x, \dots, x_n) > 0\} = (0, \theta)^n;$$
$$\log L(\theta; x, \dots, x_n) = (-n \log \theta) 1_{(0,\theta)^n}(x_1, \dots, x_n).$$

Pertanto

$$I_n(\theta) = E^\theta\Big[\Big\{\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta; X_1, \dots, X_n)\Big\}^2\Big] = E^\theta\Big[\Big\{-\frac{n}{\theta} 1_{(0,\theta)^n}(X_1, \dots, X_n)\Big\}^2\Big]$$
$$= \frac{n^2}{\theta^2} E^\theta\big[1_{(0,\theta)^n}(X_1, \dots, X_n)\big] = \frac{n^2}{\theta^2}\int_{(0,\theta)^n} \frac{1}{\theta^n}\,\mathrm{d}x = \frac{n^2}{\theta^2} \neq n I_1(\theta).$$

Esempio 6.12 Un caso bidimensionale. Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2$, con m e σ^2 entrambi sconosciuti. Calcoleremo $I(m, \sigma^2)$. Osserviamo che in questo caso si ha $\theta = (m, \sigma^2)$, cioè il parametro è bidimensionale. $I(\theta)$ sarà dunque una matrice 2×2:

$$I(\theta) = \begin{pmatrix} I_{1,1}(\theta) & I_{1,2}(\theta) \\ I_{2,1}(\theta) & I_{2,2}(\theta) \end{pmatrix}.$$

La verosimiglianza è data da

$$L(\theta; x_1, \dots, x_n) = L\big((m, \sigma^2); x_1, \dots, x_n\big) = \frac{1}{(2\pi)^{\frac{n}{2}}} \sigma^{-n} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - m)^2\Big),$$

dunque

$$\log L\big((m,\sigma^2)\big) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2;$$

da cui (in (1) e (2) si fanno gli stessi conti che nei punti (a) e (b))

$$(1)\ \frac{\partial^2}{\partial m^2}\log L\big((m,\sigma^2)\big) = \frac{\partial^2}{\partial m^2}\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\Big\} = \frac{\partial}{\partial m}\Big\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)\Big\} = -\frac{n}{\sigma^2};$$

$$(2)\ \frac{\partial^2}{\partial(\sigma^2)^2}\log L\big((m,\sigma^2)\big) = \frac{\partial^2}{\partial(\sigma^2)^2}\Big\{-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\Big\}$$

$$= \frac{\partial}{\partial(\sigma^2)}\Big\{-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-m)^2\Big\} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-m)^2;$$

$$(3)\ \frac{\partial^2}{\partial m\partial(\sigma^2)}\log L\big((m,\sigma^2)\big) = \frac{\partial}{\partial(\sigma^2)}\Big\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-m)\Big\} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-m).$$

Dunque

$$(1)\ I_{1,1} = -E^\theta\Big[\frac{\partial^2}{\partial m^2}\log L\big((m,\sigma^2)\big)\Big] = \frac{n}{\sigma^2};$$

$$(2)\ I_{2,2} = -E^\theta\Big[\frac{\partial^2}{\partial(\sigma^2)^2}\log L\big((m,\sigma^2)\big)\Big] = -E^\theta\Big[\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i-m)^2\Big]$$

$$= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^4}\sum_{i=1}^{n}\underbrace{E^\theta\Big[\Big(\frac{X_i-m}{\sigma}\Big)^2\Big]}_{=Var\left(\frac{X_i-m}{\sigma}\right)=1} = -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} = \frac{n}{2\sigma^4};$$

$$(3)\ I_{1,2} = I_{2,1} = -E^\theta\Big[\frac{\partial^2}{\partial m\partial(\sigma^2)}\log L\big((m,\sigma^2)\big)\Big] = \frac{1}{\sigma^4}E^\theta\Big[\sum_{i=1}^{n}(X_i-m)\Big] = 0,$$

e quindi la matrice di informazione è

$$I(m,\sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Adesso vogliamo vedere se la nozione di informazione di Fisher è in accordo con l'intuizione per quanto riguarda i concetti di statistica libera e statistica esaustiva.
È necessario innanzitutto definire cosa si intende per "misura dell'informazione" contenuta in una statistica T (finora abbiamo parlato solo di informazione contenuta in un modello statistico); per far questo mettiamoci nella situazione seguente.

Sia $\big(\Omega,\mathcal{F},\{P^\theta,\theta\in\Theta\}\big)$un modello statistico assegnato, dominato da una misura μ e con verosimiglianza $L(\theta)$; su di esso consideriamo la statistica T, a valori nello spazio misurabile $(E,\mathcal{E})$. Indicata con $Q^\theta = T(P^\theta)$ la legge di T (cioè la probabilità immagine di P^θ secondo T), consideriamo il modello statistico $(E,\mathcal{E},\{Q^\theta,\theta\in\Theta\})$ (detto *modello immagine di* $\big(\Omega,\mathcal{F},\{P^\theta,\theta\in\Theta\}\big)$ *secondo* T). Si tratta ovviamente di un modello dominato (una misura dominante è $\nu = T(\mu)$, supposta al solito σ-finita; indicheremo la verosimiglianza di questo modello con $L_T(\theta)$. Supporremo che sia L che L_T verifichino le ipotesi necessarie per poter definire l'informazione di Fisher; le informazioni dei due

modelli $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$e $(E, \mathcal{E}, \{Q^\theta, \theta \in \Theta\})$ saranno indicate con $I(\theta)$ e $I_T(\theta)$ rispettivamente. $I_T(\theta)$ è appunto l'*informazione contenuta in* T.

Sappiamo che una statistica libera non dà alcuna informazione sul parametro, mentre all'opposto una statistica conserva tutta l'informazione del modello. Quindi, intuitivamente ci aspettiamo che $I_T = 0$ se T è libera e $I_T = I$ se T è esaustiva. Questo è effettivamente quello che accade. Precisamente si hanno i risultati seguenti.

Proposition 6.13 (a) Se T è libera, allora $I_T(\theta) = 0$ per ogni $\theta \in \Theta$.
(b)Viceversa, se $I_T(\theta) = 0$ per ogni $\theta \in \Theta$ e l'applicazione $\theta \mapsto \nabla L_T(\theta)$ è continua, allora T è libera.

Dimostrazione. (a) T è libera se $Q^\theta = T(P^\theta)$ non dipende da θ. Dunque non dipende da θ neppure $L_T(\theta) = \frac{\mathrm{d}Q^\theta}{\mathrm{d}\nu}$, e di conseguenza $\nabla \log L_T = 0$.

(b) Indichiamo con $X : (E, \mathcal{E}) \to (E, \mathcal{E})$ l'applicazione identica $X(x) = x$, $x \in E$, e osserviamo che $I_T(\theta)$ è la matrice di covarianza del vettore aleatorio centrato $\nabla \log L_T(\theta, X)$. Dunque, se essa è identicamente nulla, allora, per ogni θ, il vettore $\nabla \log L_T(\theta, X)$ è Q^θ-quasi certamente nullo. Dunque, per ogni θ, esiste un evento $N_\theta \in \mathcal{E}$, con $Q^\theta(N_\theta) = 0$ e tale che, per ogni $x \notin N_\theta$, si ha

$$\nabla \log L_T(\theta, x) = \frac{1}{L_T(\theta, x)} \nabla L_T(\theta, x) = 0 \Longrightarrow \nabla L_T(\theta, x) = 0.$$

Osserviamo ora che $\frac{\mathrm{d}Q^\theta}{\mathrm{d}\nu} = L_T(\theta, x) > 0$ per ogni θ e per ogni x, e dunque, per ogni θ,

$$\nu(N_\theta) = \int_{N_\theta} \mathrm{d}\nu = \int_{N_\theta} \frac{\mathrm{d}\nu}{\mathrm{d}Q^\theta}\, \mathrm{d}Q^\theta == \int_{N_\theta} \frac{1}{L(\theta)}\, \mathrm{d}Q^\theta = 0$$

perché l'integrale è fatto su un insieme trascurabile rispetto a Q^θ. Sia $D \subseteq \Theta$ un sottoinsieme numerabile e denso in Θ, e poniamo

$$N = \bigcup_{\theta \in D} N_\theta.$$

Allora $\nu(N) = 0$. Sia $x \notin N$; allora $x \notin N_\theta$ per ogni $\theta \in D$ e quindi

$$\nabla L_T(\theta, x) = 0, \qquad \text{sempre per ogni } \theta \in D.$$

Per ogni θ_0 esiste una successione (θ_n) di elementi di D, tale che $\theta_n \to \theta_0$. Passando allora al limite, per la continuità di $\theta \mapsto \nabla L_T(\theta)$ si ottiene

$$\nabla L_T(\theta_0, x) = \lim_{n \to \infty} \nabla L_T(\theta_n, x) = 0,$$

per ogni $x \notin N$. Dunque $L_T(\theta, x) = \frac{\mathrm{d}Q^\theta}{\mathrm{d}\nu}$ è ν-quasi certamente costante rispetto a θ, e di conseguenza anche $Q^\theta = T(P^\theta)$ lo è : infatti, per ogni $A \in \mathcal{E}$ abbiamo

$$Q^\theta(A) = \int_A \mathrm{d}Q^\theta = \int_A \frac{\mathrm{d}Q^\theta}{\mathrm{d}\nu}\, \mathrm{d}\nu.$$

Ciò significa che T è libera.

□

Passiamo a vedere il caso di una statistica esaustiva. Premettiamo il

Lemma 6.14 *Per ogni statistica T si ha la relazione*

$$E^\theta\big[\nabla \log L(\theta)|T\big] = \nabla \log L_T(\theta, T).$$

DIMOSTRAZIONE. Il secondo membro dell'uguaglianza nell'enunciato è chiaramente T-misurabile; quindi basta verificare che, per ogni $B \in \mathcal{F}$, del tipo $B = T^{-1}(A) = \{T \in A\}$, con $A \in \mathcal{E}$ e per ogni $i = 1, \dots, k$ si ha

$$\int_B \frac{\partial \log L(\theta, \omega)}{\partial \theta_i}\, \mathrm{d}P^\theta(\omega) = \int_B \frac{\partial \log L(\theta, T(\omega))}{\partial \theta_i}\, \mathrm{d}P^\theta(\omega).$$

Infatti

$$\int_B \frac{\partial \log L(\theta)}{\partial \theta_i}\, \mathrm{d}P^\theta = \int_B \frac{\partial L(\theta)}{\partial \theta_i} \cdot \frac{1}{L(\theta)}\, \mathrm{d}P^\theta = \int_B \frac{\partial L(\theta)}{\partial \theta_i} \cdot \frac{\mathrm{d}\mu}{\mathrm{d}P^\theta}\, \mathrm{d}P^\theta$$
$$= \int_B \frac{\partial L(\theta)}{\partial \theta_i}\, \mathrm{d}\mu = \frac{\partial}{\partial \theta_i} \int_B L(\theta)\, \mathrm{d}\mu = \frac{\partial}{\partial \theta_i} \int_B \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}\, \mathrm{d}\mu = \int_B \mathrm{d}P^\theta = \frac{\partial}{\partial \theta_i} P^\theta(B) = \frac{\partial}{\partial \theta_i} Q^\theta(A).$$

Questo nel modello di partenza. Nel modello immagine avremo le stesse relazioni (fino alla penultima uguaglianza) sostituendo Q^θ, ν, A e $L_T(\theta)$ al posto di P^θ, μ, B e $L(\theta)$ rispettivamente. Troviamo

$$\int_A \frac{\partial \log L_T(\theta)}{\partial \theta_i}\, \mathrm{d}Q^\theta = \frac{\partial}{\partial \theta_i} Q^\theta(A),$$

e quindi concludiamo che

$$\int_B \frac{\partial \log L(\theta, \omega)}{\partial \theta_i}\, \mathrm{d}P^\theta(\omega) = \frac{\partial}{\partial \theta_i} Q^\theta(A) = \int_A \frac{\partial \log L_T(\theta, x)}{\partial \theta_i}\, \mathrm{d}Q^\theta(x) = \int_B \frac{\partial \log L(\theta, T(\omega))}{\partial \theta_i}\, \mathrm{d}P^\theta(\omega).$$

□

Proposition 6.15 *Per ogni statistica T si ha la relazione*

$$I_T(\theta) \le I(\theta), \qquad \forall\, \theta \in \Theta,$$

nel senso che la matrice $I(\theta) - I_T(\theta)$ è semidefinita positiva.

DIMOSTRAZIONE. Per il lemma precedente, per ogni $u \in \mathbb{R}^k$ si ha

$$\langle \nabla \log L_T(\theta, T), u\rangle^2 = \Big\{E^\theta\big[\langle \nabla \log L(\theta), u\rangle|T\big]\Big\}^2 \le E^\theta\Big[\langle \nabla \log L(\theta), u\rangle^2|T\Big], \tag{5}$$

(dove nell'ultima relazione si è applicata la diseguaglianza di Jensen per le speranze condizionali: per ogni ϕ funzione convessa si ha

$$\phi\big(E[X|\mathcal{B}]\big) \le E[\phi(X)|\mathcal{B}].$$

Nel nostro caso si prende $\phi(x) = x^2$.)

Passando allora alla speranza secondo P^θ in entrambi i membri della diseguaglianza (5), troviamo

$$\langle I_T(\theta)u, u\rangle = \langle(\mathbf{Cov}^\theta \nabla \log L_T(\theta, T)u, u\rangle \le \langle(\mathbf{Cov}^\theta \nabla \log L(\theta, X)u, u\rangle = \langle I(\theta)u, u\rangle,$$

indicando con X è l'applicazione identica di Ω in Ω e ricordando che, dai conti fatti nell'osservazione 6.4 segue che, se U è un vettore centrato, si ha

$$E\big[\langle X, u\rangle^2\big] = \langle(\mathbf{Cov}U)u, u\rangle.$$

La Proposizione è dimostrata.

□

Teorema 6.16 *Se T è esaustiva, allora $I(\theta) = I_T(\theta)$ per ogni $\theta \in \Theta$.*

DIMOSTRAZIONE. Grazie al Teorema di fattorizzazione 2.6, i può supporre che μ sia una dominante privilegiata. In tal caso lo stesso Teorema ci dice che la verosimiglianza si scrive nella forma

$$L_T(\theta, \omega) = g^\theta(T(\omega)).$$

Questo implica che la verosimiglianza nel modello immagine (rispetto alla misura dominante $\nu = T(\mu)$) ha la forma

$$L(\theta, t) = g^\theta(t).$$

Infatti, per ogni $A \in \mathcal{E}$ si ha

$$Q^\theta(A) = P^\theta(T \in A) = \int_\Omega 1_A(T)\,\mathrm{d}P^\theta = \int_\Omega 1_A(T)L(\theta)\,\mathrm{d}\mu = \int_\Omega 1_A(T)g^\theta(T)\,\mathrm{d}\mu$$
$$= \int_E 1_A(t)g^\theta(t)\,\mathrm{d}\nu(t) = \int_A g^\theta(t)\,\mathrm{d}\nu,$$

per il teorema di integrazione rispetto alla legge immagine.

Dunque, per $i, j = 1, \ldots, k$ si ha

$$I(\theta)_{i,j} = E^\theta\Big[\frac{\partial}{\partial\theta_i}\log L(\theta)\cdot\frac{\partial}{\partial\theta_j}\log L(\theta)\Big] = \int \frac{\partial}{\partial\theta_i}\log g^\theta(T)\cdot\frac{\partial}{\partial\theta_j}\log g^\theta(T)\ \underbrace{g^\theta(T)\,\mathrm{d}\mu}_{L(\theta)\,\mathrm{d}\mu=\mathrm{d}P^\theta}$$
$$= \int \frac{\partial}{\partial\theta_i}\log g^\theta(t)\cdot\frac{\partial}{\partial\theta_j}\log g^\theta(t)\ \underbrace{g^\theta(t)\,\mathrm{d}\nu(t)}_{L_T(\theta)\,\mathrm{d}\nu(t)=\mathrm{d}Q^\theta(t)} = I_T(\theta)_{i,j}$$

□

Il viceversa di questo risultato ha bisogno di qualche ipotesi supplementare. Precisamente

Teorema 6.17 *Se le applicazioni $\theta \mapsto L(\theta)$ e $\theta \mapsto L_T(\theta)$ sono continue e $I(\theta) = I_T(\theta)$ per ogni $\theta \in \Theta$, allora T è esaustiva.*

DIMOSTRAZIONE. Dire che $I(\theta) = I_T(\theta)$ significa dire che, per ogni $u \in \mathbb{R}^k$ e per ogni $\theta \in \Theta$, è nulla la speranza rispetto a P^θ di

$$E^\theta\Big[\langle\nabla\log L(\theta), u\rangle^2|T\Big] - \Big\{E^\theta\big[\langle\nabla\log L(\theta), u\rangle|T\big]\Big\}^2,$$

per il Lemma 6.14. Dato che questa quantità è sempre non negativa, (per la prima relazione che compare nella dimostrazione della Proposizione 6.15), si deduce che

$$E^\theta\Big[\langle\nabla\log L(\theta), u\rangle^2|T\Big] = \Big\{E^\theta\big[\langle\nabla\log L(\theta), u\rangle|T\big]\Big\}^2,$$

P^θ-quasi certamente, per ogni u e per ogni θ. Vale il risultato seguente (per la dimostrazione si veda [2]).

Proposition 6.18 *(*Reciproca della diseguaglianza di Jensen*). Sullo spazio $(\Omega, \mathcal{F}, P)$ sia Y un vettore aleatorio integrabile a valori in $\mathbb{R}^d$ e $\mathcal{B}$ una sotto-σ-algebra di $\mathcal{F}$. Sia $\phi : \mathbb{R}^d \to \mathbb{R}$ una funzione strettamente convessa. Se accade che*

$$E[\phi(Y)|\mathcal{B}] = \phi\big(E[Y|\mathcal{B}]\big),$$

allora $1_B Y$ è $\mathcal{B}$-misurabile per ogni $B \in \mathcal{B}$.
In particolare, se $\mathcal{B} = \{\emptyset, \Omega\}$ è la σ-algebra banale, allora $X = 1_\Omega X$ è costante.

Applicando questa proposizione con $\phi(x) = x^2$, $Y = \langle \nabla \log L(\theta), u\rangle$ e $\mathcal{B} = \sigma(T)$ si trova che, per ogni $u \in \mathbb{R}^k$ e per ogni $\theta \in \Theta$, la v.a.

$$\omega \mapsto 1_B \langle \nabla \log L(\theta, omega), u\rangle$$

è $\sigma(T)$-misurabile per ogni $B \in \sigma(T)$. Dunque, indicando con X la v.a identica su Ω, si ha

$$\begin{aligned} &1_B \langle \nabla \log L(\theta, X), u\rangle = E^\theta\big[1_B \langle \nabla \log L(\theta, X), u\rangle | T\big] = 1_B E^\theta\big[\langle \nabla \log L(\theta, X), u\rangle | T\big] \\ &= 1_B \langle \nabla \log L_T(\theta, T), u\rangle, \end{aligned}$$

dove nell'ultima uguaglianza si è applicato il Lemma 6.14. L'uguaglianza ottenuta, applicata con $B = \Omega$, dice allora che

$$\langle \nabla \log L(\theta, X), u\rangle = \langle \nabla \log L_T(\theta, T), u\rangle,$$

P^θ-quasi certamente. In altre parole, per ogni θ e per ogni u, esiste un evento $N_{\theta,u} \in \mathcal{F}$ con $P^\theta(N_{\theta,u}) = 0$ tale che, per ogni $\omega \notin N_{\theta,u}$, si ha

$$\langle \nabla \log L(\theta, \omega), u\rangle = \langle \nabla \log L_T(\theta, T(\omega)), u\rangle. \tag{6}$$

Procedendo come nella dimostrazione della Proposizione 6.13, si arriva a trovare un evento $N_u \in \mathcal{F}$, con $\mu(N_u) = 0$ tale che l'uguaglianza (6) vale per ogni $\omega \notin N_u$ e per ogni θ.
A questo punto, fissata una base $u_1, \ldots, u_k$ di $\mathbb{R}^k$ e posto $N = \cup_{i=1}^n N_{u_i}$, si ha che $\mu(N) = 0$ e, per ogni $\omega \notin N$, per linearità la relazione (6) vale per ogni $u \in \mathbb{R}^k$, e di conseguenza si ha l'uguaglianza

$$\nabla \log L(\theta, \omega) = \nabla \log L_T(\theta, T(\omega)).$$

Integrando, si trova allora che, per ogni $\omega \notin N$ e per ogni θ, risulta

$$\log L(\theta, \omega) = \log L_T(\theta, T(\omega)) + g(\omega),$$

(per un'opportuna g) e cioè

$$L(\theta, \omega) = L_T(\theta, T(\omega)) \mathrm{e}^{g(\omega)}.$$

La conclusione segue allora dal Teorema di fattorizzazione 2.6.

□

Esempio 6.19 Sia $(X_1, \ldots, X_n)$ un campione di legge $\mathcal{N}(m, 1)$ e sia $T = \overline{X}$ (la media campionaria). Il modello statistico di partenza è

$$\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{P^m, m \in \mathbb{R}\}\big),$$

dove P^m è il prodotto tensoriale di n leggi $\mathcal{N}(m, 1)$. Dall'esempio 6.10 (a) segue che $I(m) = \frac{n}{\sigma^2} = n$; inoltre, come è ben noto, la legge di $\overline{X}$ è la $\mathcal{N}(m, \frac{1}{n})$ e quindi il modello di arrivo è

$$\big(\mathbb{R}, \mathcal{B}(\mathbb{R}^n), Q^m\big),$$

dove $Q^m = \mathcal{N}(m, \frac{1}{n})$; di conseguenza, amcora per l'esempio 6.10 (a), si ha

$$I_T(m) = \frac{1}{\sigma^2} = \frac{1}{\frac{1}{n}} = n = I(m),$$

in accordo con il fatto che $T = \overline{X}$ è esaustiva (a varianza costante).

Esempio 6.20 Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$ e sia $T = S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$ (la varianza campionaria). In questo caso abbiamo $\theta = (m, \sigma^2)$. Il modello di partenza è

$$\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{P^\theta, \theta \in \mathbb{R} \times \mathbb{R}^+\}\big),$$

dove P^θ è il prodotto tensoriale di n leggi $\mathcal{N}(m, \sigma^2)$. Per stabilire il modello di arrivo, è necessario ricavare la legge di T. Vedremo nel Teorema... che la statistica

$$\frac{T(n-1)}{\sigma^2} = \sum_{i=1}^n \Big(\frac{X_i - \overline{X}}{\sigma}\Big)^2$$

ha legge $\chi^2(n-1) = \Gamma\big(\frac{n-1}{2}, \frac{1}{2}\big)$, la cui densità è

$$f(x) = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} x^{\frac{n-3}{2}} \mathrm{e}^{-\frac{x}{2}} 1_{\mathbb{R}^+}(x).$$

Dunque

$$P(T \le t) = P\Big(\frac{T(n-1)}{\sigma^2} \le \frac{t(n-1)}{\sigma^2}\Big) = \int_{-\infty}^{\frac{t(n-1)}{\sigma^2}} f(x)\,\mathrm{d}x,$$

e quindi, derivando, si trova che la densità di T è data da

$$f_T(t) = f\Big(\frac{t(n-1)}{\sigma^2}\Big) \cdot \frac{n-1}{\sigma^2} = \frac{n-1}{\sigma^2} \cdot \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} \Big(\frac{t(n-1)}{\sigma^2}\Big)^{\frac{n-3}{2}} \mathrm{e}^{-\frac{t(n-1)}{2\sigma^2}} 1_{\mathbb{R}^+}(t)$$

$$= \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} \Big(\frac{n-1}{\sigma^2}\Big)^{\frac{n-1}{2}} t^{\frac{n-3}{2}} \mathrm{e}^{-\frac{t(n-1)}{2\sigma^2}} 1_{\mathbb{R}^+}(t).$$

Indicheremo la legge di T con il simbolo $Q^{\sigma^2} (= \sigma^2 \cdot \chi^2(n-1))$. Il modello di arrivo è dunque

$$\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+, \{Q^{\sigma^2}, \sigma^2 \in \mathbb{R}^+\}).$$

Abbiamo visto in 6(b) che $I(\sigma^2) = \frac{n}{2\sigma^4}$; calcoliamo ora $I_T(\sigma^2)$. Si ha, su $\mathbb{R}^+$,

$$L_T(\sigma^2, t) = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} \Big(\frac{n-1}{\sigma^2}\Big)^{\frac{n-1}{2}} t^{\frac{n-3}{2}} \mathrm{e}^{-\frac{t(n-1)}{2\sigma^2}}.$$

Dunque

$$\log L_T(\sigma^2, t) = c + \frac{n-1}{2} \log \frac{n-1}{\sigma^2} + \frac{n-3}{2} \log t - \frac{(n-1)t}{2\sigma^2.}$$

Derivando

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)} \log L_T(\sigma^2, t) = -\frac{n-1}{2} \cdot \frac{\sigma^2}{n-1} \cdot \frac{n-1}{\sigma^4} + \frac{(n-1)t}{\sigma^4};$$

derivando ancora

$$\frac{\mathrm{d}^2}{\mathrm{d}(\sigma^2)^2} \log L_T(\sigma^2, t) = \frac{n-1}{2\sigma^4} - \frac{(n-1)t}{\sigma^6}.$$

Quindi

$$I_t(\sigma^2) = -E^{\sigma^2}\Big[\frac{n-1}{2\sigma^4} - \frac{(n-1)T}{\sigma^6}\Big] = -\frac{n-1}{2\sigma^4} + E^{\sigma^2}\Big[\frac{(n-1)T}{\sigma^6}\Big] = -\frac{n-1}{2\sigma^4} + \frac{1}{\sigma^4} E^{\sigma^2}\Big[\frac{(n-1)T}{\sigma^2}\Big].$$

Come abbiamo detto, $\frac{T(n-1)}{\sigma^2}$ ha legge $\chi^2(n-1)$, e quindi media $n-1$. Per concludere, si trova

$$I_T(\sigma^2) = \frac{n-1}{2\sigma^4} < \frac{n}{2\sigma^4} = I(m, \sigma^2);$$

dunque T non è esaustiva per σ^2.

Passiamo a dimostrare una importante diseguaglianza che, in sostanza, dice che se pretendiamo di avere uno stimatore corretto, non si può contemporaneamente pretendere di avere un rischio basso quanto vogliamo.

Ci serve qualche richiamo di algebra lineare. Sia A una matrice $n \times n$ simmetrica; si dice che A è *semidefinita positiva* ($A \geq 0$) se, per ogni $x \in \mathbb{R}^n$, $\langle Ax, x\rangle \geq 0$; si dice che A è *definita positiva* ($A > 0$) se, per ogni $x \in \mathbb{R}^n$, con $x \neq 0$, $\langle Ax, x\rangle > 0$. È noto che A è definita positiva se e solo se A è semidefinita positiva e invertibile. Nel seguito ci serviranno due risultati di algebra lineare.

Lemma 6.21 *Sia A una matrice $n \times n$ simmetrica semidefinita positiva; allora esiste una matrice B $n \times n$ simmetrica tale che $A = B\,{}^tB = B^2$. Inoltre, se A è invertibile, anche B è invertibile. B viene talvolta indicata con il simbolo $\sqrt{A}$.*

Dimostrazione. Dato che A è semidefinita positiva, i suoi autovalori $\lambda_1, \ldots, \lambda_n$ sono tutti non negativi. È noto che, posto

$$K = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}$$

esiste una matrice ortogonale O tale che $A = OK\,{}^tO$. Poniamo

$$H = \begin{pmatrix} \sqrt{\lambda}_1 & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\lambda}_n \end{pmatrix}$$

e $B = OH\,{}^tO$. B risponde alla questione. Infatti ${}^tB = {}^t(OH\,{}^tO) = OH\,{}^tO = B$ ed inoltre

$$B^2 = (OH\,{}^tO)(OH\,{}^tO) = OH({}^tOO)H\,{}^tO = OH^2\,{}^tO = OK\,{}^tO = A.$$

La seconda affermazione è ovvia, perchè se A è invertibile, allora è definita positiva e quindi tutti i suoi autovalori sono strettamente positivi, e dunque anche H è invertibile.

□

Lemma 6.22 *Siano $a \in \mathbb{R}^n$ e A una matrice $n \times n$ simmetrica e definita positiva (dunque invertibile). Allora*

$$\sup_{x \in \mathbb{R}^n} \frac{\langle x, a\rangle^2}{\langle Ax, x\rangle} = \max_{x \in \mathbb{R}^n} \frac{\langle x, a\rangle^2}{\langle Ax, x\rangle} = \langle A^{-1}a, a\rangle.$$

Dimostrazione. Il risultato è ovvio se $a = \mathbf{0}$. Consideriamo dunque il caso in cui $a \neq \mathbf{0}$. Dal Lemma 6.21 sappiamo che esiste una matrice B $n \times n$ simmetrica e invertibile tale che $A = B^2$; dunque, per ogni $x \in \mathbb{R}^n$, esiste uno e un solo $y \in \mathbb{R}^n$ tale che $x = B^{-1}y$; in particolare poniamo $b = B^{-1}a$. Allora, poiché B^{-1} è simmetrica e $AB^{-1} = B$, si ha

$$\frac{\langle x, a\rangle^2}{\langle Ax, x\rangle} = \frac{\langle B^{-1}y, a\rangle^2}{\langle AB^{-1}y, B^{-1}y\rangle} = \frac{\langle y, B^{-1}a\rangle^2}{\langle By, B^{-1}y\rangle} = \frac{\langle y, B^{-1}a\rangle^2}{\langle y, BB^{-1}y\rangle} = \frac{\langle y, b\rangle^2}{\|y\|^2} = \Big(\frac{\langle y, b\rangle}{\|y\|}\Big)^2.$$

È ben noto che la funzione $y \mapsto \frac{\langle y,b\rangle}{\|y\|}$ assume il suo massimo per $y = \frac{b}{\|b\|}$ e che tale massimo vale $\|b\|$. Pertanto il massimo del suo quadrato vale

$$\|b\|^2 = \langle B^{-1}a, B^{-1}a\rangle = \langle (B^{-1})^2 a, a\rangle = \langle (B^2)^{-1}a, a\rangle = \langle A^{-1}a, a\rangle.$$

□

Torniamo alla nostra situazione. Vale il

Teorema 6.23 *In aggiunta alle ipotesi fatte finora, supponiamo che $I(\theta)$ sia invertibile per ogni θ. Allora, per ogni θ e per ogni v.a. Y (di quadrato integarbile secondo P^θ) si ha*

$$Var^\theta(Y) \geq \langle\, I(\theta)^{-1}(\nabla E^\theta[Y]), \nabla E^\theta[Y]\,\rangle.$$

Dimostrazione. Non è restrittivo supporre che la funzione $\theta \mapsto E^\theta[Y]$ non sia costante (in tal caso la tesi è ovvia). Si ha (passando la derivata dentro la speranza)

$$\begin{aligned}
&\nabla E^\theta[Y] = \nabla E[L(\theta)Y] = E\big[\big(\nabla L(\theta)\big)Y\big] = \int \big(\nabla L(\theta)\big)Y \,\mathrm{d}\mu \\
&= \int \big(\nabla L(\theta)\big)Y \frac{\mathrm{d}\mu}{\mathrm{d}P^\theta}\,\mathrm{d}P^\theta = \int \big(\nabla L(\theta)\big)Y \frac{1}{L(\theta}\,\mathrm{d}P^\theta = E^\theta[\big(\nabla \log L(\theta)\big)Y] \\
&= E^\theta\big[\big(\nabla \log L(\theta)\big)\cdot\big(Y - E^\theta[Y]\big)\big],
\end{aligned}$$

ricordando che il vettore $\nabla \log L(\theta)$ è centrato rispetto a ogni P^θ e quindi

$$E^\theta\big[\big(\nabla \log L(\theta)\big)\cdot E^\theta[Y]\big] = 0.$$

Moltiplicando scalarmente per $x \in \mathbb{R}^k$ il primo e ultimo termine della relazione precedente, si ottiene allora

$$\langle x, \nabla E^\theta[Y]\rangle = E^\theta\big[\langle x, \big(\nabla \log L(\theta)\big)\cdot\big(Y - E^\theta[Y]\big)\rangle\big]$$

e, usando nella funzione integranda la diseguaglianza di Schwartz si trova

$$\langle x, \nabla E^\theta[Y]\rangle^2 \leq \big(Var^\theta Y\big)E^\theta\big[[\langle x, \big(\nabla \log L(\theta)\big)\rangle^2\big] = \big(Var^\theta Y\big)\langle I(\theta)x, x\rangle,$$

dove l'ultima uguaglianza segue dai conti fatti nell'Osservazione 6.4 (ved. anche la dimostrazione della Proposizione 6.15).

Dato che $I(\theta)$ è definita positiva (in quanto invertibile), la relazione precedente si può scrivere anche nella forma

$$Var^\theta Y \geq \frac{\langle x, \nabla E^\theta[Y]\rangle^2}{\langle I(\theta)x, x\rangle}$$

e questa diseguaglianza può essere ottimizzata passando al $\sup_{x\in\mathbb{R}^n}$. Utilizzando allora il Lemma 6.22 con $A = I(\theta)$ e $a = \nabla E^\theta[Y]$ si ottiene la tesi.

□

L'importanza del Teorema precedente è riposta nel

Corollario 6.24 *(*Diseguaglianza di Cramer-Rao*). Nelle ipotesi del Teorema 6.23, sia Y uno stimatore corretto e di quadrato integrabile della funzione $g(\theta)$. Allora*

$$R_Y(\theta) = Var^\theta(Y) \geq \langle\, I(\theta)^{-1}(\nabla g(\theta)), \nabla g(\theta)\,\rangle.$$

Osservazione 6.25 Nel caso particolare in cui Θ sia un intervallo della retta, la diseguaglianza di Cramer-Rao diventa

$$R_Y(\theta) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

Si vede dunque che il rischio di Y è minorato da un numero positivo, tanto più piccolo quanto più nell'intorno di θ l'informazione di Fisher è grande. Questa osservazione spiega (in parte) quanto detto all'inizio di questo paragrafo. Le cose saranno più chiare quando avremo introdotto l'informazione di Kullback (§ 7).

La diseguaglianza di Cramer-Rao dice che uno stimatore corretto ha sempre un rischio non inferiore ad una certa quantità (il termine a destra nella diseguaglianza). È quindi naturale chiedersi se il confine inferiore posto da tale diseguaglianza (detto *confine*, o *bound di Cramer-Rao*) sia raggiungibile.

Definizione 6.26 Si chiama *efficace* (*efficient* in inglese) uno stimatore Y di quadrato integrabile della funzione $g(\theta)$ il cui rischio uguagli il confine inferiore posto dalla diseguaglianza di Cramer-Rao.

Esempio 6.27 Consideriamo un modello esponenziale con $\Theta \subseteq \mathbb{R}$ e verosimiglianza della forma $L(\theta) = \exp\big(\theta T - \psi(\theta)\big)$. Ricordando il Teorema 5.11 si ha

$$I(\theta) = E^\theta\Big[\Big(\frac{\mathrm{d}}{\mathrm{d}\theta}\log L(\theta)\Big)^2\Big] = E^\theta\big[\big(T-\psi'(\theta)\big)^2\big] = E^\theta\big[\big(T-E^\theta[T]\big)^2\big] = Var^\theta(T) = \psi''(\theta).$$

T è uno stimatore corretto efficace di $g(\theta) = \psi'(\theta)$; infatti il confine di Cramer-Rao è

$$\frac{(g'(\theta))^2}{I(\theta)} = \frac{\big(\psi''(\theta)\big)^2}{\psi''(\theta)} = \psi''(\theta) = Var^\theta(T) = R_T(\theta).$$

Considerando più in generale un modello esponenziale con $L(\theta) = \exp\big(\langle\theta, T\rangle - \psi(\theta)\big)$ con Θ aperto di $\mathbb{R}^k$, con passaggi analoghi (che richiedono qualche calcolo di algebra lineare) si prova che
(a) la matrice di informazione $\big(I(\theta)_{i,j}\big)$ è data da

$$I(\theta)_{i,j} = \frac{\partial^2\psi(\theta)}{\partial\theta_i\partial\theta_j}, \qquad i,j = 1,\dots,k;$$

(b) Per ogni $i = 1,\dots,k$, T_i è uno stimatore efficace di $\frac{\partial\psi(\theta)}{\partial\theta_i}$. Parlando in termini vettoriali, si dice che T è uno stimatore efficace di $\nabla\psi(\theta)$.

Esercizio 6.28 Sia $\big(\mu^\theta\big)$ una famiglia esponenziale su $\mathbb{R}$ e si consideri un campione di taglia n e legge μ^θ. Provare che il confine di Cramer-Rao decresce come $\frac{1}{n}$.

Esercizio 6.29 Si consideri un campione di taglia n e legge $\mathcal{E}(\theta)$ $(0 < \theta < \infty)$. Provare che non esiste uno stimatore efficace di θ.
Suggerimento: ricordando che $\sum_{i=1}^n X_i$ è una statistica esaustiva completa e che, sotto P^θ, ha legge $\Gamma(n,\theta)$, verificare (ed. Esempio (8.4)) che $T = \frac{n-1}{\sum_{i=1}^n X_i}$ è uno stimatore corretto di θ (e quindi ottimale nella classe degli stimatori di θ corretti e di quadrato integrabile), che però non è efficace.

7 L'informazione di Kullback

Il significato ell'informazione di Fisher si comprende meglio se si introduce un altro concetto di informazione, dovuto a Kullback.

Sia $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$un modello statistico dominato dalla misura μ. Supponiamo che il vero valore di θ sia θ_1; ci domandiamo in quale misura il modello (o meglio, il risultato ω dell'esperimento di cui $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$è il modello) ci permetta di distinguere θ_1 da un altro valore θ_2. Consideriamo i due casi estremi ($\Theta \subseteq \mathbb{R}$ per semplicità):

(i) $L(\theta_1, \omega) = L(\theta_2, \omega)$ per ogni $\omega \in \Omega$; è evidente che in questo caso il risultato ω dell'esperimento non permette di fare alcuna distinzione fra θ_1 e θ_2;

(ii) $L(\theta_1, \omega) > 0$ per $\omega \in (a_1, b_1)$, $L(\theta_2, \omega) > 0$ per $\omega \in (a_2, b_2)$, dove (a_1, b_1) e (a_2, b_2) sono due intervalli in $\mathbb{R}$ con $(a_1, b_1) \cap (a_2, b_2) = \emptyset$. Dato che ω cade quasi certamente in (a_1, b_1), possiamo identificare θ con certezza (essendo (a_1, b_1) disgiunto da (a_2, b_2).

Ma naturalmente in generale ci troviamo in una situazione intermedia fra (i) e (ii): in qualche caso l'identificazione non è proprio possibile, come mostra la figura (1).

Invece in una situazione come quella della figura (2) sottostante, se il risultato dell'esperimento cade tra a_1 e a_2, possiamo di nuovo identificare θ_1.

Pertanto cercheremo di stabilire "in che misura" il risultato ω dell'esperimento permette di distinguere θ_1 da θ_2.

Definizione 7.1 Il *potere discriminante* tra θ_1 e θ_2 da assegnare al risultato ω è la quantità

$$i(\theta_1/\theta_2)(\omega) = \log \frac{L(\theta_1, \omega)}{L(\theta_2, \omega)},$$

definita su $\{L(\theta_1, \omega) > 0\} \cup \{L(\theta_2, \omega) > 0\}$ e con la convenzione $\log 0 = -\infty$, $\log(\frac{c}{0}) = +\infty$.

Osservazione 7.2 Se siamo nella situazione di pefetta discriminazione (caso (ii)) allora la quantità precedente vale $+\infty$; se invece non è possibile distinguere (caso (i)), allora essa vale 0. Inoltre, per i risultati ω per i quali si ha $L(\theta_2, \omega) > L(\theta_1, \omega)$, $i(\theta_1/\theta_2)$ è negativo: questo è naturale se si interpreta $L(\theta_1, \omega)$ come la probabilità di ottenere il risultato ω se il valore del parametro è θ_1 e $L(\theta_2, \omega)$ analogamente: se la probabilità di ottenere ω con il parametro θ_1 è più bassa della probabilità di ottenere ω con il parametro θ_2, saremo inclini a decidere in favore di θ_2.

Il potere discriminante dipende ovviamente da ω; quindi, se vogliamo poterlo usare come una misura della possibile "lontananza" tra θ_1 e θ_2, occorre effettuare una media. Si dà dunque la

Definizione 7.3 Si chiama *informazione di Kullback* di θ_1 contro θ_2 la quantità

$$I(\theta_1/\theta_2) := E^{\theta_1}[i(\theta_1/\theta_2)] = E^{\theta_1}\Big[\log \frac{L(\theta_1)}{L(\theta_2)}\Big] = E\Big[L(\theta_1) \log \frac{L(\theta_1)}{L(\theta_2)}\Big].$$

Osservazione 7.4 In Teoria dell'Informazione l'informazione di Kullback è nota con il nome di *entropia relativa* (di P^{θ_1} rispetto a P^{θ_2}).

È ovviamente necessario verificare che la speranza che compare nella definizione ha senso.

Teorema 7.5 *$I(\theta_1/\theta_2)$ ha senso per ogni coppia (θ_1, θ_2).*

Dimostrazione. Intanto, la v. a. $\omega \mapsto \log \frac{L(\theta_1,\omega)}{L(\theta_2,\omega)}$ è definita P^{θ_1}-quasi ovunque, poiché

$$P^{\theta_1}\big(\{L(\theta_1,\omega)>0\}\cup\{L(\theta_2,\omega)>0\}\big) \geq P^{\theta_1}\big(L(\theta_1)>0\big) = \int_{\{L(\theta_1)>0\}} \mathrm{d}P^{\theta_1} = \int_{\{L(\theta_1)>0\}} \underbrace{\frac{\mathrm{d}P^{\theta_1}}{\mathrm{d}\mu}}_{=L(\theta_1)}\,\mathrm{d}\mu$$

$$= \int L(\theta_1)\,\mathrm{d}\mu = \int \mathrm{d}P^{\theta_1} = 1.$$

Abbiamo bisogno di un

Lemma 7.6 *Sullo spazio $(\Omega, \mathcal{F}, P)$ sia Y una v.a. avente media finita (cioè tale che $E\big[|Y|\big] = E\big[Y^+\big] + E\big[Y^-\big] < +\infty$, e sia ϕ una funzione convessa. Allora $E\big[\phi(Y)^-\big] < +\infty$.*

Dimostrazione. Sia a un numero fissato. Poiché ϕ è convessa, per ogni y si ha $\phi(y) - \phi(a) \geq \kappa(a)(y-a)$, dove $\kappa(a)$ è un'opportuna costante. Dunque

$$\phi(Y) \geq \big(\phi(a) - a\kappa(a)\big) + \big(\kappa(a)Y\big). \tag{7}$$

È facile vedere che, se u e v sono due numeri reali, allora

$$(u+v)^- \leq u^- + v^-$$

e

$$(uv)^- \leq u^-v^+ + u^+v^-$$

(dimostrazione per esercizio). Utilizzando queste due relazioni nella (7), si trova che

$$\begin{aligned}\phi(Y)^- &\leq \big[\big(\phi(a) - a\kappa(a)\big) + \big(\kappa(a)Y\big)\big]^- \leq \big(\phi(a) - a\kappa(a)\big)^- + \big(\kappa(a)Y\big)^- \\ &\leq \big(\phi(a) - a\kappa(a)\big)^- + \kappa(a)^-Y^+ + \kappa(a)^+Y^-,\end{aligned}$$

e si conclude passando alla speranza.

□

Torniamo alla dimostrazione del Teorema. Applichiamo il Lemma precedente alla funzione convessa $x \mapsto -\log x$ e alla v. a. non negativa

$$\omega \mapsto Y(\omega) = \frac{L(\theta_2,\omega)}{L(\theta_1,\omega)},$$

dopo aver osservato che

$$E^{\theta_1}[Y] = \int \frac{L(\theta_2)}{L(\theta_1)}\,\mathrm{d}P^{\theta_1} = \int \frac{L(\theta_2)}{L(\theta_1)} L(\theta_1)\,\mathrm{d}\mu = \int L(\theta_2)\,\mathrm{d}\mu = \int \mathrm{d}P^{\theta_2} = 1.$$

Troviamo che

$$E^{\theta_1}\Big[\Big(\log\frac{L(\theta_1)}{L(\theta_2)}\Big)^-\Big] = E^{\theta_1}\Big[\Big(-\log\frac{L(\theta_2)}{L(\theta_1)}\Big)^-\Big] < +\infty,$$

da cui segue che $E^{\theta_1}\big[\log\frac{L(\theta_1)}{L(\theta_2)}\big]$ ha senso (eventualmente uguale a $+\infty$).

□

Esempio 7.7 (a) Sia $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P^\theta)$ il modello statistico con $P^\theta = \Pi_\theta$ (con la misura che conta i punti come misura dominante). Vogliamo calcolare $I(\theta_1/\theta_2)$.
Per $k \in \mathbb{N}$ si ha

$$\log \frac{L(\theta_1, k)}{L(\theta_2, k)} = \log \Big\{ \Big(\frac{\theta_1}{\theta_2}\Big)^k \mathrm{e}^{-(\theta_1 - \theta_2)} \Big\} = k \log \Big(\frac{\theta_1}{\theta_2}\Big) - \theta_1 + \theta_2,$$

e, integrando rispetto a P^{θ_1}, si trova

$$I(\theta_1/\theta_2) = \log \Big(\frac{\theta_1}{\theta_2}\Big) E^{\theta_1}[X] - \theta_1 + \theta_2 = \theta_1 \log \Big(\frac{\theta_1}{\theta_2}\Big) - \theta_1 + \theta_2 = \theta_1(\log \theta_1 - 1) - \theta_2(\log \theta_2 - 1).$$

(b) Più in generale, supponiamo di avere un modello esponenziale $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$con verosimiglianza

$$L(\theta) = C(\theta) \exp\big(\langle \theta, T \rangle\big).$$

Allora

$$\log \frac{L(\theta_1)}{L(\theta_2)} = \log \Big\{ \frac{C(\theta_1)}{C(\theta_2)} \exp\big(\langle \theta_1 - \theta_2, T \rangle\big) \Big\} = \log C(\theta_1) - \log C(\theta_2) + \langle \theta_1 - \theta_2, T \rangle$$

e, integrando rispetto a P^{θ_1}, si trova

$$\begin{aligned} I(\theta_1/\theta_2) &= \log C(\theta_1) - \log C(\theta_2) + \int \langle \theta_1 - \theta_2, T \rangle C(\theta_1) \exp\big(\langle \theta_1, T \rangle\big)\, \mathrm{d}\mu \\ &= \log C(\theta_1) - \log C(\theta_2) + \langle \theta_1 - \theta_2, \int T \big\{ C(\theta_1) \exp\big(\langle \theta_1, T \rangle\big) \big\}\, \mathrm{d}\mu \rangle \\ &= \log C(\theta_1) - \log C(\theta_2) + \langle \theta_1 - \theta_2, \int T\, \mathrm{d}P^{\theta_1} \rangle = \log C(\theta_1) - \log C(\theta_2) + \langle \theta_1 - \theta_2, E^{\theta_1}[T] \rangle. \end{aligned}$$

Teorema 7.8 *Per ogni coppia (θ_1, θ_2) si ha $I(\theta_1/\theta_2) \geq 0$; inoltre $I(\theta_1/\theta_2) = 0$ se e solo se $P^{\theta_1} = P^{\theta_2}$.*

Dimostrazione. Per la diseguaglianza di Jensen applicata alla funzione $x \mapsto -\log x$ si ha

$$I(\theta_1/\theta_2) = E^{\theta_1}\Big[-\log \frac{L(\theta_2)}{L(\theta_1)} \Big] \geq -\log E^{\theta_1}\Big[\frac{L(\theta_2)}{L(\theta_1)} \Big] = -\log 1 = 0.$$

Inoltre, per l'inversa della diseguaglianza di Jensen (Proposizione 6.18), nella relazione precedente vale l'uguaglianza se e solo se la funzione integranda è P^{θ_1}-quasi certamente costante, cioè se esiste un evento $N \in \mathcal{F}$ con $P^{\theta_1}(N) = 0$ tale che, per ogni $\omega \notin N$, si ha

$$\frac{L(\theta_2, \omega)}{L(\theta_1, \omega)} = c,$$

con $c \in \mathbb{R}^+$. Per individuare c, osserviamo che

$$c = E^{\theta_1}\Big[\frac{L(\theta_2)}{L(\theta_1)} \Big] = 1;$$

si ottiene quindi $L(\theta_2, \omega) = L(\theta_1, \omega)$ per $\omega \in N^c$, da cui

$$P^{\theta_2}(N^c) = \int_{N^c} L(\theta_2)\, \mathrm{d}\mu = \int_{N^c} L(\theta_1)\, \mathrm{d}\mu = P^{\theta_1}(N^c) = 1.$$

Di conseguenza, per ogni $A \in \mathcal{F}$ si ha

$$P^{\theta_2}(A) = P^{\theta_2}(A \cap N^c) + \underbrace{P^{\theta_2}(A \cap N)}_{=0} = \int_{A \cap N^c} L(\theta_2)\,\mathrm{d}\mu = \int_{A \cap N^c} L(\theta_1)\,\mathrm{d}\mu$$
$$= P^{\theta_1}(A \cap N^c) = P^{\theta_1}(A \cap N^c) + \underbrace{P^{\theta_1}(A \cap N)}_{=0} = P^{\theta_1}(A).$$

□

Vediamo infine che relazione c'è tra l'informazione di Fisher e quella di Kullback.

Teorema 7.9 *Sotto le ipotesi necessarie affinché sia possible definire l'informazione di Fisher, e se la funzione $\theta \mapsto I(\theta_1/\theta)$ è due volte derivabile sotto il segno di integrale, si ha*

$$I(\theta_1)_{i,j} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} I(\theta_1/\theta)\Big|_{\theta=\theta_1}, \qquad i,j = 1,\dots,k.$$

Dimostrazione. Derivando rispetto a θ_j l'espressione

$$I(\theta_1/\theta) = \int \log \frac{L(\theta_1)}{L(\theta)} L(\theta_1)\,\mathrm{d}\mu$$

si trova

$$\frac{\partial}{\partial\theta_j} I(\theta_1/\theta) = \int -\frac{\partial}{\partial\theta_j} \log L(\theta) L(\theta_1)\,\mathrm{d}\mu = \int -\frac{1}{L(\theta)} \frac{\partial}{\partial\theta_j} L(\theta) L(\theta_1)\,\mathrm{d}\mu, \tag{8}$$

e derivando ora rispetto a θ_i (derivata del rapporto)

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} I(\theta_1/\theta) = \int \Big[\frac{\frac{\partial L(\theta)}{\partial\theta_i} \frac{\partial L(\theta)}{\partial\theta_j}}{L^2(\theta)} - \frac{\frac{\partial^2 L(\theta)}{\partial\theta_i \partial\theta_j}}{L(\theta)} \Big] L(\theta_1)\,\mathrm{d}\mu.$$

Calcolando per $\theta = \theta_1$ si ha

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} I(\theta_1/\theta)\Big|_{\theta=\theta_1} = \int \Big[\frac{\frac{\partial L(\theta)}{\partial\theta_i} \frac{\partial L(\theta)}{\partial\theta_j}}{L^2(\theta)} - \frac{\frac{\partial^2 L(\theta)}{\partial\theta_i \partial\theta_j}}{L(\theta)} \Big] L(\theta_1)\,\mathrm{d}\mu\Big|_{\theta=\theta_1} = \int \frac{\frac{\partial L(\theta)}{\partial\theta_i} \frac{\partial L(\theta)}{\partial\theta_j}}{L^2(\theta)} L(\theta_1)\,\mathrm{d}\mu\Big|_{\theta=\theta_1} \tag{9}$$

dato che

$$\int \frac{\frac{\partial^2 L(\theta)}{\partial\theta_i \partial\theta_j}}{L(\theta)} L(\theta_1)\,\mathrm{d}\mu\Big|_{\theta=\theta_1} = \int \frac{\frac{\partial^2 L(\theta_1)}{\partial\theta_i \partial\theta_j}}{L(\theta_1)} L(\theta_1)\,\mathrm{d}\mu = \int \frac{\partial^2 L(\theta_1)}{\partial\theta_i \partial\theta_j}\,\mathrm{d}\mu = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \int L(\theta_1)\,\mathrm{d}\mu$$
$$= \frac{\partial^2}{\partial\theta_i \partial\theta_j} 1 = 0.$$

Ritornando alla relazione (9), si ha infine

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} I(\theta_1/\theta)\Big|_{\theta=\theta_1} = \int \frac{\frac{\partial L(\theta)}{\partial\theta_i} \frac{\partial L(\theta)}{\partial\theta_j}}{L^2(\theta)} L(\theta_1)\,\mathrm{d}\mu\Big|_{\theta=\theta_1} = \int \frac{\frac{\partial L(\theta)}{\partial\theta_i} \frac{\partial L(\theta)}{\partial\theta_j}}{L^2(\theta)}\,\mathrm{d}P^{\theta_1}\Big|_{\theta=\theta_1}$$
$$= E^{\theta_1}\Big[\frac{\frac{\partial L(\theta)}{\partial\theta_i} \frac{\partial L(\theta)}{\partial\theta_j}}{L^2(\theta)} \Big]\Big|_{\theta=\theta_1} = E^{\theta_1}\Big[\frac{\frac{\partial L(\theta_1)}{\partial\theta_i} \frac{\partial L(\theta_1)}{\partial\theta_j}}{L^2(\theta_1)} \Big] = I(\theta_1).$$

□

Osservazione 7.10 (a) È facile vedere che $\nabla I(\theta_1/\theta)|_{\theta=\theta_1} = 0$. Infatti, come abbiamo visto nella dimostrazione precedente (relazione (8))

$$-\nabla I(\theta_1/\theta)|_{\theta=\theta_1} = \int \nabla \log L(\theta)\big|_{\theta=\theta_1} L(\theta_1)\,\mathrm{d}\mu = \int \nabla L(\theta)\big|_{\theta=\theta_1} \cdot \frac{1}{L(\theta_1)} L(\theta_1)\,\mathrm{d}\mu$$
$$= \int \nabla L(\theta)\big|_{\theta=\theta_1}\,\mathrm{d}\mu = \nabla\Big\{\int L(\theta)\,\mathrm{d}\mu\Big\}\Big|_{\theta=\theta_1} = \nabla\Big\{\int \mathrm{d}P^\theta\Big\}\Big|_{\theta=\theta_1} = \nabla 1 = 0.$$

Questo fatto è naturale, perché, come abbiamo visto, $I(\theta_1/\theta) \geq 0$ per ogni θ e $I(\theta_1/\theta_1) = 0$. Dunque θ_1 è punto di minimo per la funzione $\theta \mapsto I(\theta_1/\theta)$.

(b)Dal Teorema appena dimostrato si deduce che l'hessiano di $\theta \mapsto I(\theta_1/\theta)$, calcolato in θ_1, è semidefinito positivo (im quanto uguale a $I(\theta_1)$, che, come sappiamo, è una matrice di covariamza). Anche questo è naturale, per gli stessi motivi del punto (a).

(c) Il Teorema precedente permette di precisare l'affermazione (fatta all'inizio del §6) che l'informazione di Fisher serva per descrivere la variazione "locale" delle leggi P^θ. Per semplicità supponiamo che il parametro θ sia unidimensionale. Per il teorema 7.9 e per il punto (a) precedente si ha, approssimando al secondo ordine con la formula di Taylor nell'intorno di θ_1

$$I(\theta_1/\theta) \approx I(\theta_1/\theta_1) + \frac{\mathrm{d}}{\mathrm{d}\theta} I(\theta_1/\theta)\Big|_{\theta=\theta_1} (\theta - \theta_1) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} I(\theta_1/\theta)\Big|_{\theta=\theta_1} (\theta - \theta_1)^2 = \frac{1}{2} I(\theta_1)(\theta - \theta_1)^2.$$

Dunque, più $I(\theta_1)$ è vicino a 0 e più il grafico della funzione $\theta \mapsto I(\theta_1/\theta)$ è "piatto " nell'intorno di θ_1, ed è quindi difficile discriminare tra il vero valore θ_1 e i valori di θ nel suo intorno. La situazione è ovviamente inversa per valori grandi di $I(\theta_1)$ (ved. figure).

8 Stimatori di massima verosimiglianza

Il metodo della massima verosimiglianza è largamente usato in statistica per la sua presentazione intuitiva e soprattutto per la sua semplicità; tuttavia le giustificazioni rigorose sono solo asintotiche, e pertanto dovrebbe essere usato solo quando si dispone di campioni molto numerosi.

Esempio 8.1 (introduttivo). Una moneta dà testa con probabilità θ; il valore di θ non è noto; si sa però che esso è uguale a $\frac{1}{2}$ oppure a $\frac{1}{10}$. Un tizio, che deve stabilire quale di questi due valori è quello giusto, decide di effettuare n lanci di una moneta, ed ottiene in ciascun lancio la faccia "testa". A questo punto, come è facile capire, egli è propenso a credere che la probabilità che la moneta dia testa è $\frac{1}{2}$. Ritiene infatti che il risultato ottenuto (n volte "testa"in n lanci) potrebbe sì verificarsi anche nell'altro caso, ma con una probabilità inferiore.

Cerchiamo di formalizzare la situazione. Il modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$è quello di un campione $(X_1, \ldots, X_n)$ avente legge $\mathcal{B}(1, \theta)$, con $\theta \in \Theta = \{\frac{1}{2}, \frac{1}{10}\}$. La verosimiglianza (rispetto alla misura che conta i punti di $\{0,1\}^n$) è , come è noto (ved. esempio (a) sulle statistiche esaustive)

$$L(\theta, \omega) = \theta^{\sum_{i=1}^n \omega_i} (1-\theta)^{n - \sum_{i=1}^n \omega_i}.$$

La probabilità del risultato

$$\omega = \underbrace{(1, 1, \ldots, 1)}_{n \text{ volte}}$$

è quindi

$$L\big(\theta, (1,1,\ldots,1)\big) = \theta^n = \begin{cases} \big(\frac{1}{10}\big)^n & \text{se } \theta = \frac{1}{10} \\ \\ \big(\frac{1}{2}\big)^n & \text{se } \theta = \frac{1}{2}. \end{cases}$$

Il tizio ha dunque deciso di considerare vero il valore di θ per il quale il risultato effettivamente osservato è più probabile. In altre parole, egli ha calcolato

$$\max_{\theta\in\theta} L\big(\theta,(1,1,\dots,1)\big) = \max_{\theta\in\Theta}\theta^n,$$

ed ha deciso per il valore del parametro in cui tale massimo è raggiunto, cioè per il punto di massimo della funzione

$$\theta \mapsto L\big(\theta,(1,1,\dots,1)\big).$$

Più in generale, se egli avesse ottenuto il risultato $\omega=(\omega_1,\dots,\omega_n)$, la sua decisione sarebbe stata quella di prendere come vero valore del parametro il punto di massimo della funzione $\theta\mapsto L(\theta,\omega)$.

Passiamo alla formalizzazione generale. Consideriamo un modello statistico $\big(\Omega,\mathcal{F},\{P^\theta,\theta\in\Theta\}\big)$dominato dalla misura μ. Per il momento supporremo che $\Theta\subseteq\mathbb{R}$. Supponiamo assegnata una scelta della verosimiglianza $L(\theta,\omega)=\frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}$.

Definizione 8.2 Sia $U:\Omega\to\Theta$ uno stimatore del parametro θ. Si dice che U è uno *stimatore di massima verosimiglianza* se, per ogni $\omega\in\Omega$, si ha

$$L\big(U(\omega),\omega\big)=\sup_{\theta\in\Theta} L(\theta,\omega)$$

(ovviamente questo estremo superiore sarà un massimo se viene assunto in qualche punto θ_0).

In generale non è affatto detto che un tale stimatore esista, oppure che sia univocamente determinato; tuttavia, quando esiste, generalmente è facile calcolarlo, e si usa denotarlo con il simbolo $\hat\theta$. Se la funzione $\theta\mapsto L(\theta,\omega)$ (a ω fissato) è differenziabile, $\hat\theta$ verifica l'equazione

$$\frac{\mathrm{d}}{\mathrm{d}\theta}L(\theta,\omega)\Big|_{\theta=\hat\theta(\omega)}=0 \tag{10}$$

(attenzione: questa è naturalmente una condizione solo necessaria). Ovviamente è inutile cercare il massimo della funzione $\theta\mapsto L(\theta,\omega)$ per gli (eventuali) ω nei quali essa vale 0. D'altra parte, se $\omega\in\{\omega: L(\theta,\omega)>0\}$, e dato che la funzione $x\mapsto\log x$ è crescente, è chiaro poi che la (10) equivale all'equazione

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log L(\theta,\omega)\Big|_{\theta=\hat\theta(\omega)}=0,$$

che viene detta *equazione di massima verosimiglianza*, e risulta spesso più maneggevole della (10) (soprattutto in presenza di un modello esponenziale, come vedremo).

Esempio 8.3 Sia $(\mu^\theta)_{\theta\in\Theta}$ una famiglia di misure di probabilità su $\mathbb{R}$, con $\Theta\subseteq\mathbb{R}$, dominata dalla misura μ. Sia $f^\theta(x)=\frac{\mathrm{d}\mu^\theta}{\mathrm{d}\mu}$ una densità, e consideriamo un campione di taglia n e legge μ^θ. La verosimiglianza (rispetto alla misura $\mu^{\otimes n}$) è

$$L\big(\theta;x_1,\dots,x_n\big)\big)=\prod_{i=1}^n f^\theta(x_i)$$

e l'equazione (10) diventa

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\prod_{i=1}^n f^\theta(x_i)\Big)\Big|_{\theta=\hat\theta(\omega)}=0;$$

dunque l'equazione di massima verosimiglianza è

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\sum_{i=1}^n \log f^\theta(x_i)\Big)\Big|_{\theta=\hat\theta(\omega)}=0.$$

Esempio 8.4 Consideriamo un campione di legge $\mathcal{E}(\theta)$, con $\theta > 0$. In questo caso si ha

$$f^\theta(x) = \begin{cases} \theta \mathrm{e}^{-\theta x} & \text{per } x > 0 \\ 0 & \text{altrove.} \end{cases}$$

La verosimiglianza è

$$L(\theta; x_1, \dots, x_n) = \begin{cases} \theta^n \mathrm{e}^{-\theta(x_1+\dots+x_n)} & \text{se } x_i > 0,\ \forall i \\ 0 & \text{altrove.} \end{cases}$$

L'equazione di massima verosimiglianza è

$$\sum_{i=1}^n \Big(\frac{1}{\theta} - x_i\Big) = 0,$$

e cioè

$$\frac{n}{\theta} - \sum_{i=1}^n x_i = 0,$$

che ha la (unica) soluzione $\theta = \frac{n}{\sum_{i=1}^n x_i}$; si verifica poi facilmente che si tratta di un punto di massimo, per cui lo stimatore di massima verosimiglianza di θ è

$$\hat\theta = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\overline{X}}.$$

(osservazione euristica: il parametro di un'esponenziale è uguale all'inversa della media, e in corrispondenza lo stimatore di massima verosimiglianza risulta essere, in questo caso, l'inverso della media campionaria).

Vogliamo vedere se $\hat\theta$ è uno stimatore corretto di θ (ved. Esercizio (6.29)). Sotto P^θ la v.a. $\sum_{i=1}^n X_i$ ha legge $\Gamma(n, \theta)$. Dunque

$$E^\theta\Big[\frac{n}{\sum_{i=1}^n X_i}\Big] = \frac{n}{\Gamma(n)} \int_0^{+\infty} \frac{1}{x} \theta^n x^{n-1} \mathrm{e}^{-\theta x}\, \mathrm{d}x = \frac{n\theta}{\Gamma(n)} \int_0^{+\infty} (\theta x)^{n-2} \mathrm{e}^{-\theta x}\, (\theta \mathrm{d}x)$$
$$= \frac{n\theta}{\Gamma(n)} \int_0^{+\infty} y^{n-2} \mathrm{e}^{-y} \mathrm{d}y = \frac{n\theta}{\Gamma(n)} \Gamma(n-1) = \frac{n\theta}{n-1};$$

si deduce che $\hat\theta$ non è corretto (lo è solo asintoticamente, cioè per $n \to \infty$), mentre è corretto lo stimatore di θ

$$U = \frac{n-1}{\sum_{i=1}^n X_i}.$$

Sappiamo dalla teoria dei modelli esponenziali che $T = \sum_{i=1}^n X_i$ è una statistica esaustiva completa. Dunque U, essendo uno stimatore di θ T-misurabile e corretto, è ottimale, per il Teorema 4.1 (di Blackwell-Rao). Osserviamo che $\hat\theta$ e U differiscono di poco; questo succede molto spesso nel caso degli stimatori di massima verosimiglianza.

Esempio 8.5 Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$, con verosimiglianza

$$L(m, \sigma^2; x_1, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\Big(- \frac{\sum_{i=1}^n (x_i - m)^2}{2\sigma^2}\Big).$$

Vogliamo calcolare gli stimatori di massima verosimiglianza di m e di σ^2. Nel caso multidimensionale l'equazione di massima verosimiglianza è ovviamente

$$\nabla \log L(\theta, \omega))\Big|_{\theta = \hat\theta(\omega)} = 0,$$

e in questo caso diventa il sistema

$$\begin{cases} \frac{\partial}{\partial m} \log L(m, \sigma^2; x_1, \dots, x_n) = 0 \\ \frac{\partial}{\partial(\sigma^2)} \log L(m, \sigma^2; x_1, \dots, x_n) = 0. \end{cases}$$

Ricordando i calcoli fatti nell' Esempio 6.10, (a) e (b), possiamo scrivere il sistema nella forma

$$\begin{cases} \frac{\sum_{i=1}^n (x_i - m)}{\sigma^2} = 0 \\ \\ -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - m)^2}{2\sigma^4} = 0, \end{cases}$$

ed ha la soluzione

$$(m, \sigma^2) = \Big(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n} \Big)$$

Lo stimatore di massima verosimiglianza di (m, σ^2) è dunque

$$\Big(\frac{\sum_{i=1}^n X_i}{n}, \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n} \Big) = \Big(\overline{X}, \frac{n-1}{n} S^2 \Big),$$

di nuovo solo asintoticamente corretto.

Con calcoli simili ai precedenti (ma più semplici) si può vedere che

(i) nel caso di varianza nota, lo stimatore di massima verosimiglianza di m è ancora $\overline{X}$;

(ii) nel caso di media nota, lo stimatore di massima verosimiglianza di σ^2 è $\frac{\sum_{i=1}^n (X_i - m)^2}{n}$.

Esercizio 8.6 (a) Calcolare lo stimatore di massima verosimiglianza del parametro $\theta > 0$ basandosi su un campione di taglia n e di legge Π_θ.

(b) Calcolare lo stimatore di massima verosimiglianza del parametro $\theta > -1$ basandosi su un campione di taglia n e legge avente densità

$$f^\theta(x) = \begin{cases} (\theta + 1) x^\theta & \text{per } x \in [0, 1] \\ 0 & \text{altrove.} \end{cases}$$

Come abbiamo accennato all'inizio, per gli stimatori di massima verosimiglianza non si possono dare che dei risultati asintotici (cioè per la taglia n del campione che tende a ∞). Bisogna allora costruire un modello statistico che idealizzi una successione infinita di esperimenti. Cominciamo con l'ammettere il risultato seguente di Teoria della Misura (è un caso particolare del *Teorema di Ionescu-Tulcea*):

Teorema 8.7 *Sia μ una probabilità su $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$. Esiste su $\mathbb{R}^{\mathbb{N}} = \prod_{n \in \mathbb{N}} \mathbb{R}_n$, munito della σ-algebra prodotto $\bigotimes_{n \in \mathbb{N}} \mathcal{B}(\mathbb{R}_n)$, una e una sola probabilità (indicata usualmente con $\mu^{\otimes \mathbb{N}}$) tale che, per ogni $k \in \mathbb{N}$ e per ogni successione finita di boreliani $A_1, \dots, A_k$ risulti*

$$\mu^{\otimes \mathbb{N}}(A_1 \times A_2 \times \cdots \times A_k \times \mathbb{R} \times \mathbb{R} \times \cdots) = \mu(A_1) \cdot \ldots \cdot \mu(A_k).$$

Inoltre, quando $\big(\mathbb{R}^{\mathbb{N}}, \bigotimes_{n \in \mathbb{N}} \mathcal{B}(\mathbb{R}_n)\big)$ è munito di questa probabilità, le proiezioni canoniche $X_i(\omega) = X_i(x_1, x_2, \dots, x_i, \dots) = x_i$ sono indipendenti ed hanno tutte legge μ.

Nei teoremi che seguiranno, considereremo una famiglia $\{\mu^\theta, \theta \in \Theta\}$ di probabilità su $\mathbb{R}$ e un campione infinito di legge μ^θ. Il modello statistico sarà

$$\Big(\mathbb{R}^{\mathbb{N}}, \bigotimes_{n\in\mathbb{N}} \mathcal{B}(\mathbb{R}_n), \{P^\theta, \theta \in \Theta\}\Big),$$

dove si è posto $P^\theta = (\mu^\theta)^{\otimes\mathbb{N}}$.

Supporremo che le misure μ^θ siano dominate da una misura μ e porremo $\frac{\mathrm{d}\mu^\theta}{\mathrm{d}\mu} = f^\theta(x)$; chiameremo *successione di stimatori di massima verosimiglianza* una successione di v.a. $\hat\theta_n$ tali che, per ogni n, $\hat\theta_n$ sia uno stimatore di massima verosimiglianza per il campione $(X_1, \dots, X_n)$ (cioè $\hat\theta_n$ è funzione solo delle v.a. $X_1, \dots, X_n$ e si ha

$$L_n(\hat\theta_n; X_1, \dots, X_n) = \sup_{\theta\in\Theta} L_n(\theta; X_1, \dots, X_n),$$

dove

$$L_n(\theta; X_1, \dots, X_n) = \prod_{i=1}^n f^\theta(X_i)$$

è la verosimiglianza del campione $(X_1, \dots, X_n)$).

Noi considereremo solo il caso di un campione infinito con legge appartenente ad un modello esponenziale su $\mathbb{R}$ con verosimiglianza

$$\frac{\mathrm{d}\mu^\theta}{\mathrm{d}\mu}(x) = f^\theta(x) = \exp\big(\theta T(x) - \psi(\theta)\big), \quad \theta \in \Theta,$$

dove Θ è un intervallo della retta (l'esame di casi più generali è possibile, ma a prezzo di serie complicazioni). L'equazione di massima verosimiglianza è in questo caso

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\sum_{i=1}^n \log f^\theta(x_i)\Big) = \sum_{i=1}^n \frac{\mathrm{d}}{\mathrm{d}\theta}\log f^\theta(x_i) = \sum_{i=1}^n \frac{\mathrm{d}}{\mathrm{d}\theta}\big(\theta T(x_i) - \psi(\theta)\big) = \sum_{i=1}^n T(x_i) - n\psi'(\theta) = 0,$$

e cioè

$$\psi'(\theta) = \frac{\sum_{i=1}^n T(x_i)}{n}. \tag{11}$$

Come sappiamo dal Teorema 5.11, punto (b), si ha

$$\psi''(\theta) = Var^\theta(T) > 0$$

(se fosse $Var^\theta(T) = 0$, T sarebbe costante e quindi $f^\theta(x)$ sarebbe costante (in x). Questo non è possibile perchè l'unica funzione costante integrabile su $\mathbb{R}$ è la funzione identicamente nulla, che non è una densità di probabilità).

Dunque ψ' è strettamente crescente, e quindi biunivoca, da Θ a $\psi'(\Theta)$, e quindi l'equazione di massima verosimiglianza (11) ha al più una soluzione. Più precisamente, la soluzione è

$$\hat\theta_n = (\psi')^{-1}\Big(\frac{\sum_{i=1}^n T(x_i)}{n}\Big), \tag{12}$$

a patto che $\frac{\sum_{i=1}^n T(x_i)}{n} \in \psi'(\Theta)$. Le ipotesi del Teorema che segue garantiscono questa condizione, almeno per valori grandi di n.

Teorema 8.8 *Supponiamo che ψ sia di classe C^2 e che $\psi'(\Theta)$ sia un intervallo aperto. Allora, per ogni fissato θ_0, è definito P^{θ_0}-q.c. per n abbastanza grande lo stimatore di massima verosimiglianza $\hat{\theta}_n$ di θ_0. Inoltre $\hat{\theta}_n \to \theta_0$ per $n \to \infty$, P^{θ_0}-q.c.*

Osservazione 8.9 Detto in termini precisi, l'enunciato precedente significa che, per P^{θ_0}-q.o. ω, esiste un intero $n_0 = n_0(\omega, \theta_0)$ tale che, per ogni $n > n_0$ l'equazione (nell'incognita θ)

$$L_n(\theta; X_1(\omega), \dots, X_n(\omega)) = \sup_{\theta \in \Theta} L_n(\theta; X_1(\omega), \dots, X_n(\omega))$$

ha una e una sola soluzione $\hat{\theta}_n(\omega)$ ed inoltre $\lim_{n\to\infty} \hat{\theta}_n(\omega) = \theta_0$.

Dimostrazione. Per la Legge Forte dei Grandi Numeri, la v.a. $\frac{\sum_{i=1}^n T(X_i)}{n}$ converge P^{θ_0}-q.c. verso $E^{\theta_0}[T(X_1)] = \psi'(\theta_0)$ (dove l'ultima uguaglianza segue dal Teorema 5.11, punto (a)). Dato che $\psi'(\Theta)$ è un intervallo aperto, per n abbastanza grande $\frac{\sum_{i=1}^n T(X_i)}{n}$ appartiene a $\psi'(\Theta)$, in quanto convergente a $\psi'(\theta_0) \in \psi'(\Theta)$. Dunque

$$\hat{\theta}_n = (\psi')^{-1}\Big(\frac{\sum_{i=1}^n T(x_i)}{n}\Big)$$

è definito (per l'osservazione fatta sopra).

Infine, per $n \to \infty$ e P^{θ_0}-q.c.,

$$\hat{\theta}_n = (\psi')^{-1}\Big(\frac{\sum_{i=1}^n T(x_i)}{n}\Big) \to (\psi')^{-1}\big(\psi'(\theta_0)\big) = \theta_0,$$

dato che $(\psi')^{-1}$ è continua.

□

Osservazione 8.10 Dal Teorema precedente si deduce che, a differenza di quanto succedeva per il campione finito (di taglia n fissata, per una famiglia esponenziale di leggi di probabilità su $\mathbb{R}$), nel campione infinito le probabilità P^θ sono tutte tra loro estranee (cioè portate da insiemi disgiunti) Infatti, sia $\theta_1 \neq \theta_2$, e per $i = 1, 2$ poniamo $A_i = \{\omega \in \Omega : \hat{\theta}_n(\omega) \to \theta_i\}$. Ovviamente A_1 e A_2 sono disgiunti, ed inoltre $P^{\theta_i}(A_i) = 1$ per il Teorema precedente. Dunque il modello del campione infinito non è regolare.

Definizione 8.11 Si dice che la successione $(\hat{\theta}_n)_{n\in\mathbb{N}}$ è *consistente* (risp. *fortemente consistente*) se, per ogni $\theta \in \Theta$, rispetto alla probabiltà P^θ, $\hat{\theta}_n$ converge a θ in probabilità (risp. quasi certamente).

Dunque

Corollario 8.12 *Nelle ipotesi del Teorema precedente, la successione $(\hat{\theta}_n)_{n\in\mathbb{N}}$ è fortemente consistente.*

Per maggiore chiarezza, ripetiamo in un caso concreto il ragionamento fatto nel Teorema 8.8.

Esempio 8.13 Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{E}(\theta)$, con $\theta \in \Theta = (0, 1)$. Abbiamo visto nell'Esempio 8.4 che l'equazione di massima verosimiglianza è

$$\sum_{i=1}^n \Big(\frac{1}{\theta} - x_i\Big) = 0,$$

e dunque l'eventuale (unica) soluzione è

$$\theta = \frac{n}{\sum_{i=1}^n X_i},$$

che, scriita in termini delle osservazioni, significa

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^n X_i}.$$

Questa soluzione esiste solo se $\frac{n}{\sum_{i=1}^n X_i} \in (0,1)$, e questo non è necessariamente vero (dipende dai valori delle osservazioni: se ottenessimo valori X_i molto piccoli, $\frac{n}{\sum_{i=1}^n X_i}$ sarebbe un valore grande, eventualmente più grande di 1). Tuttavia, la Legge Forte dei Grandi Numeri assicura che $\frac{\sum_{i=1}^n X_i}{n}$ converge P^{θ_0}-q.c. verso $\frac{1}{\theta_0}$, e dunque $\hat{\theta}_n$ converge P^{θ_0}-q.c. verso $\theta_0 \in (0,1)$. Dunque $\hat{\theta}_n$ è definito a partire da un certo $n_0 = n_0(\omega, \theta_0)$ in poi.

Vediamo ora un risultato di convergenza in legge.

Teorema 8.14 *Nelle ipotesi del Teorema 8.8, per ogni $\theta \in \Theta$, $\sqrt{n}(\hat{\theta}_n - \theta)$ converge in legge (secondo la probabilità P^θ) alla $\mathcal{N}\big(0, \frac{1}{\psi''(\theta)}\big)$.*

Dimostrazione. La dimostrazione del Teorema si basa sul Lemma seguente, la cui prova è rimandata alla fine del paragrafo.

Lemma 8.15 *(*Metodo Delta*). Sullo spazio $(\Omega, \mathcal{F}, P)$ sia $(U_n)_{n\in\mathbb{N}}$ una successione di v.a. convergente P-q.c. verso la costante a e tale che $\sqrt{n}(U_n - a)$ converga in legge verso la $\mathcal{N}(0, \sigma^2)$ (con σ^2 assegnato). Sia poi g una funzione di classe C^2 definita in un intorno di a. Allora la successione $\sqrt{n}\big(g(U_n) - g(a)\big)$ converge in legge alla $\mathcal{N}\big(0, (g'(a))^2\sigma^2\big)$.*

Ricordando l'espressione (12) dello stimatore di massima verosimiglianza, basta applicare il Metodo Delta con

$$U_n = \frac{\sum_{i=1}^n T(X_i)}{n}, \quad a = \psi'(\theta), \quad g = (\psi')^{-1},$$

osservando che le ipotesi del Metodo sono soddisfatte grazie alla Legge Forte dei Grandi Numeri (ved. inizio della dimostrazione del Teorema (8.8)) e alle seguenti considerazioni:

(a) si ha

$$\sqrt{n}(U_n - a) = \frac{\sum_{i=1}^n T(X_i) - n\psi'(\theta)}{\sqrt{n}},$$

che converge verso la $\mathcal{N}(0, \psi''(\theta))$ (ricordare che le $T(X_i)$ sono i.i.d., con $E[T(X_i)] = \psi'(\theta)$, $Var^\theta(T(X_i) = \psi''(\theta))$;

(b) si ha

$$g'(a) = \Big(\frac{\mathrm{d}}{\mathrm{d}\theta}(\psi')^{-1}\Big)(\psi'(\theta)) = \frac{1}{\psi''\Big((\psi')^{-1}(\psi'(\theta))\Big)} = \frac{1}{\psi''(\theta)}.$$

□

Passiamo alla dimostrazione del Lemma 8.15. Ci servono due Lemmi.

Lemma 8.16 *Sullo spazio $(\Omega, \mathcal{F}, P)$ sia $(X_n)_{n\in\mathbb{N}}$ una successione di v. a. convergente in legge verso una v.a. X P-q.c. finita, e $(Y_n)_{n\in\mathbb{N}}$ una successione convergente a 0 in probabilità (o, ciò che è lo stesso, in legge). Allora*

$$X_n Y_n \xrightarrow{P} 0.$$

DIMOSTRAZIONE. Dato che $X_n \to^{\mathcal{L}} X$, per ogni funzione f continua e limitata si ha

$$\int f(X_n)\,\mathrm{d}P \to \int f(X)\,\mathrm{d}P. \tag{13}$$

Sia $c > 0$, e consideriamo la funzione

$$g = 1_{(-\infty,-c)\cup(c,+\infty)}.$$

Questa funzione non è continua, ma esiste una successione decrescente $(f_m)_{m\in\mathbb{N}}$ di funzioni continue e limitate tali che $g = \inf_m f_m$. Dunque per ogni m fissato si ha $f_m \geq g$ e quindi, per la (13),

$$\int f_m(X)\,\mathrm{d}P = \lim_{n\to\infty} \int f_m(X_n)\,\mathrm{d}P \geq \limsup_{n\to\infty} \int g(X_n)\,\mathrm{d}P = \limsup_{n\to\infty} P(|X_n| > c).$$

D'altra parte, per il Teorema di convergenza monotona,

$$\lim_{m\to\infty} \int f_m(X)\,\mathrm{d}P = \int g(X)\,\mathrm{d}P = P(|X| > c)$$

e quindi si conclude che

$$P(|X| > c) \geq \limsup_{n\to\infty} P(|X_n| > c).$$

Dato che

$$\lim_{c\to+\infty} P(|X| > c) = \lim_{c\to+\infty} P(X < -c) + 1 - P(X \leq c) = 0,$$

fissato $\epsilon > 0$ esiste c_ϵ tale che

$$\limsup_{n\to\infty} P(|X_n| > c_\epsilon) \leq P(|X| > c_\epsilon) \leq \epsilon.$$

Sia ora $\delta > 0$ fissato. Si ha

$$\big\{|X_nY_n| > \delta\big\} \subseteq \big\{|X_n| > c_\epsilon\big\} \cup \Big\{|Y_n| > \frac{\delta}{c_\epsilon}\Big\}$$

e quindi

$$P\big(|X_nY_n| > \delta\big) \leq P\big(|X_n| > c_\epsilon\big) + P\Big(|Y_n| > \frac{\delta}{c_\epsilon}\Big) \leq \epsilon + P\Big(|Y_n| > \frac{\delta}{c_\epsilon}\Big).$$

Mandando $n \to \infty$, dato che Y_n converge a 0 in probabilità, si trova

$$\limsup_{n\to\infty} P\big(|X_nY_n| > \delta\big) \leq \epsilon,$$

e si conclude per l'arbitrarietà di ϵ.

□

Osservazione 8.17 Se X non è P-quasi certamente finita, la tesi può non valere. Per esempio, sia X_n una successione convergente P-quasi certamente verso $X = +\infty$. Prendiamo $Y_n = \frac{1}{X_n}$, che ovviamente converge a 0. Allora X_nY_n converge a 1.

Lemma 8.18 (TEOREMA DI SLUTSKY) *Se $(X_n)_{n\in\mathbb{N}}$ converge in legge a X e $(Y_n)_{n\in\mathbb{N}}$ converge a c in probabilità (o equivalentemente in legge), allora $X_n + Y_n$ converge a $X + c$ in legge.*

DIMOSTRAZIONE. Osserviamo prima di tutto che $\tilde{Y}_n := Y_n - c$ converge in legge a 0; supponiamo allora di aver dimostrato che $X_n + \tilde{Y}_n$ converge in legge a X. Ne segue (indicando con ϕ_U la funzione caratteristica di una v.a. U)

$$\phi_{X_n+Y_n}(t) = E\big[\mathrm{e}^{\mathrm{i}t(X_n+\tilde{Y}_n)}\big]\mathrm{e}^{\mathrm{i}tc} \to E\big[\mathrm{e}^{\mathrm{i}tX}\big]\mathrm{e}^{\mathrm{i}tc} = E\big[\mathrm{e}^{\mathrm{i}t(X+c)}\big], \qquad n \to \infty,$$

dunque $X_n + Y_n$ converge in legge a $X + c$ per il Teorema di continuità. Dunque supporremo che $c = 0$.
Per un criterio di convergenza in legge (ved, [1], Th. 18.7), $X_n + Y_n$ converge in legge a X se e solo se $\lim_{n\to\infty} E[f(X_n + Y_n)] = E[f(X)]$ per ogni funzione f continua, lipschitziana (con costante di Lipschitz L_f) e limitata(da una costante M_f). Poiché $\lim_{n\to\infty} E[f(X_n)] = E[f(X)]$ (X_n converge in legge a X), basterà far vedere che $\lim_{n\to\infty}\big(E[f(X_n + Y_n)] - E[f(X_n)]\big) = 0$. Ora

$$\begin{aligned}
&\limsup_{n\to\infty} \big|E[f(X_n + Y_n)] - E[f(X_n)]\big| \le \limsup_{n\to\infty} E\big[|f(X_n + Y_n) - f(X_n)|\big] \\
&= \limsup_{n\to\infty} \big(E\big[|f(X_n + Y_n) - f(X_n)|1_{\{Y_n\le\epsilon\}}\big] + E\big[|f(X_n + Y_n) - f(X_n)|1_{\{Y_n>\epsilon\}}\big]\big) \\
&\le L_f\epsilon + 2M_f \lim_{n\to\infty} P(|Y_n| > \epsilon).
\end{aligned}$$

Si conclude passando al limite, per l'arbitrarietà di ϵ.

□

Passiamo alla dimostrazione del Metodo Delta. Sviluppando con la formula di Taylor al secondo ordine si ha

$$g(x) - g(a) = g'(a)(x - a) + \frac{(x-a)^2}{2}g''(\xi),$$

con ξ opportuno. Quindi

$$\sqrt{n}\big(g(U_n) - g(a)\big) = \sqrt{n}g'(a)(U_n - a) + \frac{g''(M_n)}{2}(U_n - a)\sqrt{n}(U_n - a).$$

Il primo addendo converge in legge alla $\mathcal{N}\big(0, (g'(a))^2\sigma^2\big)$. Il secondo addendo è il prodotto di $\frac{g''(M_n)}{2}$, che converge a $\frac{g''(a)}{2}$, di $\sqrt{n}(U_n - a)$ (che converge in legge) e di $(U_n - a)$ (che converge quasi certamente e quindi in probabilità a 0). Per il Lemma 8.16 il secondo addendo tende a 0 in probabilità. Per il Lemma 8.18 si conclude che la somma dei due addendi tende in legge alla $\mathcal{N}\big(0, (g'(a))^2\sigma^2\big)$.

□

Osservazione 8.19 Ricordiamo che $\psi''(\theta) = I(\theta)$ (informazione di Fisher, ved. Esempio 6.27). Per un parametro $\theta \in \Theta \subseteq \mathbb{R}^k$ si può dimostrare che, se $(\hat{\theta}_n)_{n\in\mathbb{N}}$ è una successione di stimatori di massima verosimiglianza in un modello esponenziale, allora il vettore $\sqrt{n}(\hat{\theta}_n - \theta)$ converge in legge verso una $\mathcal{N}_k\big(\mathbf{0}, I^{-1}(\theta)\big)$, (legge normale k-dimensionale di vettore delle medie $\mathbf{0}$ e e matrice di covarianza $I^{-1}(\theta)$), dove $I(\theta)$ è la matrice di informazione di Fisher (matrice $k \times k$). Delle leggi normali multidimensionali parleremo nel prossimo paragrafo.

9 Variabili gaussiane e vettori gaussiani

Sullo spazio di probabilità $(\Omega, \mathcal{F}, P)$ consideriamo un campione $X = (X_1, \dots, X_n)$ di legge gaussiana $\mathcal{N}(0, \sigma^2)$. La funzione caratteristica del vettore aleatorio X è ($u \in \mathbb{R}^k$)

$$\begin{aligned}
\phi_X(u) &= E\big[\exp\big(\mathrm{i}\langle u, X\rangle\big)\big] = E\Big[\exp\Big(\mathrm{i}\sum_{j=1}^n u_jX_j\Big)\Big] = E\Big[\prod_{j=1}^n \exp\big\{\mathrm{i}(u_jX_j)\big\}\Big] \\
&= \prod_{j=1}^n E\Big[\exp\big\{\mathrm{i}(u_jX_j)\big\}\Big] = \prod_{j=1}^n \exp\Big(-\frac{\sigma^2}{2}u_j^2\Big) = \exp\Big(-\frac{\sigma^2}{2}\sum_{j=1}^n u_j^2\Big) = \exp\Big(-\frac{\sigma^2}{2}\|u\|^2\Big),
\end{aligned}$$

ricordando che la funzione caratteristica di una v.a. unidimensionale Z di legge $\mathcal{N}(m,\sigma^2)$ è

$$\phi_Z(t) = \exp\Big(\mathrm{i}mt - \frac{\sigma^2}{2}t^2\Big). \tag{14}$$

Consideriamo ora

Proposition 9.1 *Sia una matrice A $n\times n$ ortogonale (cioè invertibile e tale che ${}^tA = A^{-1}$) e sia $Y = AX$. Allora anche Y è un campione di legge $\mathcal{N}(0,\sigma^2)$ (ovvero le sue componenti sono indipendenti e di legge $\mathcal{N}(0,\sigma^2)$).*

Dimostrazione. Basta calcolare la funzione caratteristica di Y:

$$\phi_Y(u) = E\big[\exp\big(\mathrm{i}\langle u, AX\rangle\big)\big] = E\big[\exp\big(\mathrm{i}\langle {}^tAu, X\rangle\big)\big] = \exp\Big(-\frac{\sigma^2}{2}\|{}^tAu\|^2\Big) = \exp\Big(-\frac{\sigma^2}{2}\|u\|^2\Big),$$

dato che, essendo A ortogonale,

$$\|{}^tAu\|^2 = \langle {}^tAu,\, {}^tAu\rangle = \langle A\,{}^tAu, u\rangle = \langle u,u\rangle = \|u\|^2.$$

Si riconosce quindi la funzione caratteristica di un campione di legge $\mathcal{N}(0,\sigma^2)$) (calcolata sopra). □

Ricordiamo che si chiama *legge del chi quadro a n gradi di libertà* (indicata con il simbolo $\chi^2(n)$) la legge $\Gamma(\frac{n}{2},\frac{1}{2})$. Si dice che X ha legge $q\cdot\chi^2(n)$ se la v.a. $\frac{X}{q}$ ha legge $\chi^2(n)$.
L'esercizio che segue fornisce una caratterizzazione della $\chi^2(n)$ più utile per i nostri scopi.

Esercizio 9.2 (a) Sia X una v.a. avente legge $\mathcal{N}(0,1)$. Calcolare la legge della v.a. $Z = X^2$.
(b) Mostrare che la legge $\chi^2(n)$ coincide con la legge della v.a. $Z_1^2+\cdots+Z_n^2$, dove Z_i $(i=1,\dots,n)$ sono indipendenti e tutte di legge $\mathcal{N}(0,1)$.
(c)Mostrare che la media di una(qualsiasi variabile avente) legge $\chi^2(n)$vale n.

Soluzione. (a) Calcoliamo la funzione di ripartizione di Z, e poi "deriviamo". Per $t<0$ si ha evidentemente $P(Z\le t)=0$. Per $t>0$ si ha

$$P(Z\le t) = P(-\sqrt{t}\le Z\le\sqrt{t}) = \int_{-\sqrt{t}}^{\sqrt{t}} \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big)\,\mathrm{d}x.$$

Derivando, per $t>0$ troviamo

$$f_Z(t) = \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{2\sqrt{t}}\cdot\exp\big(-\frac{t}{2}\big) - \frac{1}{\sqrt{2\pi}}\cdot\frac{-1}{2\sqrt{t}}\cdot\exp\big(-\frac{t}{2}\big) = \frac{1}{\sqrt{2\pi t}}\cdot\exp\big(-\frac{t}{2}\big),$$

e quindi in conclusione

$$f_Z(t) = \begin{cases} 0 & \text{per } t<0\\ \frac{1}{\sqrt{2\pi t}}\cdot\exp\big(-\frac{t}{2}\big) & \text{per } t>0,\end{cases}$$

che non è altro che la densità $\Gamma(\frac{1}{2},\frac{1}{2})$.

(b) Sommando n variabili indipendenti, tutte di legge $\Gamma(\frac{1}{2},\frac{1}{2})$ (punto (a)), si trova, come è noto, una v.a. avente legge $\Gamma(\frac{n}{2},\frac{1}{2}) = \chi^2(n)$.

(c) Si ha

$$E[Z_1^2+\cdots+Z_n^2] = E[Z_1^2]+\cdots+E[Z_n^2] = 1+\cdots+1 = n,$$

essendo $E[Z_i^2]$ la varianza di una $\mathcal{N}(0,1)$.

Il Teorema che segue è uno dei più importanti della Statistica.

Teorema 9.3 (DI COCHRAN). *Siano $E_1 \oplus \cdots \oplus E_k$ una decomposizione ortogonale di $\mathbb{R}^n$ in k sottospazi di dimensioni rispettive $r_1, \ldots, r_k$ (con $r_1 + \cdots + r_k = n$) e X un campione di taglia n e legge $\mathcal{N}(0, \sigma^2)$; indichiamo con X_{E_i} la proiezione del vettore X sul sottospazio E_i; allora le v.a. $X_{E_1}, \ldots, X_{E_k}$ sono indipendenti e $\|X_{E_i}\|^2$ ha legge $\sigma^2 \cdot \chi^2(r_i)$, per ogni $i = 1, \ldots, k$.*

DIMOSTRAZIONE. Sia $\eta_1, \ldots, \eta_{r_1}$ una base ortonormale per il sottospazio E_1, $\eta_{r_1+1}, \ldots, \eta_{r_1+r_2}$ una base per E_2, $\ldots$, $\eta_{r_1+r_2+\cdots+r_{k-1}+1}, \ldots, \eta_{r_1+r_2+\cdots+r_k}$ una base per E_k.
Allora $\eta_1, \ldots, \eta_{r_1}, \ldots, \eta_{r_1+r_2+\cdots+r_k}$ formano una base ortonormale per $\mathbb{R}^n$ e la matrice A che ha per righe $\eta_1, \ldots, \eta_{r_1}, \ldots, \eta_{r_1+r_2+\cdots+r_k}$ è ortogonale; inoltre, evidentemente, si ha $(AX)_j = \langle X, \eta_j \rangle$, $j = 1, \ldots, n$. Quindi, per la Proposizione precedente, le v.a. $\langle X, \eta_j \rangle$, $j = 1, \ldots, n$ sono tutte tra loro indipendenti e di legge $\mathcal{N}(0, \sigma^2)$.
Essendo

$$X_{E_1} = \sum_{j=1}^{r_1} \langle X, \eta_j \rangle \eta_j,$$

si vede che X_{E_1} è funzione delle v.a. $\langle X, \eta_j \rangle$ con $j = 1, \ldots, r_1$. Analogamente si vede che X_{E_2} è funzione delle v.a. $\langle X, \eta_j \rangle$ con $j = r_1 + 1, \ldots, r_1 + r_2$ e, in generale, si riconosce che le v.a. $X_{E_1}, \ldots, X_{E_k}$ sono funzioni di gruppi separati di v.a. indipendenti, e dunque sono indipendenti tra loro. Inoltre

$$\|X_{E_1}\|^2 = \sum_{j=1}^{r_1} \langle X, \eta_j \rangle^2,$$

e quindi, per l'Esercizio 9.2 (b), $\|X_{E_1}\|^2$ ha legge $\sigma^2 \cdot \chi^2(r_1)$. Analogamente per gli altri vettori X_{E_i}.

□

Ricordiamo che si chiama *legge di Student a n gradi di libertà* (denotata con $t(n)$) la legge di una v.a. U del tipo

$$U = \frac{X}{\sqrt{Y}} \sqrt{n},$$

dove X e Y sono due v.a. indipendenti, di leggi rispettive $\mathcal{N}(0, 1)$ e $\chi^2(n)$. La legge $t(n)$ ammette densità: una sua versione può essere calcolata con il metodo indicato nell'esercizio che segue. Tale versione risulta essere una funzione pari, e dunque la $t(n)$ è una legge simmetrica (del resto questo può essere mostrato anche a partire dall'espressione di U). Tuttavia per i nostri scopi tale densità non sarà importante.

Si può anche mostrare che la $t(n)$ converge in legge verso la $\mathcal{N}(0, 1)$ quando $n \to \infty$. Siano infatti $X, Z_1, \ldots, Z_n$ v. a. indipendenti definite sullo stesso psazio $\Omega, \mathcal{F}, P)$ e tutte di legge $\mathcal{N}(0, 1)$; posto

$$W_n = \sqrt{\frac{n}{\sum_{i=1}^n Z_k^2}},$$

per l'Esercizio 9.2 (b) la v.a. XW_n ha legge $t(n)$. Inoltre per $n \to \infty$, W_n converge a 1 q.c. (si tratta della reciproca di $\frac{\sum_{i=1}^n Z_k^2}{n}$, che converge a $E[Z_1^2] = VarZ_1 = 1$ per la Legge Forte dei Grandi Numeri). Per il Lemma 8.16 $X(W_n - 1)$ converge a 0 in probabilità, e per il Lemma 8.18 $XW_n = X(W_n - 1) + X$ converge alla $\mathcal{N}(0, 1)$ in legge.

Esercizio 9.4 Siano X e Y due v.a. aventi densità congiunta f (rispetto alla misura di Lebesgue multidimensionale). Calcolare la densità (rispetto alla misura di Lebesgue sulla retta) della v.a.

$$Z = \frac{X}{Y}.$$

SOLUZIONE. Sia g una funzione boreliana. Allora

$$E[g(Z)] = E\Big[g\Big(\frac{X}{Y}\Big)\Big] = \iint_{\mathbb{R}^2} g\Big(\frac{x}{y}\Big) f(x,y)\,\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{+\infty} \mathrm{d}y \int_{-\infty}^{+\infty} g\Big(\frac{x}{y}\Big) f(x,y)\mathrm{d}x$$
$$\int_{-\infty}^{+\infty} \mathrm{d}y \int_{-\infty}^{+\infty} g(z) f(zy,y)|y|\mathrm{d}z = \int_{-\infty}^{+\infty} \mathrm{d}z\, g(z) \int_{-\infty}^{+\infty} f(zy,y)|y|\,\mathrm{d}y = \int_{-\infty}^{+\infty} \mathrm{d}z\, g(z)h(z),$$

dove si pone

$$h(z) = \int_{-\infty}^{+\infty} f(zy,y)|y|\,\mathrm{d}y,$$

che è evidentemente la (una) densità per Z. In particolare, se X e Y sono indipendenti e di densità rispettivamente f_1 e f_2, allora

$$h(z) = \int_{-\infty}^{+\infty} f_1(zy) f_2(y)|y|\,\mathrm{d}y;$$

osserviamo che, se il vettore (X,Y) prende valori in $(\mathbb{R}^+)^2$ le formule precedenti diventano rispettivamente

$$h(z) = \Big(\int_0^{+\infty} f(zy,y)y\,\mathrm{d}y\Big)1_{\mathbb{R}^+}(z), \qquad h(z) = \Big(\int_0^{+\infty} f_1(zy)f_2(y)y\,\mathrm{d}y\Big)1_{\mathbb{R}^+}(z).$$

Per un campione $(X_1,\ldots,X_n)$ (non necessariamente gaussiano), consideriamo le due statistiche

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

Si tratta della *media campionaria* (o *media empirica*) e della *varianza campionaria* (o *varianza empirica*) che abbiamo già incontrate in varie occasioni.

Dal Teorema di Cochran discende un importante corollario, che utilizzaremo varie volte.

Teorema 9.5 *Sia $X = (X_1,\ldots,X_n)$ un campione di taglia n e legge $\mathcal{N}(m,\sigma^2)$. Allora*

(i) la v.a.

$$\frac{\overline{X} - m}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1);$$

(ii) la v.a.

$$\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} = \frac{S^2(n-1)}{\sigma^2} \sim \chi^2(n-1);$$

(iii) $\frac{\overline{X}-m}{\sigma}\sqrt{n}$ e S^2 sono tra loro indipendenti (e di conseguenza lo sono anche $\overline{X}$ e S^2)

(iv) la v.a.

$$\frac{\overline{X} - m}{S}\sqrt{n} \sim t(n-1).$$

DIMOSTRAZIONE. (i) È facile vedere che $\overline{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$: segue dal fatto che $X_1 + \cdots + X_n$ (somma di n variabili indipendenti e tutte di legge $\mathcal{N}(m, \sigma^2)$ ha legge $\mathcal{N}(nm, n\sigma^2)$. Il punto (i) si ottiene allora per standardizzazione di $\overline{X}$.

(ii) Sia E_1 il sottospazio generato dal versore

$$\eta = \Big(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\Big),$$

e sia $E_2 = E_1^\perp$. Posto

$$Y_i = \frac{X_i - m}{\sigma}, \quad i = 1, \ldots, n,$$

$Y = (Y_1, \ldots, Y_n)$ è un campione di legge $\mathcal{N}(0, 1)$. Si hanno poi le relazioni

$$\overline{Y} = \frac{\sum_{i=1}^n (X_i - m)}{n\sigma} = \frac{\overline{X} - m}{\sigma}; \qquad Y_{E_1} = \langle Y, \eta\rangle \eta = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\eta = \sqrt{n}\frac{\sum_{i=1}^n Y_i}{n}\eta = \sqrt{n}\,\overline{Y}\eta. \quad (15)$$

Per la prima delle (15) si ha

$$\begin{aligned}
&\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} = \sum_{i=1}^n \Big(\frac{X_i - \overline{X}}{\sigma}\Big)^2 = \sum_{i=1}^n \Big(\frac{(X_i - m) - (\overline{X} - m)}{\sigma}\Big)^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 \\
&= \sum_{i=1}^n Y_i^2 + n\overline{Y}^2 - 2\overline{Y}\Big(\sum_{i=1}^n Y_i\Big) = \sum_{i=1}^n Y_i^2 + n\overline{Y}^2 - 2n\overline{Y}\Big(\frac{\sum_{i=1}^n Y_i}{n}\Big) = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2 \\
&= \|Y\|^2 - \|Y_{E_1}\|^2 = \|Y_{E_2}\|^2,
\end{aligned}$$

per il Teorema di Pitagora. Dunque, per il Teorema di Cochran 9.3,

$$\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} = \|Y_{E_2}\|^2$$

ha legge $\chi^2(n-1)$.

(iii) Ancora dal Teorema di Cochran sappiamo che Y_{E_1} e Y_{E_2} sono indipendenti. Dato che

$$\frac{\overline{X} - m}{\sigma}\sqrt{n} = \overline{Y}\sqrt{n} = \langle Y_{E_1}, \eta\rangle \quad \text{(funzione di } Y_{E_1}\text{)}, \qquad S^2 = \frac{\sigma^2}{n-1}\|Y_{E_2}\|^2 \quad \text{(funzione di } Y_{E_2}\text{)}$$

(dove la prima equazione segue alla seconda delle relazioni (15)), si conclude che anche $\frac{\overline{X}-m}{\sigma}\sqrt{n}$ e S^2 sono indipendenti.

(iv) Si può scrivere

$$\frac{(\overline{X} - m)\sqrt{n}}{S} = \frac{(\overline{X} - m)\sqrt{n}}{\sigma} \cdot \sqrt{\frac{\sigma^2}{S^2}} = \frac{(\overline{X} - m)\sqrt{n}}{\sigma} \cdot \sqrt{\frac{\sigma^2}{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}}} = \frac{\frac{(\overline{X}-m)\sqrt{n}}{\sigma}}{\sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}}} \cdot \sqrt{n-1},$$

che ha legge $t(n-1)$ per la definizione di legge di Student e per i tre punti precedenti di questo Teorema.

□

Osservazione 9.6 Notare la somiglianza tra le due v.a. $U := \frac{(\overline{X}-m)\sqrt{n}}{\sigma}$ e $V := \frac{(\overline{X}-m)\sqrt{n}}{S}$; la quantità σ, presente nell'espressione di U, è sostituita da S in quella di V, e quindi V, oltre che da m, dipende solo dal campione $(X_1, \ldots, X_n)$. Notiamo anche che V ha, come U, una legge nota (e che è assai simile alla legge di U per n grande). Tutto questo fa prevedere la possibilità di utilizzare V al posto di U nel caso che σ sia una quantità incognita.

Esercizio 9.7 Siano $(X_1, \dots, X_p)$ e $(Y_1, \dots, Y_q)$ due campioni gaussiani tra loro indipendenti, di leggi rispettive $\mathcal{N}(m_1, \sigma^2)$ e $\mathcal{N}(m_2, \sigma^2)$. Poniamo

$$\overline{X} = \frac{\sum_{i=1}^p X_i}{p}; \qquad S_X^2 = \frac{\sum_{i=1}^p (X_i - \overline{X})^2}{p-1}; \qquad \overline{Y} = \frac{\sum_{i=1}^q Y_i}{q}; \qquad \frac{\sum_{i=1}^q (Y_i - \overline{Y})^2}{q-1}.$$

Mostrare che

(i) la v.a. $\overline{X} - \overline{Y}$ ha legge $\mathcal{N}\Big(m_1 - m_2, \sigma^2\big(\frac{1}{p} + \frac{1}{q}\big)\Big)$;

(ii) la v.a. $S_X^2(p-1) + S_Y^2(q-1)$ ha legge $\sigma^2 \cdot \chi^2(p+q-2)$;

(iii) le v.a. $\overline{X}$, $\overline{Y}$ e

$$S_X^2(p-1) + S_Y^2(q-1) = \sum_{i=1}^p (X_i - \overline{X})^2 + \sum_{i=1}^q (Y_i - \overline{Y})^2$$

sono (globalmente) indipendenti;

(iv) la v.a.

$$\frac{\sqrt{p+q-2} \cdot \{(\overline{X} - \overline{Y}) - (m_1 - m_2)\}}{\sqrt{\frac{1}{p} + \frac{1}{q}} \cdot \sqrt{\sum_{i=1}^p (X_i - \overline{X})^2 + \sum_{i=1}^q (Y_i - \overline{Y})^2}}$$

ha legge $t(p+q-2)$.

Soluzione. Il punto (i) si ricava dall'indipendenza di $\overline{X}$ e $\overline{Y}$ e dal punto (i) del Teorema 9.5. Il punto (ii) segue dal punto (ii) del Teorema 9.5, ricordando che la legge $\chi^2(n)$ non è altro che la $\Gamma\big(\frac{n}{2}, \frac{1}{2}\big)$. Il punto (iv) segue dai punti precedenti (con lo stesso ragionamento usato per il punto (iv) del teorema 9.5. Resta da verificare il punto (i).

Per $k = 1, 2, \dots, p+q$ poniamo

$$Z_k = \begin{cases} \dfrac{X_k - m_1}{\sigma} & \text{per } 1 \le k \le p \\ \\ \dfrac{Y_{k-p} - m_2}{\sigma} & \text{per } p+1 \le k \le p+q. \end{cases}$$

Allora $Z = (z_1, \dots, Z_{p+q})$ è un campione di v.a. aventi legge $\mathcal{N}(0,1)$. Consideriamo i due versori ortogonali di $\mathbb{R}^{p+q}$

$$\eta_1 = \Big(\underbrace{\frac{1}{\sqrt{p}}, \dots, \frac{1}{\sqrt{p}}}_{p \text{ volte}}, \underbrace{0, \dots, 0}_{q \text{ volte}} \Big), \qquad \eta_2 = \Big(\underbrace{0, \dots, 0}_{p \text{ volte}}, \underbrace{\frac{1}{\sqrt{q}}, \dots, \frac{1}{\sqrt{q}}}_{q \text{ volte}} \Big);$$

siano E_1 e E_2 i sottospazi di $\mathbb{R}^{p+q}$ generati rispettivamente da η_1 e η_2 e E_3 il sottospazio ortogonale a $E_1 \oplus E_2$. Per il Teorema di Cochran 9.3, le tre proiezioni Z_{E_1}, Z_{E_2} e Z_{E_3} sono tra loro indipendenti. D'altra parte (analogamente a quanto visto nella dimostrazione del Teorema 9.5)

$$Z_{E_1} = \Big(\frac{\sum_{i=1}^p Z_i}{\sqrt{p}} \Big) \eta_1 = \sqrt{p} \Big(\frac{\overline{X} - m_1}{\sigma} \Big) \eta_1,$$

da cui

$$\frac{\overline{X} - m_1}{\sigma} = \frac{\langle Z_{E_1}, \eta_1 \rangle}{\sqrt{p}}, \qquad \overline{X} = \sigma \cdot \frac{\langle Z_{E_1}, \eta_1 \rangle}{\sqrt{p}} + m_1,$$

e in modo simile

$$Z_{E_2} = \Big(\frac{\sum_{i=p+1}^{p+q} Z_i}{\sqrt{q}}\Big)\eta_2 = \sqrt{q}\Big(\frac{\overline{Y} - m_2}{\sigma}\Big)\eta_1,$$

$$\frac{\overline{Y} - m_2}{\sigma} = \frac{\langle Z_{E_2}, \eta_2\rangle}{\sqrt{q}}, \qquad \overline{Y} = \sigma \cdot \frac{\langle Z_{E_2}, \eta_2\rangle}{\sqrt{q}} + m_2;$$

pertanto $\overline{X}$ è funzione di Z_{E_1} e $\overline{Y}$ è funzione di Z_{E_2}. Infine (riguardare i conti del Teorema 9.5)

$$S_X^2(p-1) + S_Y^2(q-1) = \sum_{i=1}^{p}(X_i - \overline{X})^2 + \sum_{i=1}^{q}(Y_i - \overline{Y})^2 = \sum_{i=1}^{p} X_i^2 - p\overline{X}^2 + \sum_{i=1}^{q} Y_i^2 - q\overline{Y}^2$$

$$= \sum_{i=1}^{p}\big(\sigma Z_i + m_1\big)^2 - p\Big(\sigma\frac{\sum_{i=1}^{p} Z_i}{p} + m_1\Big)^2 + \sum_{i=p+1}^{p+q}\big(\sigma Z_i + m_2\big)^2 - q\Big(\sigma\frac{\sum_{i=p+1}^{p+q} Z_i}{q} + m_2\Big)^2$$

$$= \sigma^2\Big\{\sum_{i=1}^{p+q} Z_i^2 - \frac{(\sum_{i=1}^{p} Z_i)^2}{p} - \frac{(\sum_{i=p+1}^{p+q} Z_i)^2}{q}\Big\} = \sigma^2\Big(\|Z\|^2 - \|Z_{E_1}\|^2 - \|Z_{E_2}\|^2\Big) = \sigma^2 \cdot \|Z_{E_3}\|^2,$$

e quindi $S_X^2(p-1) + S_Y^2(q-1)$ è funzione di Z_{E_3}. Le tre variabili $\overline{X}$, $\overline{Y}$ e

$$S_X^2(p-1) + S_Y^2(q-1)$$

sono dunque indipendenti in quanto funzioni di tre variabili indipendenti distinte.

Definizione 9.8 Si chiama *vettore aleatorio gaussiano* un vettore aleatorio $X = (X_1, \dots, X_n)$ (definito su un opportuno spazio di probabilità $(\Omega, \mathcal{F}, P)$) tale che, per ogni $u \in \mathbb{R}^n$, la v.a. $\langle u, X\rangle$ abbia legge gaussiana.

Ovviamente un campione di legge gaussiana è un vettore gaussiano. Dimostreremo che esistono vettori gaussiani che non sono campioni (altrimenti la definizione precedente sarebbe poca cosa); prima però vediamo alcune proprietà di ogni vettore gaussiano.

Indicheremo con $m = (E[X_1], \dots, E[X_n])$ il vettore delle medie delle varie componenti (cioè $m_i = E[X_i]$, $i = 1, \dots, n$) e con Γ la matrice di covarianza $\Gamma = (\Gamma_{i,j}) = \big(Cov(X_i, X_j)\big)$, $i, j = 1, \dots, n$. È facile provare che la v.a. $\langle u, X\rangle$ ha media $\langle u, m\rangle$ e varianza $\langle \Gamma u, u\rangle$:

$$E[\langle u, X\rangle] = E\Big[\sum_{i=1}^{n} u_i X_i\Big] = \sum_{i=1}^{n} u_i E\big[X_i\big] = \sum_{i=1}^{n} u_i m_i = \langle u, m\rangle;$$

$$Var\big(\langle u, X\rangle\big) = Cov\Big(\sum_{i=1}^{n} u_i X_i, \sum_{j=1}^{n} u_j X_i\Big) = \sum_{i,j=1}^{n} u_i u_j Cov(X_i, X_j) = \sum_{i,j=1}^{n} u_i u_j \Gamma_{i,j} = \langle \Gamma u, u\rangle.$$

Il seguente risultato generalizza fatti ben noti in $\mathbb{R}$:

Teorema 9.9 *(a) Sia X un vettore gaussiano a valori in $\mathbb{R}^n$. La legge di X è determinata da m e Γ, e viene denotata con $\mathcal{N}_n(m, \Gamma)$.*
(b) Se X ha legge $\mathcal{N}_n(m, \Gamma)$, A è una matrice $k \times n$ e $b \in \mathbb{R}^k$, allora il vettore $Y = AX + b$ (a valori in $\mathbb{R}^k$ ha legge $\mathcal{N}_k(Am + b, A\Gamma\,{}^tA)$.

DIMOSTRAZIONE. (a) La funzione caratteristica di X è

$$\phi_X(u) = E\big[\exp\big(\mathrm{i}\langle u, X\rangle\big)\big] = \phi_{\langle u,X\rangle}(1) = \exp\Big(\mathrm{i}\langle u, m\rangle - \frac{1}{2}\langle \Gamma u, u\rangle\Big),$$

ricordando la (15).

(b) Calcoliamo la funzione caratteristica di Y. Si ha

$$\begin{aligned}\phi_Y(u) &= E\big[\exp\big(\mathrm{i}\langle u, Y\rangle\big)\big] = E\big[\exp\big(\mathrm{i}\langle u, AX+b\rangle\big)\big] = \exp\big(\mathrm{i}\langle u, b\rangle\big)\cdot E\big[\exp\big(\mathrm{i}\langle u, AX\rangle\big)\big]\\ &= \exp\big(\mathrm{i}\langle u, b\rangle\big)\cdot E\big[\exp\big(\mathrm{i}\langle {}^tAu, X\rangle\big)\big] = \exp\big(\mathrm{i}\langle u, b\rangle\big)\cdot \exp\Big(\mathrm{i}\langle {}^tAu, m\rangle - \frac{1}{2}\langle \Gamma\, {}^tAu, {}^tAu\rangle\Big)\\ &= \exp\big(\mathrm{i}\langle u, b\rangle\big)\cdot \exp\Big(\mathrm{i}\langle u, Am\rangle - \frac{1}{2}\langle (A\Gamma\, {}^tA)u, u\rangle\Big) = \exp\Big(\mathrm{i}\langle u, Am+b\rangle - \frac{1}{2}\langle (A\Gamma\, {}^tA)u, u\rangle\Big),\end{aligned}$$

e si conclude usando il punto (a).

□

Osservazione 9.10 Sia I_n la matrice identità $n\times n$. È immediato vedere che un vettore gaussiano con legge $\mathcal{N}_n(\mathbf{0}, I_n)$ non è altro che un campione taglia n e di legge $\mathcal{N}(0,1)$.

Teorema 9.11 *(a) Fissati $m\in\mathbb{R}^n$ e Γ matrice $n\times n$ simmetrica e semidefinita positiva, esiste un vettore aleatorio avente legge $\mathcal{N}_n(m,\Gamma)$.*
(b) Se Γ è invertibile, la legge $\mathcal{N}_n(m,\Gamma)$ è assolutamente continua rispetto alla misura di Lebesgue n-dimensionale, con densità

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\Gamma|}}\exp\Big(-\frac{1}{2}\langle \Gamma^{-1}(x-m), (x-m)\rangle\Big).$$

DIMOSTRAZIONE. (a) Sia X un vettore aleatorio di legge $\mathcal{N}_n(\mathbf{0}, I_n)$, e sia $A=\sqrt{\Gamma}$ la matrice del Lemma 6.21. Posto $Y = AX+m$, per il Teorema 9.9, punto (b) , Y ha legge $\mathcal{N}_n(A\mathbf{0}+m, AI_n\,{}^tA)$= $\mathcal{N}_n(m, A^2)$= $\mathcal{N}_n(m,\Gamma)$.

(b) Poniamo $\phi(x) = Ax+m$. ϕ è invertibile e $\phi^{-1}(x) = A^{-1}(x-m)$, $\frac{\partial\phi^{-1}}{\partial x} = A^{-1}$; per la nota formula di cambio di variabili si ha (indicando con f_U la densità del vettore aleatorio U)

$$f_Y(x) = f_X\big(\phi^{-1}(x)\big)\Big|\frac{\partial\phi^{-1}}{\partial x}\Big| = \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\Big(\frac{1}{2}\|A^{-1}(x-m)\|^2\Big)\big|A^{-1}\big|.$$

Dato che $\Gamma = A^2$ si ha $|\Gamma| = |A^2| = |A|^2$ e quindi $\big|A^{-1}\big| = |A|^{-1} = \big(\sqrt{|\Gamma|}\big)^{-1}$. Inoltre

$$\begin{aligned}\big\|A^{-1}(x-m)\big\|^2 &= \langle A^{-1}(x-m), A^{-1}(x-m)\rangle = \langle {}^t(A^{-1})A^{-1}(x-m), (x-m)\rangle\\ &= \langle (A\,{}^tA)^{-1}(x-m), (x-m)\rangle = \langle (A^2)^{-1}(x-m), (x-m)\rangle = \langle \Gamma^{-1}(x-m), (x-m)\rangle.\end{aligned}$$

sostituendo si trova la formula dell'enunciato.

□

Osservazione 9.12 In realtà il punto (b) ammette anche il viceversa (che non dimostriamo): se Γ non è invertibile, allora la legge $\mathcal{N}_n(m,\Gamma)$ non è assolutamente continua rispetto alla misura di Lebesgue n-dimensionale.

10 I modelli lineari

Esempio 10.1 (introduttivo). L'esito di un certo fenomeno aleatorio (ad esempio il rendimento di un certo terreno coltivato) è la somma di una funzione $f(x)$ (dove f è sconosciuta) di una certa quantità (non aleatoria) x (ad esempio x = quantità di concime impiegato) e di un "disturbo" aleatorio W (ad esempio W = quantità di pioggia caduta nel periodo di osservazione).

Per valutare ragionevolmente f si fanno n prove (ad esempio si concimano n terreni nello stesso modo), e in questa maniera si ottengono n risultati

$$\begin{cases} Y_1 = f(x_1) + W_1 \\ Y_2 = f(x_2) + W_2 \\ \vdots \\ Y_n = f(x_n) + W_n; \end{cases}$$

Inoltre si suppone che le v.a. W_m, $m = 1, \ldots, n$ siano centrate, non correlate e con varianza σ^2, anch'essa sconosciuta.

Lo scopo dell'indagine è quello di ottenere informazioni sulle quantità non note, e cioè su f e σ^2. Per rendere il problema matematicamente trattabile, si può approssimare f con un polinomio di grado $k-1$, dove $k < n$:

$$f(x) \approx \sum_{j=1}^{k} x^{j-1}\theta_j;$$

quindi le equazioni precedenti diventano (scriviamo il sistema in forma compatta)

$$Y_m = \sum_{j=1}^{k} (x_m)^{j-1}\theta_j + W_m, \qquad m = 1, 2, \ldots, n, \qquad k < n.$$

Le v.a. W_m possono essere scritte nella forma $W_m = \sigma Z_m$, dove le Z_m sono centrate e tali che $Cov(Z_i, Z_j) = \delta_{i,j}$. In definitiva il modello è diventato

$$Y_m = \sum_{j=1}^{k} (x_m)^{j-1}\theta_j + \sigma Z_m, \qquad m = 1, 2, \ldots, n, \qquad k < n,$$

e a questo punto ottenere informazioni su f e σ significa stimare i numeri $\theta_1, \ldots, \theta_k$ e σ^2.

I modelli come quello qui indicato sono detti *modelli di regressione*, e sono un caso particolare dei cosiddetti *modelli lineari*, di cui diamo ora la definizione.

Definizione 10.2 Si chiama *modello lineare* un modello statistico nel quale l'osservazione è formata da n v.a. della forma

$$Y_m = \sum_{j=1}^{k} a_{m,j}\theta_j + \sigma Z_m, \qquad m = 1, 2, \ldots, n, \qquad k < n,$$

dove le incognite sono $\theta_1, \ldots, \theta_k$ e σ^2.

Osservazione 10.3 *da Wikipedia:*"L'origine del termine *regressione* è storicamente documentata. L'espressione *reversione* era usata nel XIX secolo per descrivere un fenomeno biologico, in base al quale la progenie di individui eccezionali tende in media a presentare caratteristiche meno notevoli di quelle dei genitori, e più simili a quelle degli antenati più remoti. Francis Galton studiò tale fenomeno, applicandovi il termine, forse improprio, di *regressione verso la media* (o la mediocrità)."

Come abbiamo detto, si suppone che il vettore aleatorio $Z = {}^t(Z_1, \dots, Z_n)$ sia centrato e con matrice di covarianza I_n. Inoltre il modello può essere rappresentato in forma vettoriale come

$$Y = A\theta + \sigma Z,$$

dove $\theta = {}^t(\theta_1, \dots, \theta_k)$, $A = \big(a_{m,j}\big)_{m\le n, j\le k}$, $Z = {}^t(Z_1, \dots, Z_n)$. Si suppone che l'applicazione $L_A : \mathbb{R}^k \to \mathbb{R}^n$ definita da $L_A(x) = Ax$ sia iniettiva, cioè che la matrice $n \times k$ A sia di rango massimo ($= \min\{k, n\} = k$).

Volendo rappresentare la situazione come un modello statistico (nel senso usuale), si prende

$$\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{P^{\theta,\sigma^2}\}\big),$$

dove P^{θ,σ^2} è la legge di Y, cioè l'immagine della legge di Z secondo l'applicazione $z \mapsto A\theta + \sigma z$.

Diremo che siamo nel *caso gaussiano* se la legge di Z è la $\mathcal{N}_n(\mathbf{0}, I_n)$ (cioè se $Z = (Z_1, \dots, Z_n)$ è un campione di legge $\mathcal{N}(0,1)$). In tal caso la legge P^{θ,σ^2} ammette densità rispetto alla misura di Lebesgue su $\mathbb{R}^n$, data da

$$f^{\theta,\sigma^2}(y) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\Big(-\frac{1}{2\sigma^2}\|y - A\theta\|^2\Big).$$

Si ha

$$\|y - A\theta\|^2 = \langle y - A\theta, y - A\theta\rangle = \|y\|^2 + \|A\theta\|^2 - 2\langle y, A\theta\rangle;$$

indichiamo con $L_A(\mathbb{R}^k)$ il sottospazio chiuso di $\mathbb{R}^n$ immagine di $\mathbb{R}^k$ secondo l'applicazione L_A e con ¶ la proiezione di $\mathbb{R}^n$ su $L_A(\mathbb{R}^k)$. Allora si ha $y = ¶y + z$, con z vettore ortogonale a $L_A(\mathbb{R}^k)$, e quindi

$$\langle y, A\theta\rangle = \langle ¶y + z, A\theta\rangle = \langle ¶y, A\theta\rangle + \langle z, A\theta\rangle = \langle ¶y, A\theta\rangle.$$

Ne segue che

$$f^{\theta,\sigma^2}(y) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big(-\frac{1}{2\sigma^2}\Big\{\|y\|^2 + \|A\theta\|^2 - 2\langle ¶y, A\theta\rangle\Big\} - n\log\sigma\Big);$$

dato che

$$\Big\langle \big(\frac{A\theta}{\sigma^2}, -\frac{1}{2\sigma^2}\big), (¶Y, \|Y\|^2)\Big\rangle = \frac{\langle ¶y, A\theta\rangle}{\sigma^2} - \frac{\|Y\|^2}{2\sigma^2},$$

si riconosce che siamo in presenza di un modello esponenziale e che $(¶Y, \|Y\|^2)$ è una statistica esaustiva completa. In particolare $¶Y$ è uno stimatore di θ (osservare che ¶ manda $\mathbb{R}^n$ in $L_A(\mathbb{R}^k)$, che è un sottospazio di $\mathbb{R}^n$ a dimensione k (= numero dei θ_i), in quanto A è di rango massimo).

In analogia al caso gaussiano, in un modello lineare considereremo solo degli stimatori lineari di θ, cioè del tipo VY, con V matrice $k \times n$ (e di conseguenza $L_V : \mathbb{R}^n \to \mathbb{R}^k$).

Cominciamo con un lemma.

Lemma 10.4 *Siano A una matrice $n \times k$ con $k < n$ di rango massimo ($= k$) e $y \in \mathbb{R}^n$. Allora il vettore $x \in \mathbb{R}^k$ che rende minima la funzione $x \mapsto \|y - Ax\|^2$ è dato da*

$$x = \mathcal{U}y, \qquad \textit{dove} \quad \mathcal{U} = \big({}^tAA\big)^{-1}\big({}^tA\big).$$

DIMOSTRAZIONE. È noto che il vettore $z \in L_A(\mathbb{R}^k)$ (cioè del tipo $z = Ax$ per qualche $x \in \mathbb{R}^k$) che rende minima la distanza di y da $L_A(\mathbb{R}^k)$ è $z = \P y$, dove $\P$ è la proiezione ortogonale di $\mathbb{R}^n$ su $L_A(\mathbb{R}^k)$. Dunque dovrà essere

$$Ax = z = \P y. \tag{16}$$

Questa relazione implica che x è unico (infatti, se fosse $\P y = Ax_1 = Ax_2$, allora $x_1 = x_2$ perché A è di rango massimo). Dunque possiamo definire l'applicazione $\mathcal{U} : \mathbb{R}^n \to \mathbb{R}^k$ tale che il vettore x si possa mettere nella forma $x = \mathcal{U}y$. Dalla (16) segue allora che $A\mathcal{U}y = \P y$, per ogni $y \in \mathbb{R}^n$. In altre parole si ha

$$A\mathcal{U} = \P. \tag{17}$$

Poiché A non è una matrice quadrata, non possiamo invertire la relazione precedente scrivendo $\mathcal{U} = A^{-1}\P$; dunque, per identificare $\mathcal{U}$ dobbiamo fare ulteriori passaggi.

Cominciamo osservando che ${}^t\P = \P$ (infatti, siano u e v due vettori di $\mathbb{R}^n$. Allora $u = \P u + a$, dove a è ortogonale a $\in L_A(\mathbb{R}^k)$, e quindi $\langle u, \P v\rangle = \langle \P u + a, \P v\rangle = \langle \P u, \P v\rangle$. Analogamente si ha $\langle \P u, v\rangle = \langle \P u, \P v\rangle$, e quindi $\langle u, \P v\rangle = \langle \P u, v\rangle$). Dato che ovviamente $A = \P A$, abbiamo, per la (17),

$${}^tA = {}^t(\P A) = {}^tA\,{}^t\P = {}^tA\P = {}^tA(A\mathcal{U}) = ({}^tAA)\mathcal{U}.$$

Da questa relazione segue la tesi, a patto che ${}^tAA$ sia invertibile. Per dimostrarlo, osserviamo prima di tutto che $L_{{}^tAA} : \mathbb{R}^k \to \mathbb{R}^k$; dunque ${}^tAA$ è invertibile se e solo se $L_{{}^tAA}$ è iniettiva, ovvero se e solo se la relazione $({}^tAA)x = 0$ implica che $x = 0$. Ma se $({}^tAA)x = 0$, allora

$$0 = \langle ({}^tAA)x, x\rangle = \langle Ax, Ax\rangle = \|Ax\|^2,$$

da cui segue che $Ax = 0$ e quindi $x = 0$ perché L_A è iniettiva (essendo A di rango massimo).

□

Diremo che uno stimatore *vettoriale* è *corretto* se è corretto componente per componente.

Teorema 10.5 *(*DI GAUSS-MARKOV*). Lo stimatore $\mathcal{U}Y$ è uno stimatore corretto di θ, ottimale nella classe degli stimatori lineari corretti di θ, e $\|A\mathcal{U}Y - Y\|^2$ è uno stimatore corretto di $(n-k)\sigma^2$. Nel caso gaussiano, tali stimatori sono ottimali tra tutti gli stimatori corretti (lineari e non, per quanto riguarda θ) di θ e $(n-k)\sigma^2$ rispettivamente.*

DIMOSTRAZIONE. Sia VY uno stimatore lineare. VY è uno stimatore corretto di θ se e solo se

$$\theta = E^{\theta,\sigma^2}[VY] = E^{\theta,\sigma^2}[V(A\theta + \sigma Z)] = VA\theta + \sigma V E^{\theta,\sigma^2}[Z] = VA\theta,$$

perché Z è centrato. La relazione precedente dice che, se VY è corretto, allora deve aversi

$$VA = I_k, \tag{18}$$

ed in effetti $\mathcal{U}$ ha questa proprietà. Infatti, per il Lemma 10.4, risulta

$$\mathcal{U}A = \big\{\big({}^tAA\big)^{-1}\big({}^tA\big)\big\}A = \big({}^tAA\big)^{-1}\big({}^tAA\big) = I_k.$$

Quindi $\mathcal{U}Y$ è uno stimatore corretto. Valutiamo adesso il rischio di un generico stimatore corretto VY. Esso è uguale a

$$E^{\theta,\sigma^2}\big[\|VY - \theta\|^2\big] = E^{\theta,\sigma^2}\big[\|V(A\theta + \sigma Z) - \theta\|^2\big] = E^{\theta,\sigma^2}\big[\|V(A\theta) + \sigma VZ - \theta\|^2\big] = \sigma^2 E^{\theta,\sigma^2}\big[\|VZ\|^2\big],$$

per la (18). D'altra parte

$$\|VZ\|^2 = \sum_i \Big(\sum_j v_{i,j}Z_j\Big)^2 = \sum_i \Big(\sum_j v_{i,j}Z_j\Big)\Big(\sum_k v_{i,k}Z_k\Big) = \sum_i \Big(\sum_{j,k} v_{i,j}v_{i,k}Z_jZ_k\Big),$$

e, passando alla speranza,

$$E^{\theta,\sigma^2}\big[\|VZ\|^2\big] = \sum_i \Big(\sum_{j,k} v_{i,j}v_{i,k}E^{\theta,\sigma^2}[Z_jZ_k]\Big) = \sum_i \Big(\sum_{j,k} v_{i,j}v_{i,k}\delta_{j,k}\Big) = \sum_i \Big(\sum_j v_{i,j}^2\Big)$$
$$= \sum_{i,j} v_{i,j}^2 = \|V\|^2,$$

per cui si ottiene infine

$$E^{\theta,\sigma^2}\big[\|VY-\theta\|^2\big] = \sigma^2\|V\|^2.$$

Dunque, volendo trovare lo stimatore ottimale (cioè di rischio minimo) nella classe degli stimatori lineari corretti, occorre minimizare la quantità $\|V\|^2$ sotto la condizione (18). Mostriamo che lo stimatore $\mathcal{U}Y$ è appunto quello che minimizza tale quantità. Osserviamo che per ogni V tale che $VA = I_k$, dato che $A\mathcal{U} = \P$ (ved. equazione (17)) si ha

$$\mathcal{U} = I_k\mathcal{U} = (VA)\mathcal{U} = V(A\mathcal{U}) = V\P.$$

Di conseguenza

$${}^t\mathcal{U} = {}^t(V\P) = {}^t\P\,{}^tV = \P\,{}^tV,$$

da cui

$$\|\mathcal{U}\|^2 = \|\,{}^t\mathcal{U}\|^2 = \|\P\,{}^tV\|^2 \le \|\,{}^tV\|^2 = \|V\|^2,$$

poiché la proiezione diminuisce la norma.

Passiamo a considerare lo stimatore $\|A\mathcal{U}Y - Y\|^2$. Si ha, per la (17),

$$A\mathcal{U}Y - Y = \P Y - Y = \P(A\theta+\sigma Z) - (A\theta+\sigma Z) = (\P A\theta - A\theta) + \sigma(\P Z - Z) = \sigma(\P Z - Z),$$

ricordando che $\P A = A$. Quindi

$$E^{\theta,\sigma^2}\big[\|A\mathcal{U}Y-Y\|^2\big] = \sigma^2 E^{\theta,\sigma^2}\big[\|\P Z - Z\|^2\big].$$

Per dimostrare che $\|A\mathcal{U}Y-Y\|^2$ è uno stimatore corretto di $\sigma^2(n-k)$, bisogna allora far vedere che

$$E^{\theta,\sigma^2}\big[\|\P Z - Z\|^2\big] = n-k.$$

Studiamo dapprima il caso particolare in cui $\P$ è la proiezione sul sottospazio generato dalle prime k coordinate. Allora abbiamo

$$\P Z - Z = (\underbrace{0,\dots,0}_{k\,volte}, -Z_{k+1},\dots,-Z_n),$$

e la tesi è ovvia perché

$$E^{\theta,\sigma^2}\big[\|\P Z - Z\|^2\big] = E^{\theta,\sigma^2}\big[Z_{k+1}^2 + \cdots + Z_n^2\big] = n-k,$$

ricordando che $E^{\theta,\sigma^2}[Z_i^2] = \delta_{i,i} = 1$.

Passiamo ora al caso generale. Premettiamo un

Lemma 10.6 *Siano $B = (b_{i,j})_{i,j=1,\dots,n}$ una matrice $n\times n$ ortogonale e Z un vettore aleatorio, definito su $(\Omega,\mathcal{F},P)$ e a valori in $\mathbb{R}^n$, centrato e con matrice di covarianza $\Gamma_Z = \big(\gamma_{i,j}\big)_{i,j=1,\dots,n}$. Allora la matrice di covarianza del vettore $W = BZ$ è data da*

$$\Gamma_W = B\Gamma_Z\,{}^tB.$$

Dimostrazione (del Lemma). Risulta

$$E[W_iW_j] = E\Big[\Big(\sum_h b_{i,h}Z_h\Big)\Big(\sum_k b_{j,k}Z_k\Big)\Big] = E\Big[\sum_{h,k} b_{i,h}b_{j,k}Z_hZ_k\Big] = \sum_{h,k} b_{i,h}b_{j,k}E\big[Z_hZ_k\big]$$
$$= \sum_{h,k} b_{i,h}b_{j,k}\gamma_{h,k} = \sum_{h,k} b_{i,h}\gamma_{h,k}\tilde{b}_{k,j},$$

dove $\tilde{b}_{k,j} = b_{j,k}$. Si riconosce che ${}^tB = \big(\tilde{b}_{k,j}\big) = \big(b_{j,k}\big)$, e quindi abbiamo ottenuto

$$E[W_iW_j] = (B\Gamma_Z\,{}^tB)_{i,j}.$$

□

Applicando il Lemma precedente nel nostro caso, in cui $\Gamma_Z = I_n$, otteniamo, per una data matrice B ortogonale,

$$\Gamma_W = B\Gamma_Z\,{}^tB = BI_n\,{}^tB = B\,{}^tB = I_n,$$

cioè il vettore $W = BZ$ ha le stesse proprietà di Z. Dunque ci si riconduce al caso particolare precedente se consideriamo la matrice ortogonale B che cambia la base in modo che i primi k elementi generino il sottospazio (a dimensione k) $L_A(\mathbb{R}^k)$ su cui ¶ proietta.

Se siamo nel caso gaussiano, allora $(\P Y, \|Y\|^2)$ è una statistica esaustiva completa, come abbiamo visto prima. Dunque, per i risultati generali sulle statistiche esaustive complete, per vedere che $\mathcal{U}Y$ e $\|A\mathcal{U}Y - Y\|^2$ sono stimatori ottimali, basta far vedere che sono funzioni della statistica esaustiva completa. Osserviamo prima di tutto che

$$\mathcal{U} = \big({}^tAA\big)^{-1}\big({}^tA\big) = \big({}^tAA\big)^{-1}{}^t\big(\P A\big) = \big({}^tAA\big)^{-1}\big({}^tA\big)\,{}^t\P = \big\{\big({}^tAA\big)^{-1}\big({}^tA\big)\big\}\P = \mathcal{U}\P;$$

(lo abbiamo anche già dimostrato sopra, perché abbiamo visto che per ogni V tale che $VA = I_k$, abbiamo $\mathcal{U} = V\P$).
Dunque $\mathcal{U}Y = \mathcal{U}(\P Y)$ è funzione di $\P Y$. Inoltre, per il Teorema di Pitagora,

$$\|Y\|^2 = \|\P Y\|^2 + \|Y - \P Y\|^2,$$

da cui, ricordando la relazione (17),

$$\|A\mathcal{U}Y - Y\|^2 = \|\P Y - Y\|^2 = \|Y\|^2 - \|\P Y\|^2,$$

e quindi anche $\|A\mathcal{U}Y - Y\|^2$ è funzione della statistica completa.

□

Osservazione 10.7 Una volta effettuato l'esperimento, che ha prodotto il risultato ω, per il calcolo effettivo di $\mathcal{U}Y(\omega)$ si procede così. Noi abbiamo a disposizione l'osservazione $Y(\omega)$, e sappiamo dal Teorema di Gauss-Markov 10.5 che la stima lineare di θ deve essere $\mathcal{U}Y(\omega)$. Da Lemma 10.4 sappiamo che $\mathcal{U}Y(\omega)$ minimizza $\|Y(\omega) - A(\mathcal{U}Y(\omega))\|^2$. Dunque si cerca il vettore $\theta_0 \in \mathbb{R}^k$ in cui si realizza il minimo della funzione

$$\theta \mapsto \|Y(\omega) - A(\mathcal{U}Y(\omega))\|^2 = \sum_{m=1}^{n}\Big(Y_m(\omega) - \sum_{j=1}^{k} a_{m,j}\theta_j\Big)^2,$$

e si pone $\mathcal{U}Y(\omega) = \theta_0$. In casi come questo si usa dire che si stima θ nel senso dei *minimi quadrati*. In pratica si annullano le derivate parziali rispetto a $\theta_1, \ldots, \theta_k$

Esempio 10.8 Torniamo al modello di regressione che abbiamo visto all'inizio, e cioè

$$Y_m = \sum_{j=1}^{k} (x_m)^{j-1}\theta_j + \sigma Z_m, \qquad m = 1, 2, \ldots, n,$$

con $k < n$ e $x_1 \neq x_2 \neq \cdots \neq x_n$. Allora è ben noto che la matrice A data da

$$\begin{pmatrix} 1 & x_1 & \ldots & x_1^{k-1} \\ 1 & x_2 & \ldots & x_2^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^{k-1} \end{pmatrix}$$

ha rango massimo. Annullando le derivate parziali si arriva alle equazioni

$$d_i = \sum_{j=1}^{k} c_{i,j}\theta_j,$$

dove

$$d_i = \sum_{m=1}^{n} Y_m(\omega) x_m^{i-1}, \qquad c_{i,j} = \sum_{m=1}^{n} x_m^{i+j-2}.$$

11 Cenni sulle regioni di fiducia

Premettiamo la definizione di *quantile*.

Sia F una funzione di ripartizione. Per ogni $\alpha \in (0,1)$, poniamo

$$S_\alpha = \big\{x \in \mathbb{R} : F(x) \geq \alpha\big\}, \qquad F^{\leftarrow}(\alpha) = \inf S_\alpha = \inf\big\{x \in \mathbb{R} : F(x) \geq \alpha\big\}.$$

Notiamo che, grazie alla non decrescenza di F, S_α è una semiretta destra in $\mathbb{R}$, di cui dunque $F^{\leftarrow}(\alpha)$ è l'origine.

Definizione 11.1 La funzione $F^{\leftarrow} : (0,1) \to \mathbb{R}$ così definita si chiama *inversa generalizzata* di F. Tuttavia in Statistica si usa chiamarla preferibilmente *funzione quantile* di F (meglio: della legge di cui F è la funzione di ripartizione). Se $\alpha \in (0,1)$, il numero $F^{\leftarrow}(\alpha)$ si chiama *quantile di ordine* α *di* F.

Osservazione 11.2 È immediato verificare che, se F è continua e strettamente crescente nell'insieme $H_F = \{x \in \mathbb{R} : 0 < F(x) < 1\}$, allora $F^{\leftarrow}$ non è altro che la funzione inversa F^{-1} di F; in questo caso si hanno quindi le relazioni

$$F\big(F^{\leftarrow}(\alpha)\big) = \alpha, \quad \forall\, \alpha \in (0,1); \qquad F^{\leftarrow}\big(F(x)\big) = x, \quad \forall\, x \in H_F.$$

La seguente Proposizione elenca le principali proprietà di $F^{\leftarrow}$:

Proposition 11.3 *Valgono le seguenti proprietà:*

(i) $F^{\leftarrow}$ *è non decrescente.*

(ii) $F(x_1) < \alpha \leq F(x_2) \iff x_1 < F^{\leftarrow}(\alpha) \leq x_2$.

(iii) $F\big(F^{\leftarrow}(\alpha)\big) \geq \alpha$ *per ogni* $\alpha \in (0,1)$. *In più, se* F *è continua, la diseguaglianza è un'uguaglianza.*

(iv) $F^{\leftarrow}\big(F(x)\big) \leq x$ *per ogni* $x \in H_F$. *In più, se* F *è strettamente crescente, la diseguaglianza è un'uguaglianza.*

(v) F *è continua se e solo se* $F^{\leftarrow}$ *è strettamente crescente.*

(vi) F *è strettamente crescente se e solo se* $F^{\leftarrow}$ *è continua .*

(vii) Se la v.a. X *ha funzione di ripartizione* F, *allora* $P\big(F^{\leftarrow}\big(F(X)\big) \neq X\big) = 0$.

Si lascia la dimostrazione per esercizio; tuttavia, dato che saranno usate nel seguito (nella dimostrazione del Teorema 23.3), a titolo di esempio mostriamo direttamente le due proprità seguenti:

(1) Se $\alpha \leq \beta$, allora $F^{\leftarrow}(\alpha) \leq F^{\leftarrow}(\beta)$ (questa non è altro che la (i) della Proposizione precedente)
Infatti in tal caso $S_\beta \subseteq S_\alpha$ e quindi $F^{\leftarrow}(\alpha) = \inf S_\alpha \leq \inf S_\beta = F^{\leftarrow}(\beta)$.

(2) $F\big(F^{\leftarrow}(\alpha)^-\big) \leq \alpha \leq F\big(F^{\leftarrow}(\alpha)\big)$.
Infatti

(a) Sia $t < F^{\leftarrow}(\alpha)$. Allora $t \notin S_\alpha$, e quindi $F(t) < \alpha$. Passando al limite per $t \uparrow F^{\leftarrow}(\alpha)$, si ha la prima delle diseguaglianze da dimostrare.
(b) Per definizione di estremo inferiore, esiste una successione (x_n) di elementi di S_α tale che $x_n \downarrow F^{\leftarrow}(\alpha)$. Dato che $x_n \in S_\alpha$, si ha $\alpha \leq F(x_n)$ e, passando al limite in n, si trova $\alpha \leq \lim_{n\to\infty} F(x_n) = \lim_{t\downarrow F^{\leftarrow}(\alpha)} F(t) = F\big(F^{\leftarrow}(\alpha)\big)$, per la continuità a destra di F.

Utilizzeremo simboli particolari per i quantili (di ordine α) delle tre leggi fondamentali dell Statistica, e cioè

- i quantili della $\mathcal{N}(0,1)$ saranno indicati con ϕ_α;
- i quantili della $t(n)$ saranno indicati con $t_\alpha(n)$;
- i quantili della $\chi^2(n)$ saranno indicati con $\chi^2_\alpha(n)$.

Osservazione 11.4 Sia F la funzione di ripartizione continua, strettamente crescente e simmetrica, cioè tale che $F(-x) = 1 - F(x)$, $\forall\, x \in \mathbb{R}$. Allora

$$F^{-1}(\alpha) = -F^{-1}(1-\alpha), \qquad \forall\, \alpha \in (0,1). \tag{19}$$

(In realtà si potrebbe dimostrare che, se F è simmetrica e continua, allora $F^{\leftarrow}(\alpha) = -F^{\leftarrow}(1-\alpha)$).

Per dimostrare la relazione (19), dato che F è iniettiva, basta verificare che

$$F\big(F^{-1}(\alpha)\big) = F\big(-F^{-1}(1-\alpha)\big).$$

Infatti

$$F\big(F^{-1}(\alpha)\big) = \alpha;$$
$$F\big(-F^{-1}(1-\alpha)\big) = 1 - F\big(F^{-1}(1-\alpha)\big) = 1 - (1-\alpha) = \alpha.$$

La proprietà (19) vale in particolare per i quantili della $\mathcal{N}(0,1)$ e della $t(n)$, per cui si hanno le formule

$$\phi_\alpha = -\phi_{1-\alpha}; \qquad t_\alpha(n) = -t_{1-\alpha}(n). \tag{20}$$

Mentre nei problemi di stima puntuale studiati nel §3 l'obbiettivo è quello di identificare il parametro che regola un certo fenomeno aleatorio, nella teoria delle *regioni di fiducia* si cerca di ottenere, in base al risultato dell'esperimento, un sottoinsieme (aleatorio) di Θ a cui il parametro appartiene con una probabilità abbastanza alta. Più precisamente si dà la seguente

Definizione 11.5 Siano $\alpha \in (0,1)$, $g : \Theta \to \mathbb{R}$ una funzione e $S : \Omega \to \mathcal{P}(\Theta)$ una funzione tale che, per ogni θ, l'insieme $\{\omega : S(\omega) \text{ non contiene } g(\theta)\}$ (che si scrive preferibilmente, anche se in modo un po' improprio, nella forma $\{g(\theta) \notin S\}$) appartenga alla σ- algebra $\mathcal{F}$. Si dice che S è una *regione di fiducia* (o *di confidenza*) *di livello* $1-\alpha$ per $g(\theta)$ se, per ogni θ, risulta

$$P^\theta(g(\theta) \notin S) \le \alpha, \quad \text{o, equivalentemente,} \quad P^\theta(g(\theta) \in S) \ge 1-\alpha.$$

Osservazione 11.6 Naturalmente, i valori tipici per α sono piccoli, ad esempio $\alpha = 0,05$ oppure $\alpha = 0,01$.

Osservazione 11.7 Se $\Theta \subseteq \mathbb{R}$, le regioni di fiducia che si costruiscono nella pratica sono in genere degli intervalli della retta (limitati o no).

Esempio 11.8 Costruire un intervallo di fiducia di livello 0,95 per il parametro $\theta > 0$ dell'esponenziale, basato su una sola osservazione X (cioè il modello è quello di un campione unidimensionale X).

Significa che si devono trovare due funzioni $T_1(X)$ e $T_2(X)$ tali che, per ogni $\theta > 0$, risulti

$$P^\theta\big(T_1(X) \le \theta \le T_2(X)\big) \ge 0,95.$$

Partiamo da questa semplice osservazione: se X ha legge $\mathcal{E}(\theta)$, allora la v.a. $Q = \theta X$ ha legge $\mathcal{E}(1)$. Infatti, per ogni $t > 0$ si ha

$$P^\theta(Q \le t) = P^\theta(\theta X \le t) = P^\theta\Big(X \le \frac{t}{\theta}\Big) = 1 - \mathrm{e}^{-t}.$$

Di conseguenza, per ogni coppia $a, b > 0$, con $a < b$ si ottiene

$$P^\theta(a \le Q \le b) = (1 - \mathrm{e}^{-b}) - (1 - \mathrm{e}^{-a}) = \mathrm{e}^{-a} - \mathrm{e}^{-b},$$

il che equivale a

$$P^\theta\Big(\frac{a}{X} \le \theta \le \frac{b}{X}\Big) = \mathrm{e}^{-a} - \mathrm{e}^{-b}.$$

Allora poniamo

$$T_1(X) = \frac{a}{X}, \qquad T_2(X) = \frac{b}{X},$$

dove le costanti a e b sono scelte in modo che valga l'uguaglianza $\mathrm{e}^{-a} - \mathrm{e}^{-b} = 0,95$.

Osservazione 11.9 Il metodo qui seguito, e che verrà usato sistematicamente nel seguito, è quello cosiddetto *della quantità pivotale*, che consiste nel determinare una funzione di $X_1, \dots, X_n$ e del parametro θ, $Q(X_1, \dots, X_n; \theta)$ (X e θ nell'esempio), invertibile rispetto alla variabile θ, in modo che la sua legge $(A \mapsto P^\theta\big(Q(X_1, \dots, X_n; \theta) \in A\big))$ non dipenda da θ.

Nota. Nei paragrafi seguenti $X_1, \dots, X_n$ sarà un campione di taglia n e di legge $\mathcal{N}(m, \sigma^2)$; inoltre gli intervalli che costruiremo saranno tutti di livello $1-\alpha$ fissato. Indicheremo con Φ la f.d.r. della $\mathcal{N}(0,1)$. Ricordiamo anche che in questo caso si ha $\phi_\alpha = \Phi^{-1}(\alpha)$.

Intervalli di fiducia per la media della Normale con varianza nota.

Si parte dall'osservazione che la v.a.

$$Q = Q(X_1, \dots, X_n; m) = \frac{\overline{X} - m}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1),$$

come ci assicura il Teorema 9.5 (i). Pertanto Q può essere scelta come quantità pivotale.

(a) Intervallo bilaterale. Dalla relazione

$$P^m(a \le Q \le b) = \Phi(b) - \Phi(a)$$

si ricava

$$\begin{aligned}&\Phi(b) - \Phi(a)\\ &= P^m\Big(a \le \frac{\overline{X} - m}{\sigma}\sqrt{n} \le b\Big) = P^m\Big(\overline{X} - \frac{\sigma}{\sqrt{n}}b \le m \le \overline{X} - \frac{\sigma}{\sqrt{n}}a\Big).\end{aligned}$$

Pertanto basterà trovare a e b in modo che

$$\Phi(b) - \Phi(a) = 1 - \alpha.$$

Siano β e γ due numeri reali $\in (0,1)$ tali che $b = \phi_\beta$ e $a = \phi_\gamma$; allora si ha

$$\Phi(b) - \Phi(a) = \beta - \gamma,$$

e quindi basterà scegliere β e γ in modo che risulti $\beta - \gamma = 1 - \alpha$. Una scelta possibile è $\beta = 1 - \frac{\alpha}{2}$, $\gamma = \frac{\alpha}{2}$, cioè

$$b = \phi_{1-\frac{\alpha}{2}}, \qquad a = \phi_{\frac{\alpha}{2}} = -\phi_{1-\frac{\alpha}{2}},$$

dove l'ultima uguaglianza segue dalla prima delle (20). L'intervallo risultante è allora

$$\Big(\overline{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \le m \le \overline{X} + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\Big). \tag{21}$$

Osservazione 11.10 La scelta fatta per β e γ non è ovviamente l'unica possibile: ad esempio $\beta = 1 - \frac{\alpha}{3}$, $\gamma = \frac{2}{3}\alpha$ va ancora bene. Tuttavia l'intervallo (21) è quello di ampiezza minima (cioè dà la stima migliore possibile, al livello $1 - \alpha$ assegnato), come mostra l'esercizio che segue.

Esercizio 11.11 . Calcolare il

$$\begin{cases}\min(b - a)\\ \Phi(b) - \Phi(a) = 1 - \alpha\end{cases}$$

Soluzione. Poniamo $\Phi(b) = \beta$, $\Phi(a) = \gamma$. Il problema si può allora riformulare nel modo seguente:

$$\begin{cases}\min(\Phi^{-1}(\beta) - (\Phi^{-1}(\gamma))\\ \beta - \gamma = 1 - \alpha\end{cases}$$

Porremo per semplicità $k = 1 - \alpha$. Dunque si tratta di trovare il minimo della funzione

$$\gamma \mapsto \Phi^{-1}(\gamma + k) - \Phi^{-1}(\gamma). \tag{22}$$

La derivata è

$$\frac{1}{\Phi'(\Phi^{-1}(\gamma + k))} - \frac{1}{\Phi'(\Phi^{-1}(\gamma))},$$

ed è ≥ 0 per

$$\Phi'(\Phi^{-1}(\gamma + k)) \le \Phi'(\Phi^{-1}(\gamma)),$$

cioè

$$\exp\Big(-\frac{\{\Phi^{-1}(\gamma+k)\}^2}{2}\Big) \le \exp\Big(-\frac{\{\Phi^{-1}(\gamma)\}^2}{2}\Big).$$

La disequazione

$$\exp\Big(-\frac{m^2}{2}\Big) \le \exp\Big(-\frac{n^2}{2}\Big)$$

ha le soluzioni $m^2 \ge n^2$, ovvero

$$-|m| \le n \le |m|.$$

Nel nostro caso ciò significa

$$-|\Phi^{-1}(\gamma+k)| \le \Phi^{-1}(\gamma) \le |\Phi^{-1}(\gamma+k)|$$

Dunque, applicando la Φ ai tre membri di queste diseguaglianze, si trova

$$\Phi\big(-|\Phi^{-1}(\gamma+k)|\big) \le \Phi\big(\Phi^{-1}(\gamma)\big) \le \Phi\big(|\Phi^{-1}(\gamma+k)|\big). \tag{23}$$

Non è difficile vedere che, in generale, si ha

$$\Phi\big(|\Phi^{-1}(u)|\big) = \frac{1}{2} + \Big|u - \frac{1}{2}\Big|,$$

(verifica per esercizio); dunque la (23) diventa (tenuto conto dell'identità $\Phi(-t) = 1 - \Phi(t)$)

$$1 - \frac{1}{2} - \Big|\gamma + k - \frac{1}{2}\Big| \le \gamma \le \frac{1}{2} + \Big|\gamma + k - \frac{1}{2}\Big|,$$

e, sistemando i calcoli, si trova

$$-\Big|\gamma + k - \frac{1}{2}\Big| \le \gamma - \frac{1}{2} \le \Big|\gamma + k - \frac{1}{2}\Big|$$

ovvero

$$\Big(\gamma - \frac{1}{2}\Big)^2 \le \Big(\gamma - \frac{1}{2} + k\Big)^2$$

La soluzione di questa disequazione è

$$\gamma \ge \frac{1-k}{2} = \frac{\alpha}{2}.$$

Ricapitolando, abbiamo trovato che la funzione (22) è crescente (risp. decrescente) per $\gamma \ge \frac{\alpha}{2}$ (risp. *gamma* $\ge \frac{\alpha}{2}$), e quindi $\gamma = \frac{\alpha}{2}$ è punto di minimo.

(b) Intervallo Unilaterale Destro. Il termine significa che si vuole trovare una limitazione per m solo dal basso, cioè del tipo $h < m$ (il termine "destro" si spiega osservando che in tal caso $m \in (h, +\infty)$, semiretta destra).

Questa volta partiamo dalla relazione

$$P^m(Q \le b) = \Phi(b),$$

che equivale a

$$P^m\Big(\overline{X} - \frac{\sigma}{\sqrt{n}} b \le m\Big) = \Phi(b).$$

Se $b = \phi_\beta$, allora basterà che

$$1 - \alpha = \Phi(b) = \beta,$$

ovvero semplicemente $b = \phi_{1-\alpha}$, e l'intervallo è

$$\Big(\overline{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha}, +\infty\Big).$$

(c) Intervallo Unilaterale Sinistro. Non ripeteremo i calcoli, che sono analoghi ai precedenti. Si trova l'intervallo

$$\Big(-\infty, \overline{X} - \frac{\sigma}{\sqrt{n}}\phi_\alpha\Big) = \Big(-\infty, \overline{X} + \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha}\Big),$$

ricordando la prima delle relazioni (20).

Osservazione 11.12 La scelta del tipo di intervallo da considerare è in genere legata alla situazione pratica (se si deve avere una stima sia da destra che da sinistra calcoleremo un intervallo bilaterale, se invece occorre stimare il parametro solo dal basso cercheremo un intervallo unilaterale destro, e così via).

Osservazione 11.13 Per quanto riguarda le stime da sinistra unilaterale e bilaterale si ha

$$\overline{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \leq \overline{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha}.$$

Per le stime da destra vale ovviamente la diseguaglianza inversa (dimostrazione per esercizio).

Intervalli di fiducia per la media della Normale con varianza non nota.

Nella pratica gli intervalli del paragrafo precedente sono di scarsa utilità, perché nelle formule che li definiscono interviene la varianza σ^2, che in genere non si conosce. In questo caso si può sostituire σ^2 con

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

(che ne è uno stimatore), e applicare di nuovo il metodo della quantità pivotale partendo dalla v.a.

$$\frac{\overline{X} - m}{S}\sqrt{n} \sim t(n-1)$$

(Teorema 9.5) (iv). Ricordando la seconda delle relazioni (20), risulta chiaro che tutto ciò che abbiamo detto nel paragrafo precedente si può ripetere, semplicemente sostituendo σ con S e i quantili della $\mathcal{N}(0,1)$ con quelli della $t(n-1)$. Per comodità di chi legge, riportiamo comunque le formule finali.

(a) Intervallo bilaterale.

$$\Big(\overline{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1), \overline{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\Big).$$

(b) Intervallo unilaterale destro.

$$\Big(\overline{X} - \frac{S}{\sqrt{n}}t_{1-\alpha}(n-1), +\infty\Big).$$

(c) Intervallo unilaterale sinistro.

$$\Big(-\infty, \overline{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1)\Big).$$

Intervalli di fiducia per la varianza della Normale con media nota.

Qui si parte osservando che

$$Q = Q(X_1, \dots, X_n; \sigma^2) = \frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2} \sim \chi^2(n).$$

Posto

$$U^2 = \frac{\sum_{i=1}^n (X_i - m)^2}{n},$$

(che, come sappiamo, è uno stimatore della varianza), si può scrivere

$$Q = \frac{nU^2}{\sigma^2}.$$

(a) Intervallo bilaterale. Indichiamo con F_n la funzione di ripartizione della $\chi^2(n)$. Allora si ha

$$F_n(b) - F_n(a) = P^{\sigma^2}(a \le Q \le b) = P^{\sigma^2}\Big(a \le \frac{nU^2}{\sigma^2} \le b\Big) = P^{\sigma^2}\Big(\frac{nU^2}{b} \le \sigma^2 \le \frac{nU^2}{a}\Big).$$

Quindi, se al solito poniamo $b = \chi^2_\beta(n)$, $a = \chi^2_\gamma(n)$, avremo

$$1 - \alpha = F_n(b) - F_n(a) = \beta - \gamma.$$

Una scelta possibile è $\beta = 1 - \frac{\alpha}{2}$, $\gamma = \frac{\alpha}{2}$, si ottiene l'intervallo

$$\Big(\frac{nU^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}, \frac{nU^2}{\chi^2_{\frac{\alpha}{2}}(n)}\Big).$$

Gli intervalli unilaterali si ottengono in modo analogo. Riportiamo solo le formule finali.

(b) Intervallo unilaterale destro.

$$\Big(\frac{nU^2}{\chi^2_{1-\alpha}(n)}, +\infty\Big).$$

(c) Intervallo unilaterale sinistro.

$$\Big(0, \frac{nU^2}{\chi^2_{\alpha}(n)}\Big).$$

Intervalli di fiducia per la varianza della Normale con media non nota.

Poiché normalmente la media non è nota, si può cercare di sostituire m con il suo stimatore $\overline{X}$, usando, al posto di U^2, la v.a.

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

(che è ancora uno stimatore della varianza). Quindi applicheremo il metodo della quantità pivotale a partire dalla v.a.

$$Q = Q(X_1, \dots, X_n; \sigma^2) = \frac{(n-1)S^2}{\sigma^2},$$

che, dal Teorema 9.5 (ii), sappiamo avere legge $\chi^2(n-1)$. Dunque, per avere i tre nuovi intervalli, basterà sostituire nelle formule del paragrafo precedente S^2 al posto di U^2 e $n-1$ al posto di n (e i quantili della $\chi^2(n-1)$ al posto dei quantili della $\chi^2(n)$). Si ottengono così le espressioni che seguono:

(a) Intervallo bilaterale.

$$\Big(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\Big).$$

(b) Intervallo unilaterale destro.

$$\Big(\frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}, +\infty\Big).$$

(c) Intervallo unilaterale sinistro.

$$\Big(0, \frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}\Big).$$

12 Teoria dei test: generalità

Fare un test statistico significa formulare un'ipotesi riguardante il parametro $\theta \in \Theta$ (che non è noto) e pianificare un esperimento per decidere se tale ipotesi può essere ritenuta vera, e quindi accettata.

Un'*ipotesi statistica* (usualmente indicata con il simbolo H_0 e talvolta denominata *ipotesi nulla*) e la sua negazione, cioè l'ipotesi alternativa (indicata con il simbolo H_1) si formalizzano nel modo seguente.

Dato il modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$, si assegna una partizione dell'insieme Θ dei parametri in due sottoinsiemi non vuoti Θ_0 e Θ_1 (dunque $\Theta_0 \cup \Theta_1 = \Theta$ e $\Theta_0 \cap \Theta_1 = \emptyset$). Si pone poi $H_0 : \theta \in \Theta_0$ e $H_1 : \theta \in \Theta_1$.

Esempio 12.1 Nell'esempio iniziale del controllo di qualità, supponiamo di voler sottoporre a test l'ipotesi H_0 : *la probabilità che un generico pezzo sia difettoso è* < 0.1. La formalizzazione è la seguente:

$$\Theta = (0,1), \qquad \Theta_0 = (0, 0.1), \qquad \Theta_1 = [0.1, 1)$$

e dobbiamo pianificare un esperimento che ci permetta di discriminare tra l'ipotesi $H_0 : \theta \in \Theta_0$ (cioè $\theta < 0.1$) e l'alternative $H_1 : \theta \in \Theta_1$ (cioè $\theta \geq 0, 1$).

Definizione 12.2 Si chiama *funzione di test* (o semplicemente *test* una funzione $\Phi : \Omega \to \{0, 1\}$ misurabile, cioè della forma $\Phi = 1_D$, con $D \in \mathcal{F}$. L'evento $D = \{\omega \in \Omega : \Phi(\omega) = 1\}$ si chiama *regione di rigetto* o *di rifiuto*, o anche *regione critica* del test. Essa va interpretata come l'insieme dei risultati dell'esperimento pianificato che inducono (megli sarebbe dire "obbligano") lo sperimentatore a ritenere che H_0 sia falsa (se $\Phi(\omega) = 1$ significa che scegliamo H_1) e, come tale, a respingerla. In modo simmetrico, l'evento D^c si chiama *regione di accettazione* del test.

Osservazione 12.3 Identificheremo spesso un test con la sua regione critica: diremo quindi "il test D" invece che "il test di regione critica D".

Esempio 12.4 Ancora nell'esempio del controllo di qualità (con l'ipotesi H_0 stabilita in 12.1), supponiamo che l'esperimento da effettuare consista nel provare 100 pezzi, e che si sia deciso di rifiutare l'ipotesi nel caso che il numero di pezzi difettosi risulti maggiore (strettamente) di 11. La formalizzazione della situazione è allora la seguente.
Si prende il modello statistico del campione $X_1, \dots, X_{100}$ di legge $\mathcal{B}(1, \theta)$; si pone poi $T = \sum_{i=1}^{100} X_i$ e

$$\Phi(\omega) = \begin{cases} 0 & \text{se } T(\omega) \leq 11 \\ 1 & \text{se } T(\omega) \geq 12. \end{cases}$$

Evidentemente Φ è la funzione di test e si può scrivere $\Phi = 1_D$, dove $D = \{\omega \in \Omega : T(\omega) \geq 12\}$ è la regione critica del test.

Osservazione 12.5 Poniamo

$$g(\theta) = \begin{cases} 0 & \text{se } \theta \in \Theta_0 \\ 1 & \text{se } \theta \in \Theta_1. \end{cases}$$

(cioè $g(\theta)$ è l'indice di quello tra i due sottoinsiemi Θ_0 e Θ_1 a cui θ appartiene). Allora la funzione $\Phi = 1_D$ sopra introdotta non è altro che uno stimatore di $g(\theta)$ (nel senso della definizione data a suo tempo): significa che, nell'eventualità che $\omega \in D^c$ (cioè $\Phi(\omega) = 0$) allora decidiamo che $\theta \in \Theta_0$ (e cioè $g(\theta) = 0$), e quindi accettiamo H_0, mentre se $\omega \in D$ (cioè $\Phi(\omega) = 1$) allora diciamo che $\theta \in \Theta_1$ (e cioè $g(\theta) = 1$), e quindi rifiutiamo H_0.

Generalmente si ha a disposizione un certo numero di test, e bisogna decidere qual è il più affidabile. L'osservazione precedente (cioè l'interpretazione della funzione di test come uno stimatore) ci dice che ciò è possibile introducendo anche in questo caso (come per gli stimatori) un *costo*.

Definizione 12.6 Il *costo* in un problema di test è la funzione $C : \Theta \times \{0, 1\} \to \mathbb{R}$ definita da

$$C(\theta, a) = \begin{cases} a & \text{se } \theta \in \Theta_0 \\ 1 - a & \text{se } \theta \in \Theta_1. \end{cases}$$

Ricordiamo che il costo relativo alla stima $U(\omega)$ è definito come $C(\theta, U(\omega))$, e dunque nel nostro caso vale $C(\theta, 1_D(\omega))$: supponiamo che $\theta \in \Theta_0$; se decidiamo che $\theta \in \Theta_0$, allora $C(\theta, 1_D(\omega)) = C(\theta, 0) = 0$ (non paghiamo alcun costo, avendo fatto la scelta giusta); se invece decidiamo che $\theta \in \Theta_1$, allora $C(\theta, 1_D(\omega)) = C(\theta, 1) = 1$ (paghiamo un costo unitario, avendo fatto la scelta sbagliata). Le cose vanno in modo esattamente simmetrico se $\theta \in \Theta_1$.

Di conseguenza il *rischio* dello "stimatore" 1_D è

$$R_{1_D}(\theta) = E^\theta[C(\theta, 1_D)] = \begin{cases} E^\theta[1_D] = P^\theta(D) & \text{per } \theta \in \Theta_0 \\ 1 - E^\theta[1_D] = P^\theta(D^c) & \text{per } \theta \in \Theta_1. \end{cases} \tag{24}$$

La decisione che prenderemo ($\theta \in \Theta_0$ oppure $\theta \in \Theta_1$) dipende dal risultato dell'esperimento, ed è dunque aleatoria. C'è dunque una probabilità di prendere una decisione sbagliata. Precisamente:

Definizione 12.7 (a) Si chiama *errore di prima specie* l'errore che consiste nel respingere a torto l'ipotesi H_0 (l'esperimento che abbiamo effettuato ha dato un risultato $\omega \in D$, ma in realtà $\theta \in \Theta_0$). La probabilità di commettere questo tipo di errore è $P^\theta(D)$, per $\theta \in \Theta_0$.

(b) Si chiama *errore di seconda specie* l'errore che consiste nell' accettare a torto l'ipotesi H_0 (l'esperimento che abbiamo effettuato ha dato un risultato $\omega \in D^c$, ma in realtà $\theta \in \Theta_1$). La probabilità di commettere questo tipo di errore è $P^\theta(D^c)$, per $\theta \in \Theta_1$.

Seguendo ancora la terminologia introdotta per gli stimatori, possiamo dare la definizione seguente:

Definizione 12.8 Un test (di regione critica) D è *preferibile* ad un test (di regione critica) D^* se $R_{1_D}(\theta) \leq R_{1_{D^*}}(\theta)$ per ogni$\theta \in \Theta$. Ricordando l'espressione del rischio, si deve cioè avere

$$\begin{array}{rcll} P^\theta(D) & \leq & P^\theta(D^*) & \text{per} \quad \theta \in \Theta_0, \\ P^\theta(D) & \geq & P^\theta(D^*) & \text{per} \quad \theta \in \Theta_1, \end{array}$$

(ovvero, se l'ipotesi è vera, è meno probabile respingerla con il test D piuttosto che con il test D^*.

In modo analogo si possono dare le definizioni di test *strettamente preferibile*, test *ammissibile* e test *ottimale*.

Con la relazione di preferibilità si ottiene solo un ordinamento parziale fra test: esistono infatti test tra loro non confrontabili (come accadeva per gli stimatori).

Esiste un criterio di ordinamento tra test nel quale, a differenza del precedente, H_0 e H_1 non giocano un ruolo simmetrico. Spesso infatti si ritiene che sia meglio accettare un'ipotesi falsa piuttosto che respingere un'ipotesi vera. Come esempio, si può pensare ad un test del DNA usato per decidere la colpevolezza di una persona accusata di omicidio in uno stato dove per questo reato è prevista la pena di morte: se l'ipotesi H_0 corrisponde all'affermazione "l'imputato è innocente" e l'alternativa H_1 all'affermazione "l'imputato è colpevole", accettare l'ipotesi H_0 falsa equivale a mandare libero un assassino, ma respingerla quando è vera significa condannare a morte un innocente. Questo è chiaramente un errore più grave, non solo dal punto di vista morale, ma anche per il semplice fatto che al primo errore si può in qualche modo porre rimedio, mentre al secondo ovviamente no.
In altri termini, si considera meno grave commettere un errore di seconda specie piuttosto che uno di prima specie. Questo conduce ad una nuova relazione di preordinamento sulle funzioni di test, che traduce il fatto che lo sperimentatore ha la necessità di cautelarsi contro la possibilità di errore di prima specie.

Definizione 12.9 (a) Si chiama *taglia* del test la "massima probabilità di errore di prima specie", cioè il numero

$$\sup_{\theta \in \Theta_0} P^\theta(D).$$

(b) Fissato $\alpha \in (0,1)$, si dice che il test è *di livello* α se la sua taglia è minore o uguale ad α:

$$\sup_{\theta \in \Theta_0} P^\theta(D) \leq \alpha.$$

Contrariamente al caso in cui $\theta \in \Theta_0$, nel caso in cui $\theta \in \Theta_1$, $P^\theta(D)$ rappresenta la probabilità di prendere la decisione giusta. Questo giustifica la definizione seguente:

Definizione 12.10 La funzione definita su Θ_1 da $\theta \mapsto P^\theta(D)$ si chiama *potenza del test*.

Osservazione 12.11 In genere, H_0 è un'ipotesi che si spera di poter respingere, cioè si tratta di un'ipotesi per così dire "allarmante"; in quest'ottica, la quantità $1 - P^\theta(D)$, con $\theta \in \Theta_1$ (probabilità di accettare a torto H_0) è spesso chiamata probabilità di "falso allarme", o probabilità di "falso positivo". In particolare la seconda espressione è comune (e nota a tutti) per i test medici, e si usa anche nell'ambito degli studi sulla sicurezza informatica.

Allo scopo di garantirsi contro la possibilità di errore di prima specie, si fissa per prima cosa un livello α; in altri termini, il numero α è la "massima probabilità di errore di prima specie che lo sperimentatore è disposto a tollerare". D'altro canto, è anche auspicabile che la potenza non sia

troppo bassa (perché questo significherebbe una probabilità bassa di fare la scelta corretta nel caso che sia vera l'alternativa). Dunque, se $\mathcal{D}_\alpha$ è la classe dei test di livello α, si dice che il test (di regione critica) D è *uniformemente il più potente* (abbreviato in UPP, o UMP, *uniformly most powerful* in inglese) tra i test di livello α se $D \in \mathcal{D}_\alpha$ e D è *più potente* di ogni altro test D^* appartenente a $\mathcal{D}_\alpha$, cioè se, per ogni $D^* \in \mathcal{D}_\alpha$, si ha

$$P^\theta(D) \geq P^\theta(D^*), \qquad \forall\, \theta \in \Theta_1.$$

In altre parole si cerca di scegliere il test D la cui potenza sia la più alta possibile (tra i test di $\mathcal{D}_\alpha$).

Osservazione 12.12 La teoria delle regioni di fiducia può essere considerata un caso particolare della teoria di test. Fissiamo infatti $\theta_0 \in \Theta$ e consideriamo un test per l'ipotesi $H_0 : \theta = \theta_0$ contro l'alternativa $H_1 : \theta \neq \theta_0$. Supponiamo che $D(\theta_0)$ sia la regione critica di un test di livello α dell'ipotesi H_0 contro l'alternativa H_1

Supponiamo di avere scelto $D(\theta)$ con questo procedimento per ogni $\theta \in \Theta$ (mantenendo fisso α); consideriamo la regione di fiducia $S : \Omega \to \mathcal{P}(\Theta)$ definita da

$$S(\omega) = \{\theta \in \Theta : \omega \notin D(\theta)\}.$$

Si ha subito

$$P^\theta(\theta \notin S) = P^\theta(\{\omega : \omega \in D(\theta)\}) = P^\theta\big(D(\theta)\big) \leq \alpha,$$

e quindi S è una regione di fiducia di livello $1 - \alpha$.

Viceversa, se S è una regione di fiducia di livello $1 - \alpha$, l'insieme $D(\theta_0) = \{\omega : \theta_0 \notin S(\omega)\}$ è la regione critica di un test di livello α per l'ipotesi $H_0 : \theta = \theta_0$ contro l'alternativa $\theta \neq \theta_0$. Infatti in questo caso $\Theta_0 = \{\theta_0\}$ e quindi

$$\sup_{\theta \in \Theta_0} P^\theta\big(D(\theta_0)\big) = P^{\theta_0}\big(D(\theta_0)\big) = P^{\theta_0}(\theta_0 \notin S) \leq \alpha.$$

13 Il lemma di Neyman-Pearson

In questo paragrafo studieremo il caso in cui $\Theta = \{0, 1\}$ (due soli elementi). I risultati che otteremo potranno essere poi applicati anche a casi più generali.

Si dice che l'ipotesi (risp. l'alternativa) è *semplice* se Θ_0 (risp. Θ_1) è formato da un solo punto. Supponendo che sia l'ipotesi che l'alternativa siano semplici, Θ può convenzionalmente essere rappresentato nella forma $\Theta = \{0, 1\}$. In questo caso è possibile dire esattamente quali sono i "buoni" test. Si tratta della teoria di Neyman-Pearson, di cui ci occupiamo adesso.

In questo paragrafo si suppone dunque che il modello statistico sia del tipo $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \{0, 1\}\}\big)$, cioè sia formato da due sole probabilità, P^0 e P^1. Cominciamo con l'osservare che un tale modello è certamente dominato (una misura dominante è per esempio $\mu = P^0 + P^1$); scelta allora una misura dominante μ, indichiamo con p^i $(i = 0, 1)$ una scelta delle densità $\frac{\mathrm{d}P^i}{\mathrm{d}\mu}$.

Lemma 13.1 (di Neyman-Pearson). *Sia C un numero reale strettamente positivo. L'evento $D = \{p^1 > Cp^0\}$ è la regione critica di un test per l'ipotesi $H_0 : \theta = 0$ contro l'alternativa $H_1 : \theta = 1$, con taglia $\alpha = P^0(D)$. Questo test è ammissibile e UPP tra i test di livello α.*

Osservazione 13.2 La funzione $\frac{p^1}{p^0}$ viene detta *rapporto di verosimiglianza.*

Osservazione 13.3 Questo test rifiuta H_0 se la densità di P^1 è molto più grande di quella di P^0: intuitivamente, ricordando che valori grandi della densità corrispondono a valori dell'osservazione più probabili, ciò significa che se p^1 è grande (rispetto a p^0) allora siamo portati a credere che sia vera l'alternativa (e infatti per $\frac{p^1}{p^0}$ grande siamo nella zona di rigetto). Se invece p^0 è grande (rispetto a p^1) allora penseremo che è vera l'ipotesi (e infatti per $\frac{p^1}{p^0}$ piccolo siamo nella regione di accettazione).

Dimostrazione. Sia D^* un'altra regione critica; vale allora la diseguaglianza

$$(1_{D^*} - 1_D)(p^1 - Cp^0) \leq 0. \tag{25}$$

Infatti, se $p^1 - Cp^0 > 0$ (risp. ≤ 0), allora $1_D = 1$ (risp. $1_D = 0$); integrando la (25) rispetto a μ otteniamo

$$\Big(\int_{D^*} p^1 \mathrm{d}\,\mu - \int_D p^1 \mathrm{d}\,\mu\Big) - C\Big(\int_{D^*} p^0 \mathrm{d}\,\mu - \int_D p^0 \mathrm{d}\,\mu\Big) \leq 0,$$

e cioè

$$P^1(D^*) - P^1(D) \leq C\big(P^0(D^*) - P^0(D)\big). \tag{26}$$

Da questa diseguaglianza segue che

(i) D è ammissibile;

(ii) D è più potente di D^* se D^* ha livello $\alpha = P^0(D)$.

Infatti

(i) D è ammissibile se non esiste alcun test D^* strettamente preferibile. Se per assurdo tale D^* esistesse, esso sarebbe intanto preferibile a D, e quindi, ricordando l'espressione del rischio (24), avremmo

$$P^0(D^*) \leq P^0(D) \qquad \text{e} \qquad P^1(D^*) \geq P^1(D).$$

Ma, per la diseguaglianza (26), queste due disequazioni sono compatibili se e solo se

$$P^0(D^*) = P^0(D) \qquad \text{e} \qquad P^1(D^*) = P^1(D).$$

Dunque D^* non può essere strettamente preferibile a D, perché in almeno uno dei due casi ($\theta = 0$ e $\theta = 1$) dovrebbe valere la diseguaglianza stretta.

(ii) D^* ha livello $\alpha = P^0(D)$ se e solo se $P^0(D^*) = \alpha = P^0(D)$, e quindi, ancora per la (26), si ha $P^1(D) \geq P^1(D^*)$, il che significa che la potenza di D non è inferiore alla potenza di D^*.

□

Osservazione 13.4 Si verifica facilmente che il lemma precedente rimane valido anche per $C = 0$ (e in tal caso $D = \{p^1 > 0\}$) oppure per $C = +\infty$, con la convenzione $0 \cdot (+\infty) = 0$ (e in tal caso si pone $D = \{p^0 = 0, p^1 > 0\}$).

Osservazione 13.5 Si presenta ora il problema di determinare C, se si assegna il livello $\alpha \in (0,1)$ a priori. Il test di regione critica $D = \{p^1 > Cp^0\}$ del Lemma di Neyman-Pearson 13.1 sarà di livello α se C è tale che $P^0(p_1 > Cp^0) = \alpha$, ovvero se, posto, per ogni numero reale positivo t, $f(t) = P^0(p_1 > tp^0)$, esiste una soluzione dell'equazione $f(t) = \alpha$ (detto in altre parole, se la funzione $f : \mathbb{R}^+ \to (0,1)$ è surgettiva). Purtroppo, in generale questo non è vero: infatti, si vede facilmente che la funzione f è non crescente, continua a destra ed inoltre $\lim_{t\to+\infty} f(t) = 0$ (verifica

per esercizio; osservare che la funzione $g(t) = 1 - f(t)$ è una f.d.r.). Tuttavia f in generale non è continua e dunque, se α appartiene ad un intervallo del tipo $(f(t_0), \lim_{\tau \to t_0^-} f(t))$, con t_0 punto di discontinuità di f, la soluzione dell'equazione precedente non esiste. Infatti si vede facilmente che valgono solo le diseguaglianze

$$\lim_{t \to C^-} f(t) \geq \alpha \geq f(C),$$

e cioè

$$P^0(p^1 \geq Cp^0) \geq \alpha \geq P^0(p_1 > Cp^0).$$

Ne segue che condizione necessaria e sufficiente affinché l'equazione $f(x) = \alpha$ abbia soluzione è che f sia continua o, equivalentemente, che $P^0(p_1 = tp^0) = 0$ per ogni t (verifica per esercizio; comunque si tratta solo della nota condizione affinché una f.d.r. sia continua).

Notiamo per inciso che, se f è continua e strettamente decrescente, la soluzione esiste ed è anche unica (grazie alla stretta monotonia di f).

Notiamo anche che la funzione $\alpha \mapsto \inf\{t : f(t) \leq \alpha\}$ non è altro che $f^{\leftarrow}(\alpha)$, secondo la definizione di inversa generalizzata data a suo tempo in (11.1), con la sola differenza che in questo caso la funzione f è non crescente anziché non decrescente. Dunque l'equazione $f(C) = \alpha$ può scriversi anche nella forma

$$C = f^{\leftarrow}(\alpha).$$

Anticipiamo infine che, proprio per poter costruire un test che abbia esattamente la taglia desiderata (e anche per poter caratterizzare in modo preciso tutti i test ammissibili nel caso di ipotesi e alternativa semplici) è stata costruita la teoria dei test aleatori, di cui ci occuperemo in seguito.

Qui diamo intanto un esempio di test deterministico ad ipotesi ed alternativa semplici.

Esempio 13.6 Si vuole verificare se è stato inviato un segnale deterministico noto, funzione del tempo, $s(t)$, con $0 \leq t \leq T$, al quale, tuttavia, si sovrappone un rumore aleatorio, anch'esso funzione del tempo, $B(t)$ (cioè il segnale risultante dopo l'invio è $U(t) = s(t) + B(t)$). A questo scopo si campiona il segnale in n istanti $t_1 < t_2 < \cdots < t_n$, e poniamo $s_i = s(t_i)$, $B_i = B(t_i)$.
Supponiamo che le v.a. B_i siano indipendenti e di legge $\mathcal{N}(0, \sigma^2)$ (con σ^2 noto). Sia poi $(U_1, \ldots, U_n)$ la sequenza degli n segnali ricevuti, e sia H_0 l'ipotesi "il segnale è stato effettivamente inviato". Sotto H_0, le v.a. U_i sono dunque indipendenti e di legge $\mathcal{N}(s_i, \sigma^2)$, mentre sotto l'alternativa H_1 (che significa che il segnale non è stato inviato, e dunque si è ricevuto solo il rumore aleatorio B), le U_i sono ancora indipendenti ma di legge $\mathcal{N}(0, \sigma^2)$.

Rispetto alla misura di Lebesgue n- dimensionale, le due verosimiglianze sono rispettivamente

$$p^0(u_1, \ldots, u_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^n (u_i - s_i)^2 \Big);$$

$$p^1(u_1, \ldots, u_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^n u_i^2 \Big);$$

sappiamo inoltre dal Lemma di Neyman-Pearson 13.1 che le buone regioni critiche sono del tipo (qui $\lambda = \frac{1}{C}$)

$$D_\lambda = \Big\{ \frac{p^0}{p^1} < \lambda \Big\}.$$

Dato che

$$\frac{p^0}{p^1}(u_1, \ldots, u_n) = \exp\Big\{ \frac{1}{\sigma^2} \Big(\sum_{i=1}^n s_i u_i - \frac{1}{2} \sum_{i=1}^n s_i^2 \Big) \Big\},$$

con qualche conto si vede che

$$D_\lambda = \Big\{(u_1,\dots,u_n) : \sum_{i=1}^n s_i u_i < \sigma^2 \log\lambda + \frac{1}{2}\sum_{i=1}^n s_i^2\Big\} = \Big\{(u_1,\dots,u_n) : \sum_{i=1}^n s_i u_i < \tilde\lambda\Big\}$$
$$= \Big\{\omega\in\Omega : \sum_{i=1}^n s_i U_i(\omega) < \tilde\lambda\Big\},$$

dove ovviamente $\tilde\lambda = \sigma^2\log\lambda + \frac{1}{2}\sum_{i=1}^n s_i^2$. A questo punto, supponendo assegnato il livello α, dobbiamo determinare $\tilde\lambda$ in modo che la nostra regione critica abbia probabilità (sotto P^0) minore o uguale a α. D'altra parte, sotto H_0 la v.a $\sum_{i=1}^n s_i U_i$ ha legge $\mathcal{N}\big(\sum_i s_i^2, \sigma^2(\sum_i s_i^2)\big)$ e quindi la taglia del test è

$$P^0(D_\lambda) = P^0\Big(\sum_{i=1}^n s_i U_i < \tilde\lambda\Big) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^r \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}\,x = \Phi(r),$$

dove $r = \frac{\tilde\lambda - \sum_i s_i^2}{\sigma\sqrt{\sum_i s_i^2}}$ e Φ è la funzione di ripartizione della legge Normale standard. $\tilde\lambda$ si troverà dunque (dalle tavole della Normale standard), imponendo che sia

$$\Phi(r) = \Phi\Big(\frac{\tilde\lambda - \sum_i s_i^2}{\sigma\sqrt{\sum_i s_i^2}}\Big) \le \alpha,$$

e cioè

$$\tilde\lambda \le \phi_\alpha\sigma\sqrt{\sum_i s_i^2} + \sum_i s_i^2.$$

Si può anche calcolare la potenza di questo test. Infatti, sotto l'alternativa H_1, la v.a. $\sum_{i=1}^n s_i U_i$ ha legge $\mathcal{N}\big(0, \sigma^2(\sum_i s_i^2)\big)$ e quindi la potenza vale

$$P^1(D_\lambda) = P^1\Big(\sum_{i=1}^n s_i U_i < \tilde\lambda\Big) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^s \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}\,x = \Phi(s),$$

dove $s = \frac{\tilde\lambda}{\sigma\sqrt{\sum_i s_i^2}}$.
Se per esempio vogliamo che l'errore di seconda specie sia inferiore a β, dovremo imporre che

$$1 - \Phi(s) = 1 - \Phi\Big(\frac{\tilde\lambda}{\sigma\sqrt{\sum_i s_i^2}}\Big) \le \beta,$$

ovvero

$$\tilde\lambda \ge \phi_{1-\beta}\sigma\sqrt{\sum_i s_i^2}.$$

Pertanto troviamo per $\tilde\lambda$ le relazioni

$$\phi_{1-\beta}\sigma\sqrt{\sum_i s_i^2} \le \tilde\lambda \le \phi_\alpha\sigma\sqrt{\sum_i s_i^2} + \sum_i s_i^2,$$

per cui dovrà essere

$$\phi_{1-\beta}\sigma\sqrt{\sum_i s_i^2} < \phi_\alpha\sigma\sqrt{\sum_i s_i^2} + \sum_i s_i^2,$$

e cioè

$$\big(\phi_{1-\beta} - \phi_\alpha\big)\sigma < \sqrt{\sum_i s_i^2}.$$

Se vogliamo essere sicuri che la relazione precedente sia valida per ogni $(s_1, \dots, s_n)$ (anche piccoli), dovremo imporre che $\phi_{1-\beta} - \phi_\alpha \le 0$, da cui $\beta \ge 1 - \alpha$. Ad esempio, se $\alpha = 0,05$, dovrà essere $\beta \ge 0.95$, il che fornisce una probabilità di errore di seconda specie troppo alta. Dunque questo test non è soddisfacente nella pratica.

Esercizio 13.7 Rifare i conti dell'Esempio precedente, prendendo come ipotesi H_0 : "il segnale non è stato inviato". Si troverà che è possibile rendere contemporaneamente bassa la taglia del test e alta la potenza.

Esempio 13.8 L'esempio precedente è parametrico. In realtà il metodo di Neyman-Pearson funziona anche in casi non parametrici, per esempio per testare una contro l'altra due leggi perfettamente specificate, come mostra l'esempio che segue.

Si ha un campione di taglia $n = 2$ con legge μ su $\mathbb{R}$, e supponiamo di voler eseguire un test per l'ipotesi $H_0 : \mu = \mu_0 := \mathcal{N}(0,1)$ contro $H_1 : \mu = \mu_1 := \mathcal{U}(0,2)$, al livello $\alpha = 0,05$. Rispetto alla misura di Lebesgue bidimensionale, le due verosimiglianze sono

$$p^0(x_1,x_2) = \frac{1}{2\pi}\exp\Big(-\frac{x_1^2+x_2^2}{2}\Big); \qquad p^1(x_1,x_2) = \frac{1}{4}1_{(0,2)}(x_1)\cdot 1_{(0,2)}(x_2);$$

$$\frac{p^1}{p^0}(x_1,x_2) = \frac{\pi}{2}1_{(0,2)}(x_1)\cdot 1_{(0,2)}(x_2)\exp\Big(\frac{x_1^2+x_2^2}{2}\Big)$$

Se $(x_1,x_2) \notin (0,2)\times(0,2)$, allora si decide per H_0, poiché in questo caso la verosimiglianza p^1 è nulla. Se invece $(x_1,x_2) \in (0,2)\times(0,2)$, la regione critica deve essere

$$\Big\{\frac{\pi}{2}\exp\Big(\frac{x_1^2+x_2^2}{2}\Big) > C\Big\} \cap \big\{(0,2)\times(0,2)\big\} = \big\{x_1^2+x_2^2 > \lambda\big\} \cap \big\{(0,2)\times(0,2)\big\},$$

con $\lambda = 2\log\frac{2C}{\pi}$. Scritta in termini di v.a la regione critica è

$$\big\{X_1^2 + X_2^2 > \lambda, 0 < X_1 < 2, 0 < X_2 < 2\big\}.$$

Calcoliamo λ dalla relazione

$$P^0\big(X_1^2 + X_2^2 > \lambda, 0 < X_1 < 2, 0 < X_2 < 2\big) \le 0,05.$$

Poiché $\big\{X_1^2 + X_2^2 > \lambda, 0 < X_1 < 2, 0 < X_2 < 2\big\} \subseteq \big\{X_1^2 + X_2^2 > \lambda\big\}$, è sufficiente imporre che

$$P^0\big(X_1^2 + X_2^2 > \lambda\big) \le 0,05.$$

Sotto H_0, la v.a. $X_1^2 + X_2^2$ ha legge $\chi^2(2)$, dunque, indicata con F la f.d.r. della $\chi^2(2)$, abbiamo

$$0,05 \ge P^0\big(X_1^2 + X_2^2 > \lambda\big) = 1 - F(\lambda),$$

da cui si ricava

$$\lambda \ge \chi^2_{1-0,05}(2) = \chi^2_{0,95}(2) = 5,99.$$

Per calcolare la potenza di questo test, bisognerebbe calcolare, per il valore di λ scelto,

$$P^1\big(X_1^2 + X_2^2 > \lambda, 0 < X_1 < 2, 0 < X_2 < 2\big),$$

ma tralasciamo questo calcolo perché necessita della conoscenza della legge di $X_1^2 + X_2^2$ con X_1 e X_2 uniformi e indipendenti, cosa troppo complicata.

14 Test aleatori: il Teorema di Neyman-Pearson

Definizione 14.1 *Un test per il quale la funzione di test è a valori in* <u>tutto</u> $[0,1]$ *($\Phi : \Omega \to [0,1]$) si chiama test aleatorio*

Osservazione 14.2 In altre parole, se $\Phi(\omega) = p$ (con $p \in [0,1]$), allora si rifiuta H_0 con probabilità p; in particolare, se $p = 0$ (risp. $p = 1$) l'ipotesi viene rifiutata con probabilità 0, e cioè accettata (risp. rifiutata con probabilità 1, e cioè rifiutata). Si può immaginare che la decisione di accettare o meno H_0 dipenda dal risultato del lancio (indipendente dall'osservazione) di una moneta che dà "testa" con probabilità p, il quale porta a decidere H_1 se esce "testa", H_0 se esce "croce".

La funzione costo è la stessa che nel caso di ipotesi e alternativa semplici, e cioè

$$C(\theta, a) = \begin{cases} a & \text{se } \theta \in \Theta_0 \\ 1 - a & \text{se } \theta \in \Theta_1, \end{cases}$$

ma con la differenza che adesso $a \in [0,1]$. Il rischio è

$$R_\Phi(\theta) = \begin{cases} E^\theta[\Phi] & \text{se } \theta \in \Theta_0 \\ 1 - E^\theta[\Phi] & \text{se } \theta \in \Theta_1. \end{cases}$$

Si conserva per i test aleatori tutto il vocabolario usato per i test deterministici (test preferibile, ammissibile, livello, potenza etc.. . .).

Torniamo alla situazione iniziale di un modello statistico con due sole probabilità, P^0 e P^1.

Definizione 14.3 Si chiama *test di Neyman-Pearson* un test aleatorio per il quale esista una costante positiva C tale che Φ valga 1 sull'evento $\{p^1 > Cp^0\}$ e 0 sull'evento $\{p^1 < Cp^0\}$, μ- quasi ovunque.

Osservazione 14.4 Per quanto ovvio, sottolineiamo il fatto che la taglia di un test Φ di Neyman-Pearson vale $\alpha = E^0[\Phi]$.

Teorema 14.5 (di Neyman-Pearson) *(a) Per ogni $\alpha \in (0,1)$, si possono determinare due costanti $C \geq 0$ e $\gamma \in [0,1]$ tali che il test (di Neyman-Pearson)*

$$\Phi = 1_{\{p^1 > Cp^0\}} + \gamma 1_{\{p^1 = Cp^0\}}$$

abbia taglia α.

(b) Ogni test di Neyman-Pearson è ammissibile e UPP tra tutti i test di taglia $E^0[\Phi]$.

(c) Condizione necessaria e sufficiente affinché Φ sia di Neyman-Pearson è che, per ogni Φ^* tale che $E^0[\Phi^*] \leq E^0[\Phi]$, si abbia $E^1[\Phi^*] \leq E^1[\Phi]$ (cioè se Φ^* ha taglia più bassa di Φ, allora ha anche potenza più bassa).

(d) Ogni test ammissibile è di Neyman-Pearson.

Dimostrazione. (a) Sia f la funzione definita da $f(x) = P^0(p^1 > xp^0)$. Abbiamo visto in precedenza (Osservazione 13.5) che, posto $C = f^{\leftarrow}(\alpha)$, valgono le diseguaglianze

$$P^0(p^1 \geq Cp^0) \geq \alpha \geq P^0(p^1 > Cp^0)$$

e che $\alpha = P^0(p^1 > Cp^0)$ se e solo se $P^0(p^1 = Cp^0) = 0$. Un test della forma $\Phi = 1_{\{p^1 > Cp^0\}} + \gamma 1_{\{p^1 = Cp^0\}}$ ha taglia

$$E^0[\Phi] = P^0(p^1 > Cp^0) + \gamma P^0(p^1 = Cp^0)$$

Se $P^0(p^1 = Cp^0) \neq 0$, allora $E^0[\Phi] = \alpha$ se e solo se

$$\gamma = \frac{\alpha - P^0(p^1 > Cp^0)}{P^0(p^1 = Cp^0)}.$$

Se invece $P^0(p^1 = Cp^0) = 0$, si ha

$$E^0[\Phi] = P^0(p^1 > Cp^0) = \alpha,$$

qualunque sia γ.

(b) Supponiamo che Φ sia di Neyman-Pearson, e si Φ^* un altro test. Sia C la costante relativa a Φ fornita dalla Definizione 14.3. Si parte dalla diseguaglianza

$$\Big(\Phi^*(\omega) - \Phi(\omega)\Big)\Big(p^1(\omega) - Cp^0(\omega)\Big) \leq 0;$$

(si dimostra come quella del Lemma di Neyman-Pearson 13.1); integrando rispetto a μ si trova la relazione

$$E^1[\Phi^*] - E^1[\Phi] \leq C\big(E^0[\Phi^*] - E^0[\Phi]\big); \tag{27}$$

si prosegue poi esattamente come nella dimostrazione del Lemma di Neyman-Pearson 13.1.

(c) Se Φ è di Neyman-Pearson e Φ^* un altro test, la relazione (27) dice che, se $E^0[\Phi^*] \leq E^0[\Phi]$,allora $E^1[\Phi^*] \leq E^1[\Phi]$.

Viceversa, sia Φ un test per il quale vale la condizione enunciata in (c), e poniamo $\alpha = E^0[\Phi]$. Per il punto (a), esiste un test Φ^* di Neyman-Pearson avente taglia α, il che significa che $E^0[\Phi^*] = E^0[\Phi]$. Inoltre Φ^*, essendo di Neyman-Pearson, soddisfa la condizione enunciata in (c) grazie alla prima parte di questa dimostrazione; quindi, dato che $E^0[\Phi] = E^0[\Phi^*]$, si ricava che $E^1[\Phi] \leq E^1[\Phi^*]$. D'altra parte per ipotesi anche Φ verifica la condizione enunciata in (c), e dunque deve essere anche $E^1[\Phi^*] \leq E^1[\Phi]$. Allora le ultime due uguaglianze sono compatibili se e solo se $E^1[\Phi^*] = E^1[\Phi]$. Ne deduciamo che, per ogni costante C, si ha

$$\int \Big(\Phi^*(\omega) - \Phi(\omega)\Big)\Big(p^1(\omega) - Cp^0(\omega)\Big)\,\mathrm{d}\,\mu = \big(E^1[\Phi^*] - E^1[\Phi]\big) - C\big(E^0[\Phi^*] - E^0[\Phi]\big) = 0.$$

Sia ora C^* la costante associata a Φ^* dalla definizione 14.3 (ricordiamo che Φ^* è di Neyman-Pearson); sappiamo (relazione dimostrata nel punto (b)) che

$$\Big(\Phi^*(\omega) - \Phi(\omega)\Big)\Big(p^1(\omega) - C^*p^0(\omega)\Big) \leq 0;$$

dunque questa funzione è non positiva e con integrale nullo su Ω, e di conseguenza è μ-q.o. nulla; ciò significa che sull'evento $\{p^1 \neq C^*p^0\}$ si deve avere $\Phi = \Phi^*$; quindi anche Φ è di Neyman-Pearson (esattamente come Φ^*, Φ vale 1 su $\{p^1 > C^*p^0\}$ e vale 0 su $\{p^1 < C^*p^0\}$).

(d) Sia Φ ammissibile, e sia Φ^* un test di Neyman-Pearson con taglia $\alpha = E^0[\Phi]$ (l'esistenza è garantita dal punto (a)). Dunque $E^0[\Phi] = E^0[\Phi^*]$, il che significa che $R_\Phi(0) = R_{\Phi^*}(0)$. Poiché Φ è ammissibile, non può accadere che $R_{\Phi^*}(1) < R_\Phi(1)$, e cioè dovrà essere $E^1[\Phi^*] \leq E^1[\Phi]$. D'altra parte anche Φ^* è ammissibile per il punto (b), e quindi sarà anche $E^1[\Phi] \leq E^1[\Phi^*]$. Queste due relazioni sono compatibili tra loro se e solo se $E^1[\Phi^*] = E^1[\Phi]$. A questo punto si procede come nella dimostrazione del punto (c).

□

Corollario 14.6 *Se Φ è un test di Neyman-Pearson, allora $E^1[\Phi] \geq E^0[\Phi]$ (cioè la potenza è non inferiore alla taglia).*

DIMOSTRAZIONE. Poniamo $\alpha = E^0[\Phi]$ e consideriamo il test $\Phi^* \equiv \alpha$. Allora, dato che Φ è UPP, si ha

$$E^1[\Phi] \geq E^1[\Phi^*] = \alpha = E^0[\Phi].$$

□

Esempio 14.7 Sia $(X_1, \ldots, X_n)$ un campione di taglia n e legge Π_θ, e sia $\alpha \in (0,1)$ assegnato. Si vuole eseguire il test di ipotesi nulla $H_0 : \theta = \theta_0$ contro l'alternativa $H_1 : \theta = \theta_1$ dove $\theta_0 \neq \theta_1$ sono due valori assegnati. In base al Lemma di Neyman-Pearson 13.1, la regione critica deve essere del tipo

$$\{p_1 > Cp_0\} = \Big\{\frac{p^1}{p^0} > C\Big\}.$$

Dato che

$$\frac{p^1}{p^0}(k_1, \ldots, k_n) = \frac{\dfrac{\theta_1^{\sum_{i=1}^n k_i}}{\prod_{i=1}^n (k_i!)} \mathrm{e}^{-n\theta_1}}{\dfrac{\theta_0^{\sum_{i=1}^n k_i}}{\prod_{i=1}^n (k_i!)} \mathrm{e}^{-n\theta_0}} = \Big(\frac{\theta_1}{\theta_0}\Big)^{\sum_{i=1}^n k_i} \mathrm{e}^{n(\theta_0 - \theta_1)}$$

si ha

$$\begin{aligned}
&\Big\{(k_1, \ldots, k_n) : \frac{p^1}{p^0}(k_1, \ldots, k_n) > C\Big\} = \Big\{(k_1, \ldots, k_n) : \Big(\frac{\theta_1}{\theta_0}\Big)^{\sum_{i=1}^n k_i} \mathrm{e}^{n(\theta_0 - \theta_1)} > C\Big\} \\
&= \Big\{(k_1, \ldots, k_n) : \big(\sum_{i=1}^n k_i\big)\Big(\log \frac{\theta_1}{\theta_0}\Big) > \log C - n(\theta_0 - \theta_1)\Big\} \\
&= \begin{cases} \Big\{(k_1, \ldots, k_n) : \sum_{i=1}^n k_i > \lambda\Big\} & \text{se} \quad \theta_1 > \theta_0 \\ \Big\{(k_1, \ldots, k_n) : \sum_{i=1}^n k_i < \lambda\Big\} & \text{se} \quad \theta_1 < \theta_0, \end{cases}
\end{aligned}$$

con $\lambda = \frac{\log C - n(\theta_0 - \theta_1)}{\log \frac{\theta_1}{\theta_0}}$.

Vediamo un esempio numerico. Prendiamo $n = 2$, $\theta_0 = 1$, $\theta_1 = 2$, $\alpha = 0,05$. La regione critica è

$$\big\{X_1 + X_2 > \lambda\big\},$$

con λ da determinare usando la relazione

$$P^{\theta_0}(X_1 + X_2 > \lambda) = 0.05.$$

Sotto H_0, $X_1 + X_2$ ha legge Π_2, e le tavole della Π_2 danno

$$\begin{aligned}
P^{\theta_0}(X_1 + X_2 > 0) &= 0,865 \\
P^{\theta_0}(X_1 + X_2 > 1) &= 0,594 \\
P^{\theta_0}(X_1 + X_2 > 2) &= 0,323 \\
P^{\theta_0}(X_1 + X_2 > 3) &= 0,143 \\
P^{\theta_0}(X_1 + X_2 > 4) &= 0,053 \\
P^{\theta_0}(X_1 + X_2 > 5) &= 0,017.
\end{aligned}$$

Dunque il valore di λ che cerchiamo sta tra 4 e 5.

Se vogliamo usare il test deterministico, allora dobbiamo prendere $\lambda = 5$, e la probabilità di errore di prima specie è addirittura $0,017$ (assai minore di quanto richiesto, $\alpha = 0,05$).
Calcoliamo per esercizio anche la potenza, che è $P^{\theta_1}(X_1 + X_2 > 5)$. Sotto H_1, $X_1 + X_2$ ha legge Π_4, e dalle tavole della Π_4 si ricava

$$P^{\theta_1}(X_1 + X_2 > 5) = 0,2149.$$

Questo significa che si ha una probabilità del 21% di accettare a ragione H_1.

Se invece vogliamo utilizzare un test aleatorio, possiamo usare il Teorema di Neyman-Pearson 13.1, cercando Φ del tipo

$$\Phi = 1_{\{p^1 > Cp^0\}} + \gamma 1_{\{p^1 = Cp^0\}} = 1_{\{X_1+X_2>\lambda\}} + \gamma 1_{\{X_1+X_2=\lambda\}}$$

e si deve scegliere $\lambda = f^{\leftarrow}(0,05)$, con $f(\lambda) = P^{\theta_0}(X_1 + X_2 > \lambda)$. Si ha, dai calcoli precedenti,

$$\inf\{\lambda : P^{\theta_0}(X_1 + X_2 > \lambda) \leq 0,05\} = 5,$$

dunque

$$\Phi = 1_{\{X_1+X_2>5\}} + \gamma 1_{\{X_1+X_2=5\}},$$

e γ si trova dalla relazione

$$P^{\theta_0}(X_1 + X_2 > 5) + \gamma P^{\theta_0}(X_1 + X_2 = 5) = 0,05,$$

cioè

$$0,017 + \gamma(0,053 - 0,017) = 0,05$$

che fornisce $\gamma = 0,917$. Dunque il test aleatorio significa che

(i) se $X_1 + X_2 > 5$ si deve respingere H_0;

(ii) se $X_1 + X_2 < 5$ si deve accettare H_0 ;

(iii) se $X_1 + X_2 = 5$, si deve lanciare una moneta che dà "testa"con probabilità 0,917 e respingere (risp. accettare) H_0 se esce "testa" (risp. "croce").

15 Test unilaterali e bilaterali

Supponiamo assegnato un modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$nel quale l'insieme dei parametri Θ è un intervallo (in senso lato) della retta reale.

Definizione 15.1 Si chiama *test unilaterale* un test (per un'ipotesi nulla) della forma $H_0 : \theta \leq \theta_0$ contro $H_1 : \theta > \theta_0$ (dove θ_0 è un elemento fissato di Θ).

Osservazione 15.2 Anche un test del tipo $H_0 : \theta \geq \theta_0$ contro $H_1 : \theta < \theta_0$ è unilaterale nel senso della definizione precedente: basta cambiare il parametro con la trasformazione $\theta \mapsto -\theta$.

Esempio 15.3 *Il test sul controllo di qualità trattato all'inizio è un test unilaterale.*

I test di questo tipo si trattano bene se il modello ha la proprietà seguente.

Definizione 15.4 Supponiamo il modello dominato da una misura μ. Si dice che la famiglia delle verosimiglianze è *a rapporto di verosimiglianza crescente* (risp. *decrescente*) se esiste una v.a. reale T e, per ogni coppia (θ_1, θ_2) con $\theta_1 < \theta_2$ una funzione crescente (risp. decrescente) $f_{\theta_1,\theta_2} : \mathbb{R} \to [0, +\infty]$ tali che

$$\frac{L(\theta_2)}{L(\theta_1)} = f_{\theta_1,\theta_2}(T), \qquad \mu\text{-q.o.}$$

Basta ovviamente considerare il caso di rapporto di verosimiglianza crescente: nel caso decrescente ci si riporta all'altro utilizzando la v. a. $-T$ al posto di T.

Esempio 15.5 (a) Per un modello esponenziale con verosimiglianza $L(\theta) = \exp\big(\theta T - \psi(\theta)\big)$ si ha

$$\frac{L(\theta_2)}{L(\theta_1)} = \exp\big(\psi(\theta_1) - \psi(\theta_2)\big) \cdot \exp\big((\theta_2 - \theta_1)T\big),$$

e la funzione $f_{\theta_1,\theta_2} : x \mapsto \exp\big((\theta_2 - \theta_1)x\big)$ è crescente.
(b) Sia $(X_1, \ldots, X_n)$ un campione di legge $\mathcal{U}\big([0, \theta]\big)$. La verosimiglianza rispetto alla misura di Lebesgue, come sappiamo (ved. Esempio (e) sulle statistiche sufficienti), si può scrivere nella forma

$$L(\theta) = \frac{1}{\theta^n} 1_{\{\max_{1\le i\le n} X_i \le \theta\}}.$$

Sia $T = \max_{1\le i\le n} X_i$; allora

$$\frac{L(\theta_2)}{L(\theta_1)} = \begin{cases} \left(\dfrac{\theta_1}{\theta_2}\right)^n & \text{per} \quad T \le \theta_1 \\ +\infty & \text{per} \quad T > \theta_1. \end{cases},$$

e la funzione

$$f_{\theta_1,\theta_2} : x \mapsto \begin{cases} \left(\dfrac{\theta_1}{\theta_2}\right)^n & \text{per} \quad x \le \theta_1 \\ +\infty & \text{per} \quad x > \theta_1 \end{cases}$$

è chiaramente crescente.

Osservazione 15.6 Come abbiamo appena visto, $L(\theta_1)$ può essere nullo. In tal caso si pone $\frac{a}{0} = +\infty$ se $a > 0$ e $\frac{0}{0} =$ qualsiasi valore ci faccia comodo, dato che l'unica cosa che interessa è che valga la relazione $L(\theta_2) = L(\theta_1) f_{\theta_1,\theta_2}(T)$. Nell'esempio (b) precedente, se $\theta_1 < T < \theta_2$ si ha $\frac{L(\theta_2)}{L(\theta_1)} = \frac{a}{0} = +\infty$. Invece, se $T \ge \theta_2$, si ha $\frac{L(\theta_2)}{L(\theta_1)} = \frac{0}{0}$, e si prende il valore $+\infty$ perché questo è l'unico modo per rendere f_{θ_1,θ_2} crescente.

Per un test unilaterale, se il modello è a rapporto di verosimiglianza crescente, si possono individuare le "buone " regioni critiche. Infatti

Lemma 15.7 *Supponiamo che il modello sia a rapporto di verosimiglianza crescente, e sia*

$$\Phi = 1_{\{T>C\}} + \gamma 1_{\{T=C\}},$$

($C \ge 0$ e $0 \le \gamma \le 1$), dove T è la statistica della Definizione 15.4. Allora, per ogni (θ_1, θ_2) con $\theta_1 < \theta_2$, Φ è un test di Neyman-Pearson per l'ipotesi $H_0 : \theta = \theta_1$ contro $H_1 : \theta = \theta_2$.

DIMOSTRAZIONE. Sia $C^* = f_{\theta_1,\theta_2}(C)$. Dato che f_{θ_1,θ_2} è crescente, si ha

$$\{T > C\} = \big\{f_{\theta_1,\theta_2}(T) > f_{\theta_1,\theta_2}(C)\big\} = \Big\{\frac{L(\theta_2)}{L(\theta_1)} > C^*\Big\} = \big\{L(\theta_2) > C^* L(\theta_1)\big\}$$

e similmente

$$\{T < C\} = \big\{L(\theta_2) < C^* L(\theta_1)\big\}.$$

Pertanto

$$\Phi = 1_{\{L(\theta_2) > C L(\theta_1)\}} + \gamma 1_{\{L(\theta_2) = C L(\theta_1)\}},$$

che è una funzione di test di Neyman-Pearson.

□

Osservazione 15.8 Notiamo che può capitare che sia $C^* = 0$ oppure $C^* = \infty$; per questo nella dimostrazione del Lemma 13.1 di Neyman-Pearson avevamo preso in considerazione anche questi casi.

Osservazione 15.9 Per il Corollario 14.6, si ha $E^{\theta_2}[\Phi] \geq E^{\theta_1}[\Phi]$, e cioè la funzione $\theta \mapsto E^{\theta}[\Phi]$ è crescente.

Teorema 15.10 *(a) Fissati $C \in \mathbb{R}$ e $0 \leq \gamma \leq 1$, il test aleatorio*

$$\Phi = 1_{\{T > C\}} + \gamma 1_{\{T = C\}}$$

è un test per l'ipotesi $H_0 : \theta \leq \theta_0$ contro $H_1 : \theta > \theta_0$, ammissibile, di taglia $E^{\theta_0}[\Phi]$ e UPP tra i test di taglia $E^{\theta_0}[\Phi]$.
(b) Fissato $0 < \alpha < 1$, si possono determinare C e γ in modo tale che la taglia di Φ sia esattamente α.

DIMOSTRAZIONE. Poiché la funzione $\theta \mapsto E^{\theta}[\Phi]$ è crescente, la taglia del test è

$$\sup_{\theta \leq \theta_0} E^{\theta}[\Phi] = E^{\theta_0}[\Phi].$$

Dimostriamo che Φ è UPP e ammissibile. Per ogni $\theta > \theta_0$, Φ è un test di Neyman-Pearson di ipotesi e alternativa semplici. Dunque, per il Teorema di Neyman-Pearson 14.5, è ammissibile e UPP tra i test di taglia $E^{\theta_0}[\Phi]$ (il Teorema di Neyman-Pearson richiede la taglia $E^0[\Phi]$, che, in questo caso, per il Lemma precedente è proprio $E^{\theta_0}[\Phi]$).
(b) Questo punto si dimostra esattamente come nel Teorema di Neyman-Pearson.

□

Osservazione 15.11 Il Teorema precedente dice che in un modello a rapporto di verosimiglianza crescente le "buone " regioni critiche sono del tipo $\{T > C\}$ (caso $\gamma = 0$ oppure $\{T \geq C\}$ (caso $\gamma = 1$= (o, come abbiamo visto, qualcosa di "intermedio", caso $0 < \gamma < 1$).

Osservazione 15.12 Supponiamo di avere una regione critica del tipo $\{T > C\}$ ($\gamma = 0$). Volendo calcolare C in funzione del livello α desiderato, si deve risolvere rispetto a C l'equazione

$$E^{\theta_0}[\Phi] = P^{\theta_0}(T > C) = 1 - P^{\theta_0}(T \leq C) = \alpha,$$

e cioè

$$P^{\theta_0}(T \leq C) = 1 - \alpha,$$

quindi bisogna conoscere la legge di T sotto P^{θ_0}. Se, sotto P^{θ_0}, T ha f.d.r. F_T continua e strettamente crescente (almeno nell'insieme $H_{F_T} = \{x \in \mathbb{R} : 0 < F_T(x) < 1\}$), l'equazione precedente ha l'unica soluzione $C = q_{1-\alpha}$ (quantile di ordine $1 - \alpha$ della legge di T secondo P^{θ_0}).

Osservazione 15.13 (a) Se la famiglia delle verosimiglianze è a rapporto di verosimiglianza decrescente, si prendono test della forma $\Phi = 1_{\{T<C\}} + \gamma 1_{\{T=C\}}$.
(b) Per un test della forma $\theta \geq \theta_0$ contro $H_1 : \theta < \theta_0$ si prenderanno ancora test della forma $\Phi = 1_{\{T<C\}} + \gamma 1_{\{T=C\}}$.
(c) Non è strettamente necessario che Θ sia un intervallo. Ad esempio, se $\Theta = [0,1] \cup [2,+\infty)$, si può ugualmente utilizzare il Teorema precedente per il test $H_0 : \theta \leq 1$ contro $H_1 : \theta \geq 2$.

Esempio 15.14 Consideriamo un campione di taglia n e legge $\mathcal{E}(\theta)$, con $\theta > 0$, e il test di ipotesi $H_0 : \theta \leq 1$. Sappiamo che

$$L(\theta, X_1, \ldots, X_n) = \theta^n \exp\big(-\theta \sum_{i=1}^n X_i \big) 1_{\{X_i>0, \forall i\}}.$$

Quindi

$$\frac{L(\theta_2)}{L(\theta_1)}(X_1, \ldots, X_n) = \begin{cases} \Big(\dfrac{\theta_2}{\theta_1}\Big)^n \exp\big(-(\theta_2 - \theta_1) \displaystyle\sum_{i=1}^n X_i \big) & \text{se} \quad X_i > 0, \forall i \\ \\ +\infty & \text{altrove} \end{cases}$$

è a rapporto di verosimiglianza crescente in $T = -\sum_{i=1}^n X_i$. Pertanto le "buone" regioni critiche sono della forma

$$\Big\{ -\sum_{i=1}^n X_i > C \Big\} = \Big\{ \sum_{i=1}^n X_i < -C \Big\}, \qquad C < 0.$$

Calcoliamo la taglia del test che, per il Teorema 15.10, e posto $A = -C$, è data da

$$P^1\Big(\sum_{i=1}^n X_i < A \Big).$$

Sotto P^1, le v.a. $X_i \sim \mathcal{E}(1)$ e dunque $\sum_{i=1}^n X_i \sim \Gamma(n,1)$. pertanto, integrando successivamente per parti si ha

$$P^1\Big(\sum_{i=1}^n X_i < A \Big) = \frac{1}{(n-1)!} \int_0^A x^{n-1} \mathrm{e}^{-x} \, \mathrm{d}x = -\frac{1}{(n-1)!} x^{n-1} \mathrm{e}^{-x} \Big|_0^A + \frac{1}{(n-2)!} \int_0^A x^{n-2} \mathrm{e}^{-x} \, \mathrm{d}x$$
$$= \cdots = \mathrm{e}^{-A} \Big\{ -\frac{A^{n-1}}{(n-1)!} - \frac{A^{n-2}}{(n-2)!} - \cdots - A \Big\} + 1.$$

Se vogliamo fissare il livello α, si deve allora risolvere rispetto ad A l'equazione

$$= \mathrm{e}^{-A} \Big\{ -\frac{A^{n-1}}{(n-1)!} - \frac{A^{n-2}}{(n-2)!} - \cdots - A \Big\} + 1 = \alpha.$$

Si può dimostrare che questa equazione ammette una e una sola soluzione A, e in tal modo la regione critica è determinata.

Esempio 15.15 Consideriamo un campione come nell'esempio (c) sulle statistiche esaustive. Vogliamo eseguire il test $H_0 : \theta \geq 0$ contro $H_1 : \theta < 0$. Il rapporto di verosimiglianza è

$$\frac{L(\theta_2)}{L(\theta_1)}(X_1, \ldots, X_n) = \begin{cases} \Big(\dfrac{\theta_2 + 1}{\theta_1 + 1}\Big)^n \Big(\displaystyle\prod_{i=1}^n X_i \Big)^{\theta_2 - \theta_1} & \text{se} \quad X_i \in (0,1), \forall i \\ \\ +\infty & \text{altrove,} \end{cases}$$

ed è crescente in $T = \prod_{i=1}^n X_i$. Quindi le buone regioni critiche sono del tipo

$$\Big\{ \prod_{i=1}^n X_i \le C \Big\}.$$

Volendo calcolare C in funzione del livello desiderato α, osserviamo che, sotto P^0, la v.a. $-\log X_i \sim \mathcal{E}(1)$ (verifica per esercizio), e dunque $-\sum_{i=1}^n X_i \sim \Gamma(n,1)$. Dunque si deve risolvere rispetto a C l'equazione

$$\alpha = P^0\Big(\prod_{i=1}^n X_i \le C \Big) = P^0\Big(-\sum_{i=1}^n \log X_i \ge -\log C \Big) = \cdots .$$

Esercizio 15.16 Esaminare il test $H_0 : \theta \ge \frac{1}{2}$ contro $H_1 : \theta < \frac{1}{2}$ per un campione di legge

(i) Π_θ;

(ii) $\mathcal{B}(1,\theta)$. Questo secondo esempio è quello del controllo di qualità (esempio iniziale).

Definizione 15.17 Si chiama *test bilaterale* un test per l'ipotesi della forma $H_0 : \theta \in [\theta_1, \theta_2]$ contro $H_1 : \theta \notin [\theta_1, \theta_2]$, dove $\theta_1 < \theta_2$ sono due punti interni a Θ. Può accadere che $\theta_1 = \theta_2$ ed in tal caso ipotesi e alternativa si scrivono rispettivamente nella forma $H_0 : \theta = \theta_1$, $H_1 : \theta \ne \theta_1$.

Per i test bilaterali vale un risultato meno forte del Teorema sui test unilaterali, (Teorema 15.10), e in condizioni molto più restrittive. Bisogna innanzitutto limitarsi al caso di un modello esponenziale con verosimiglianza $L(\theta) = \exp(\theta T - \psi(\theta))$, e bisogna supporre che, per ogni x, si abbia $\mu(T = x) = 0$ (μ misura dominante del modello).
Occorre poi introdurre un'ulteriore

Definizione 15.18 Un test Φ si dice *corretto* se la sua potenza è non inferiore alla sua taglia, cioè se, comunque si scelgano $\theta_0 \in \Theta_0$ e $\theta_1 \in \Theta_1$, si ha

$$E^{\theta_1}[\Phi] \ge E^{\theta_0}[\Phi].$$

Osservazione 15.19 Abbiamo visto (Corollario 14.6) che i test di Neyman-Pearson sono corretti. Si può dimostrare che sono corretti anche i test del Teorema 15.10.

Vale allora il seguente risultato (di cui omettiamo la dimostrazione, alquanto complessa):

Teorema 15.20 *Nelle ipotesi sopra precisate, assegnato $\alpha \in (0,1)$, si possono determinare due numeri reali $C_1 < C_2$ tali che l'insieme $D = \{T \notin [C_1, C_2]\}$ sia la soluzione del sistema*

$$(i) \qquad \begin{cases} P^{\theta_1}(D) = \alpha \\ P^{\theta_2}(D) = \alpha \end{cases}$$

se $\theta_1 \ne \theta_2$, oppure del sistema

$$(ii) \qquad \begin{cases} P^{\theta_1}(D) = \alpha \\ \\ \dfrac{\mathrm{d}P^{\theta}(D)}{\mathrm{d}\theta}\Big|_{\theta=\theta_1} = 0 \end{cases}$$

se $\theta_1 = \theta_2$.
In tal caso il test (deterministico) di regione critica D è un test corretto, di taglia α, per l'ipotesi $H_0 : \theta \in [\theta_1, \theta_2]$ contro $H_1 : \theta \notin [\theta_1, \theta_2]$, UPP tra i test corretti di taglia α.

Esempio 15.21 Test sulla media di un campione gaussiano con varianza nota.

Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$, dove σ^2 è un valore noto. La verosimiglianza (rispetto alla misura di Lebesgue n-dimensionale) ha la forma

$$L(m) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\Big(- \frac{\sum_{i=1}^n (X_i - m)^2}{2\sigma^2}\Big)$$

e con qualche calcolo si vede che

$$\frac{L(m_2)}{L(m_1)}(X_1, \dots, X_n) = c(m_1, m_2) \exp\Big(\frac{m_2 - m_1}{\sigma^2} \cdot \big(\sum_{i=1}^n X_i\big)\Big) = c(m_1, m_2) \exp\Big(n\frac{m_2 - m_1}{\sigma^2} \cdot \overline{X}\Big),$$

e quindi è a rapporto di verosimiglianza crescente rispetto a $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$.

Per un'ipotesi della forma $H_0 : m \le m_0$ scegliamo un test (deterministico) di regione critica $\{\overline{X} > C\}$, con C tale che $P^{m_0}(\overline{X} > C) = \alpha$. Sotto P^{m_0} la legge di $\overline{X}$ è, come sappiamo, la $\mathcal{N}(m_0, \frac{\sigma^2}{n})$, e quindi

$$\alpha = P^{m_0}(\overline{X} > C) = 1 - P^{m_0}(\overline{X} \le C) = 1 - \Phi\Big(\frac{C - m_0}{\sigma}\sqrt{n}\Big),$$

quindi

$$\frac{C - m_0}{\sigma}\sqrt{n} = \phi_{1-\alpha},$$

da cui si ricava facilmente C.

Osserviamo che lo stesso risultato si ottiene scegliendo un test (ancora deterministico) di regione critica $\{\overline{X} \ge C\}$, poiché gli eventi del tipo $\{\overline{X} = C\}$ sono trascurabili rispetto a P^{m_0}.

Per un test bilaterale del tipo $H_0 : m = m_0$ contro $H_1 : m \ne m_0$ si deve dapprima risolvere in C_1 e C_2 il sistema

$$\begin{cases} P^{m_0}\big(\overline{X} \notin [C_1, C_2]\big) = \alpha \\ \\ \dfrac{\mathrm{d}}{\mathrm{d}m} P^m\big(\overline{X} \notin [C_1, C_2]\big)\Big|_{m=m_0} = 0. \end{cases}$$

Notiamo che

$$P^m\big(\overline{X} \notin [C_1, C_2]\big) = P^m\Big(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \notin \Big[\frac{C_1 - m}{\frac{\sigma}{\sqrt{n}}}, \frac{C_2 - m}{\frac{\sigma}{\sqrt{n}}}\Big]\Big) = \Phi\Big(\frac{C_1 - m}{\frac{\sigma}{\sqrt{n}}}\Big) + 1 - \Phi\Big(\frac{C_2 - m}{\frac{\sigma}{\sqrt{n}}}\Big),$$

(Φ = f.d.r. della Normale standard). Dunque

$$\frac{\mathrm{d}}{\mathrm{d}m} P^m\big(\overline{X} \notin [C_1, C_2]\big) = -\frac{\sqrt{n}}{\sigma}\varphi\Big(\frac{C_1 - m}{\frac{\sigma}{\sqrt{n}}}\Big) + \frac{\sqrt{n}}{\sigma}\varphi\Big(\frac{C_2 - m}{\frac{\sigma}{\sqrt{n}}}\Big),$$

(φ = densità Normale standard). Quindi l'equazione

$$\frac{\mathrm{d}}{\mathrm{d}m} P^m\big(\overline{X} \notin [C_1, C_2]\big)\Big|_{m=m_0} = 0$$

diventa

$$\varphi\Big(\frac{C_1 - m_0}{\frac{\sigma}{\sqrt{n}}}\Big) = \varphi\Big(\frac{C_2 - m_0}{\frac{\sigma}{\sqrt{n}}}\Big).$$

L'equazione $\varphi(x) = \varphi(y)$ ha le soluzioni $x = y$ e $x = -y$, che per noi significano rispettivamente $C_1 = C_2$ e $C_1 + C_2 = 2m_0$. La prima di queste relazioni non è utilizzabile (altrimenti sarebbe $P^{m_0}\big(\overline{X} \notin [C_1, C_1]\big) = 1 > \alpha$). La seconda equivale a $C_1 = m_0 - C$ e $C_2 = m_0 + C$, con $C \in \mathbb{R}^+$. A questo punto si può determinare C (per mezzo delle tavole della Normale standard) dalla relazione

$$\Phi\Big(\frac{C_1 - m_0}{\frac{\sigma}{\sqrt{n}}}\Big) + 1 - \Phi\Big(\frac{C_2 - m_0}{\frac{\sigma}{\sqrt{n}}}\Big) = \alpha,$$

ovvero

$$2\left\{1 - \Phi\left(\frac{C}{\frac{\sigma}{\sqrt{n}}}\right)\right\} = \alpha,$$

che ha la soluzione

$$C = \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}.$$

Osservazione 15.22 Considerato il tipo di calcoli che abbiamo fatto, spesso, invece della statistica $\overline{X}$, si prende come statistica del test

$$\frac{\overline{X} - m_0}{\sigma}\sqrt{n}.$$

Si trova allora la regione critica

$$\Big\{\Big|\frac{\overline{X} - m_0}{\sigma}\sqrt{n}\Big| > \phi_{1-\frac{\alpha}{2}}\Big\}.$$

Esempio 15.23 Test sulla varianza di un campione gaussiano con media nota.

Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$, dove m è un valore noto. La verosimiglianza (rispetto alla misura di Lebesgue n-dimensionale) è

$$L(\sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\Big(-\frac{\sum_{i=1}^n (X_i - m)^2}{2\sigma^2}\Big)$$

e con qualche calcolo si vede che il rapporto di verosimiglianza si scrive nella forma

$$\frac{L(\sigma_2^2)}{L(\sigma_1^2)}(X_1, \dots, X_n) = \exp\left(-n\log\frac{\sigma_2}{\sigma_1} + \sum_{i=1}^n (X_i - m)^2\Big(\frac{1}{2\sigma_1^2} - \frac{1}{\sigma_2^2}\Big)\right)$$

ed è crescente rispetto a $T = \sum_{i=1}^n (X_i - m)^2$. Per un test unilaterale del tipo $H_0 : \sigma^2 \le \sigma_0^2$ scegliamo dunque una regione critica del tipo $\{\sum_{i=1}^n (X_i - m)^2 > C\}$, con $P^{\sigma_0^2}(\sum_{i=1}^n (X_i - m)^2 > C) = \alpha$. Ricordando che la v.a. $\sum_{i=1}^n (X_i - m)^2$, sotto $P^{\sigma_0^2}$ ha legge $\sigma_0^2 \cdot \chi^2(n)$, si trova

$$\alpha = P^{\sigma_0^2}\big(\sum_{i=1}^n (X_i - m)^2 > C\big) = P^{\sigma_0^2}\Big(\frac{\sum_{i=1}^n (X_i - m)^2}{\sigma_0^2} > \frac{C}{\sigma_0^2}\Big) = 1 - F_n\Big(\frac{C}{\sigma_0^2}\Big),$$

(F_n = f.d.r. della $\chi^2(n)$) e quindi, dalle tavole dei quantili della $\chi^2(n)$, si ottiene

$$C = \sigma_0^2 \cdot \chi^2_{1-\alpha}(n).$$

Per un test bilaterale del tipo $H_0 : \sigma^2 = \sigma_0^2$, si prende una regione critica della forma $\{\sum_{i=1}^n (X_i - m)^2 \notin [C_1, C_2]\}$, dove, per il Teorema 15.20, C_1 e C_2 sono determinate dalle relazioni (α= livello desiderato)

$$\begin{cases} P^{\sigma_0^2}\big(\sum_{i=1}^n (X_i - m)^2 \notin [C_1, C_2]\big) = \alpha \\ \\ C_1^{\frac{n}{2}}\exp\Big(-\frac{C_1}{2\sigma_0^2}\Big) = C_2^{\frac{n}{2}}\exp\big(-\frac{C_2}{2\sigma_0^2}\Big), \end{cases}$$

ma il calcolo effettivo è quasi impossibile.

16 Test in presenza di un parametro fantasma

Supponiamo che, in un problema di test, Θ sia un prodotto cartesiano della forma $\Theta = \Lambda \times M$, ed inoltre che sia $\Theta_0 = \Lambda_0 \times M$, $\Theta_1 = \Lambda_1 \times M$, dove Λ_0 e Λ_1 sono una partizione di Λ; nel parametro del modello statistico $\theta = (\lambda, m)$ la componente m si chiama *parametro fantasma* (o anche *importuno*) per il problema di test indicato, e non è noto.

La teoria generale in questa situazione è piuttosto complicata. Faremo quindi solo due esempi, di largo impiego.

Come regola generale diciamo solo che conviene cercare una statistica la cui legge non dipenda da m (ovviamente) e sia diversa per valori diversi di λ. In questo modo spesso il problema diventa semplice.

Esempio 16.1 Test(di Fisher-Snedecor) sulla varianza di un campione gaussiano, con media sconosciuta.

Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{N}(m, \sigma^2)$. Vogliamo effettuare un test unilatero per l'ipotesi $H_0 : \sigma^2 \le \sigma_0^2$ contro $H_1 : \sigma^2 > \sigma_0^2$.

Ricordiamo che, per il Teorema 9.5 (ii) la v.a.

$$T = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

ha legge $\sigma^2 \cdot \chi^2(n-1)$, la cui densità è

$$f^{\sigma^2}(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} \cdot \left(\frac{1}{2}\right)^{\frac{n-1}{2}} \sigma^{-(n-1)} x^{\frac{n-1}{2}-1} \mathrm{e}^{-\frac{x}{2\sigma^2}}.$$

Si ha in particolare

$$\frac{f^{\sigma_2^2}}{f^{\sigma_1^2}}(x) = c(\sigma_1, \sigma_2) \mathrm{e}^{\frac{x}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)},$$

e quindi questo modello è a rapporto di verosimiglianza crescente rispetto alla statistica $X = T$. Il test $H_0 : \sigma^2 \le \sigma_0^2$ contro $\sigma^2 > \sigma_0^2$ basato sulla statistica T avrà pertanto una regione critica del tipo

$$\{T > C\} = \Big\{ \sum_{i=1}^{n} (X_i - \overline{X})^2 > C \Big\},$$

dove C è scelto, in funzione dellla taglia α desiderata, in modo tale che

$$P^{\sigma_0^2}\Big(\sum_{i=1}^{n} (X_i - \overline{X})^2 > C \Big) = \alpha.$$

Dato che, come abbiamo detto sopra, sotto $P^{\sigma_0^2}$ la v.a. $\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2}$ ha legge $\chi^2(n-1)$, si può scrivere

$$\alpha = P^{\sigma_0^2}\Big(\sum_{i=1}^{n} (X_i - \overline{X})^2 > C \Big) = P^{\sigma_0^2}\Big(\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2} > \frac{C}{\sigma_0^2} \Big) = 1 - F_{n-1}\Big(\frac{C}{\sigma_0^2} \Big),$$

dove con F_{n-1} indichiamo la f.d.r. della legge $\chi^2(n-1)$. Si trova dunque $C = \sigma_0^2 \cdot \chi^2_{1-\alpha}(n-1)$.

Esempio 16.2 Test(di Student) sulla media di un campione gaussiano, con varianza sconosciuta.

Premettiamo una

Definizione 16.3 Si chiama *legge di Student a n gradi di libertà decentrata di a* la legge di una v.a. del tipo

$$\sqrt{n}\frac{X}{\sqrt{Y}},$$

dove X la legge $\mathcal{N}(a,1)$, Y ha legge $\chi^2(n)$, e X e Y sono indipendenti.

Osservazione 16.4 Si può verificare (ma omettiamo i calcoli, che sono laboriosi) che queste leggi sono a rapporto di verosimiglianza crescente rispetto al parametro a.

Consideriamo allora un campione $(X_1,\dots,X_n)$ di legge $\mathcal{N}(m,\sigma^2)$. Vogliamo eseguire un test per l'ipotesi $H_0 : m \le 0$ contro $H_1 : m > 0$, basato sulla statistica

$$T = \sqrt{n}\frac{\overline{X}}{\sqrt{S^2}}.$$

Possiamo scrivere T nella forma

$$T = \sqrt{n}\frac{\overline{X}}{\sqrt{S^2}} = \frac{\frac{\overline{X}}{\sigma}\sqrt{n}}{\sqrt{\frac{S^2(n-1)}{\sigma^2}}}\sqrt{n-1};$$

osserviamo poi che, per il Teorema 9.5, si ha

$$\frac{\overline{X}}{\sigma}\sqrt{n} \sim \mathcal{N}\Big(\frac{\sqrt{n}}{\sigma}m, 1\Big), \qquad \frac{S^2(n-1)}{\sigma^2} \sim \chi^2(n-1),$$

ed inoltre le due v.a. indicate sopra sono tra loro indipendenti. Si ottiene allora che T ha legge di Student a $n-1$ gradi di libertà decentrata di $\frac{\sqrt{n}}{\sigma}m$.
Conviene prendere come parametri della legge normale $\big(\frac{m}{\sigma},\sigma^2\big)$ invece di (m,σ^2); l'ipotesi e l'alternativa diventano allora rispettivamente $H_0 : \frac{m}{\sigma} \le 0$ e $H_1 : \frac{m}{\sigma} > 0$, e σ^2 è un parametro fantasma. Tuttavia, utilizzando la statistica T (la cui legge dipende solo da $\frac{m}{\sigma}$), il test diventa un test unilaterale, e quindi avrà regione critica del tipo $\{T > C\}$; per calcolare C in funzione della taglia desiderata α, si tiene presente che, in base alla teoria, la taglia è

$$P^{\frac{m}{\sigma}=0}(T > C),$$

e che, se $\frac{m}{\sigma} = 0$, allora T ha legge di Student a $n-1$ gradi di libertà; C si ricava con i soliti conti.

Se il test è $H_0 : m \le m_0$ contro $H_1 : m > m_0$, si osserva che, se il campione $(X_1,\dots,X_n)$ ha legge $\mathcal{N}(m,\sigma^2)$, allora il vettore $(Y_1,\dots,Y_n) = (X_1 - m_0,\dots,X_n - m_0)$ è un campione di legge $\mathcal{N}(m-m_0,\sigma^2)$, e ci si può ricondurre al caso precedente. Osservando che $\overline{Y} = \overline{X} - m_0$ e che

$$S_Y^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n \big((X_i - m_0) - (\overline{X} - m_0)\big)^2}{n-1} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} = S^2$$

la statistica del test è

$$T = \frac{\overline{Y}}{S_Y}\sqrt{n} = \frac{\overline{X} - m_0}{S}\sqrt{n}, \tag{28}$$

e si avrà così una regione critica della forma

$$\Big\{\frac{\overline{X} - m_0}{S}\sqrt{n} > C\Big\}.$$

Osservazione 16.5 Notare che la forma (28) della statistica T è simile a quella del test con varianza nota (ved. Osservazione 15.22); la differenza consiste nel fatto che la varianza σ^2, che adesso non è nota, viene sostituita dal suo stimatore corretto S^2.

17 Test del rapporto di verosimiglianza

Sia $H_0 : \theta = 0$ contro $H_1 : \theta = 1$ un test a ipotesi e alternativa semplici. Abbiamo visto dal Lemma di Neyman-Pearson 13.1 che in questo caso una buona regione critica è $\{p^0 < Cp^1\}$. In generale

Definizione 17.1 Assegnate l'ipotesi $H_0 : \theta \in \Theta_0$ contro l'alternativa $H_1 : \theta \in \Theta_1$, e supposto che le due variabili $\sup_{\theta\in\Theta_0} L(\theta)$ e $\sup_{\theta\in\Theta_1} L(\theta)$ siano misurabili, si chiama *test del rapporto di verosimiglianza* un test di regione critica

$$\big\{ \sup_{\theta\in\Theta_0} L(\theta) < C \sup_{\theta\in\Theta_1} L(\theta)\big\}.$$

Se $\sup_{\theta\in\Theta_1} L(\theta)$ è sempre strettamente positiva, allora la regione critica si può scrivere nella forma

$$\Big\{\frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta_1} L(\theta)} < C\Big\},$$

e questo spiega il nome di *rapporto di verosimiglianza.*

Osservazione 17.2 La giustificazione per la scelta di una regione critica di questo tipo è simile a quella che abbiamo dato a suo tempo per i test di Neyman-Pearson: se è vera l'alternativa, significa che almeno una delle $L(\theta)$, con $\theta \in \Theta_1$ è grande, e quindi il $\sup_{\theta\in\Theta_1} L(\theta)$ è grande, e dunque il rapporto di verosimiglianza è piccolo.

Osservazione 17.3 Se l'ipotesi è semplice ($H_0 : \theta = \theta_0$) e se $\theta \mapsto L(\theta, \omega)$ è continua per ogni $\omega \in \Omega$, in molti casi si può mostrare che la regione critica diventa

$$\big\{L(\theta_0) < C \sup_{\theta\in\Theta} L(\theta)\big\},$$

e quindi, se si ha a disposizione uno stimatore di massima verosimiglianza $\hat\theta$, essa diventa

$$\big\{L(\theta_0) < CL(\hat\theta)\big\}$$

Anche se le proprietà di questo test non sono ben chiare, esso in generale è facile da costruire (come mostra l'osservazione precedente) e porta a delle procedure che si rivelano soddisfacenti in molti casi particolari; si potrebbe anche mostrare che, sotto opportune condizioni, ha delle buone proprietà asintotiche.

Esempio 17.4 Ancora il test di Student sulla media di un campione gaussiano, con varianza sconosciuta.

Su un campione $(X_1, \dots, X_n)$ di legge $\mathcal{N}(m, \sigma^2)$ vogliamo verificare l'ipotesi $H_0 : m = 0$ contro l'alternativa $H_1 : m \neq 0$. Messo in questa forma, non si tratta di un test bilaterale (perché c'è anche il parametro fantasma σ^2); lo diventa se si segue la procedura già usata precedentemente nel caso dell'ipotesi $H_0 : m \leq 0$, ma è piuttosto complicato. Qui seguiremo lo schema del test del rapporto di verosimiglianza.

Partiamo allora dalla verosimiglianza rispetto alla misura di Lebesgue n-dimensionale che è

$$L(m, \sigma^2; X_1, \dots, X_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\Big(- \frac{\sum_{i=1}^n (X_i - m)^2}{2\sigma^2}\Big)$$

per calcolare

$$\sup_{\sigma\in\mathbb{R}^+} L(0, \sigma^2), \qquad \sup_{m\in\mathbb{R},\sigma\in\mathbb{R}^+} L(m, \sigma^2).$$

Questo si può fare facilmente usando gli stimatori di massima verosimiglianza che abbiamo trovato a suo tempo (ved. Esempio 8.5):

$$\sup_{\sigma\in\mathbb{R}^+} L(0,\sigma^2) = L\Big(0, \frac{\sum_{i=1}^n X_i^2}{n}\Big) = \frac{1}{(2\pi)^{\frac{n}{2}}}\cdot\frac{1}{\Big\{\big(\frac{\sum_{i=1}^n X_i^2}{n}\big)^{\frac{1}{2}}\Big\}^n}\exp\left(-\frac{\sum_{i=1}^n X_i^2}{2(\frac{\sum_{i=1}^n X_i^2}{n})}\right)$$
$$= \frac{1}{(2\pi)^{\frac{n}{2}}}\cdot\exp\Big(-\frac{n}{2}\Big)\Big(\frac{\sum_{i=1}^n X_i^2}{n}\Big)^{-\frac{n}{2}}$$

$$\sup_{m\in\mathbb{R},\sigma\in\mathbb{R}^+} L(m,\sigma^2) = L\Big(\overline{X}, \frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n}\Big) = \frac{1}{(2\pi)^{\frac{n}{2}}}\cdot\frac{1}{\Big\{\big(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n}\big)^{\frac{1}{2}}\Big\}^n}\exp\left(-\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{2(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n})}\right)$$
$$= \frac{1}{(2\pi)^{\frac{n}{2}}}\cdot\exp\Big(-\frac{n}{2}\Big)\Big(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n}\Big)^{-\frac{n}{2}},$$

e il rapporto di verosimiglianza diventa di conseguenza

$$\frac{\sup_{\sigma\in\mathbb{R}^+} L(0,\sigma^2)}{\sup_{m\in\mathbb{R},\sigma\in\mathbb{R}^+} L(m,\sigma^2)} = \Big(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sum_{i=1}^n X_i^2}\Big)^{\frac{n}{2}} = \Big(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sum_{i=1}^n (X_i-\overline{X})^2 + n\overline{X}^2}\Big)^{\frac{n}{2}}$$
$$= \left(n\left\{\frac{\overline{X}}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}}\right\}^2 + 1\right)^{-\frac{n}{2}},$$

e la regione critica è

$$\Big\{\frac{\sup_{\sigma\in\mathbb{R}^+} L(0,\sigma^2)}{\sup_{m\in\mathbb{R},\sigma\in\mathbb{R}^+} L(m,\sigma^2)} < C\Big\} = \left\{\left(n\left\{\frac{\overline{X}}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}}\right\}^2 + 1\right)^{-\frac{n}{2}} < C\right\}$$
$$= \left\{\frac{|\overline{X}|}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}} > a\right\},$$

con $a = a(C)$ opportuno. Per calcolare a (e di conseguenza C) in funzione del livello α desiderato, osserviamo che

$$\frac{\overline{X}}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}} = \frac{1}{\sqrt{n(n-1)}}\cdot\underbrace{\frac{\sqrt{n-1}}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}}}_{=\frac{1}{S}}(\sqrt{n}\ \overline{X}) = \frac{1}{\sqrt{n(n-1)}}\Big(\sqrt{n}\cdot\frac{\overline{X}}{S}\Big),$$

e, sotto l'ipotesi $H_0 : m = 0$, la v.a. $\sqrt{n}\cdot\frac{\overline{X}}{S}$ ha legge di Student a $n-1$ gradi di libertà (per il Teorema 9.5 , $\sqrt{n}\cdot\frac{\overline{X}-m}{S}$ ha legge $t(n-1)$ e sotto l'ipotesi $m=0$ si tratta appunto della variabile $\sqrt{n}\cdot\frac{\overline{X}}{S}$. Ved. anche Esempio 16.2).

Esempio 17.5 Sia μ^θ $(0 < \theta < +\infty)$ la legge definita su $\mathbb{R}^+$, continua rispetto alla misura di Lebesgue con densità

$$f^\theta(x) = \mathrm{e}^{-(x-\theta)}1_{[\theta,+\infty)}(x)$$

(si tratta di una densità esponenziale di parametro 1 traslata di θ e può essere interpretata come la durata di una lampadina accesa all'istante θ).

Consideriamo un campione $(X_1, \dots, X_n)$di legge μ^θ; la verosimiglianza rispetto alla misura di Lebesgue n-dimensionale è

$$L(\theta; X_1, \dots, X_n) = \prod_{i=1}^n \mathrm{e}^{-(X_i-\theta)} 1_{[\theta,+\infty)}(X_i) = \mathrm{e}^{(-\sum_{i=1}^n X_i + n\theta)} 1_{[\theta,+\infty)}(\min_{1\le i\le n} X_i).$$

Consideriamo il test del rapporto di verosimiglianza per l'ipotesi $H_0 : \theta = 1$ contro $\theta \neq 1$. La regione critica è

$$\big\{L(1) < C \sup_{\theta\neq 1} L(\theta)\big\}.$$

Per calcolare $\sup_{\theta\neq 1} L(\theta)$ si può fare un calcolo diretto (studio della funzione $\theta \mapsto L(\theta)$) oppure osservare che lo stimatore di massima verosimiglianza del parametro θ è $\hat\theta = \min_{1\le i\le n} X_i$, e quindi

$$\sup_{\theta\neq 1} L(\theta) = \sup_{0<\theta<+\infty} L(\theta) = L(\hat\theta) = \mathrm{e}^{n(\min_{1\le i\le n} X_i) - \sum_{i=1}^n X_i}.$$

Quindi

$$\frac{L(1)}{\sup_{\theta\neq 1} L(\theta)} = \begin{cases} 0 & \text{se } \min_{1\le i\le n} X_i < 1 \\ \mathrm{e}^{n(1-\min_{1\le i\le n} X_i)} & \text{se } \min_{1\le i\le n} X_i \ge 1. \end{cases}$$

La regione critica è

$$D = \Big\{\frac{L(1)}{\sup_{\theta\neq 1} L(\theta)} < C\Big\};$$

se $C = 0$ si ha $D = \emptyset$; se $C > 1$ si ha invece $D = \Omega$ (perché $\sup_{\theta\neq 1} L(\theta) = \sup_{0<\theta<+\infty} L(\theta) \ge L(1)$). Quindi ha senso considerare solo i valori di C con $0 < C < 1$, e in tal caso si ha

$$\begin{aligned} D &= \big\{ \min_{1\le i\le n} X_i < 1\big\} \cup \big\{ \min_{1\le i\le n} X_i \ge 1, \mathrm{e}^{n(1-\min_{1\le i\le n} X_i)} < C\big\} \\ &= \big\{ \min_{1\le i\le n} X_i < 1\big\} \cup \Big\{ \min_{1\le i\le n} X_i \ge 1 - \frac{\log C}{n}\Big\}. \end{aligned}$$

Si può calcolare C in base alla taglia α desiderata conoscendo la legge di $\min_{1\le i\le n} X_i$ sotto P^1, che è l'oggetto del seguente

Esercizio 17.6 Mostrare che sotto P^θ la densità di $\min_{1\le i\le n} X_i$ è data da

$$g^\theta(x) = n\mathrm{e}^{-n(x-\theta)} 1_{[\theta,+\infty)}.$$

Soluzione. Basta osservare che, se $x < \theta$ si ha $P(\min_{1\le i\le n} X_i > x) = 0$, mentre, se $x \ge \theta$

$$P(\min_{1\le i\le n} X_i > x) = P\Big(\bigcap_{i=1}^n \{X_i > x\}\Big) = \prod_{i=1}^n P(X_i > x) = \prod_{i=1}^n \int_x^{+\infty} \mathrm{e}^{-(t-\theta)}\,\mathrm{d}t = \mathrm{e}^{-n(x-\theta)},$$

e si ottiene la densità cercata "per derivazione".

Esercizio 17.7 Si consideri un campione $(X_1, \dots, X_n)$ di legge $\mathcal{E}(\theta)$, con $\theta \in (0, +\infty)$ e si studi il test $H_0 : \theta = 1$ contro $H_1 : \theta \neq 0$ con il metodo dei test bilaterali e con il metodo del rapporto di verosimiglianza.

Il primo metodo suggerisce una regione critica della forma

$$\big\{ \sum_{i=1}^n X_i \le a\big\} \cup \big\{ \sum_{i=1}^n X_i \ge b\big\},$$

dove a e b sono legati dall'equazione

$$a^{n-1}\mathrm{e}^{-a} = b^{n-1}\mathrm{e}^{-b}.$$

Il test del rapporto di verosimiglianza dà invece una regione critica del tipo

$$\{\sum_{i=1}^{n} X_i \le c\} \cup \{\sum_{i=1}^{n} X_i \ge d\},$$

con c e d tali che

$$c^n\mathrm{e}^{-c} = d^n\mathrm{e}^{-d}.$$

18 Cenni all'Analisi Della Varianza (ANalysis Of VAriance = ANOVA)

Supponiamo di avere vari campioni di legge gaussiana, tra loro indipendenti: $(X_{1,1}, \dots, X_{1,n_1})$, campione di taglia n_1 e legge $\mathcal{N}(m_1, \sigma_1^2)$, $(X_{2,1}, \dots, X_{2,n_2})$ campione di taglia n_2 e legge $\mathcal{N}(m_2, \sigma_2^2)$, e così via.

Prende il nome di *analisi della varianza* quella parte della statistica che costruisce dei test per verificare delle ipotesi riguardanti i parametri dei diversi campioni (ad esempio l'uguaglianza delle medie). I campioni sono pensati estratti da popolazioni differenti, e l'ANOVA viene impiegata per confrontare tra loro queste popolazioni.

Diciamo subito che si impiegano le leggi normali sia perché in mancanza di ulteriori informazioni sono le più usate in statistica (grazie per esempio al Teorema Limite Centrale), sia perché in ipotesi diverse i conti diventano impossibili.

L'ANOVA è un capitolo lungo e molto complicato; qui ci limitiamo ad illustrare due esempi classici.

Esempio 18.1 Il problema di Behrens-Fisher (Test per l'uguaglianza delle medie per campioni indipendenti).

Siano $(X_1, \dots, X_p)$ e $(Y_1, \dots, Y_q)$ due campioni indipendenti di leggi rispettivamente $\mathcal{N}(m_1, \sigma_1^2)$ e $\mathcal{N}(m_2, \sigma_2^2)$. Vogliamo costruire un test per l'ipotesi $H_0 : m_1 = m_2$ contro $H_1 : m_1 \neq m_2$. La soluzione generale del problema è decisamente difficile. Un caso trattabile (anche se non realistico) è quello in cui $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (sconosciuta).

Siano

$$\overline{X} = \frac{\sum_{i=1}^{p} X_i}{p}, \qquad \overline{Y} = \frac{\sum_{i=1}^{q} Y_i}{q}$$

le medie campionarie dei due campioni. Per l'Esercizio 9.7 (iv), la v.a.

$$Z_{p,q} = \frac{\sqrt{p+q-2}\cdot\{(\overline{X}-\overline{Y})\}}{\sqrt{\frac{1}{p}+\frac{1}{q}}\cdot\sqrt{\sum_{i=1}^{p}(X_i-\overline{X})^2+\sum_{i=1}^{q}(Y_i-\overline{Y})^2}}$$

ha legge di Student a $p+q-2$ gradi di libertà decentrata di $\frac{m_1-m_2}{\sigma\sqrt{\frac{1}{p}+\frac{1}{q}}}$ Abbiamo visto a proposito del test di Student (Esempio 16.2) che queste leggi sono a rapporto di verosimiglianza crescente rispetto a $(m_1 - m_2)$; ancora il test di Student ci suggerisce come proseguire: per il test $H_0 : m_1 \le m_2$ sceglieremo una regione critica del tipo $\{Z_{p,q} > C\}$, per il test $H_0 : m_1 = m_2$ una regione critica del tipo $\{|Z_{p,q}| > C\}$.
Per calcolare C in funzione del livello desiderato, ricordiamo che se $m_1 = m_2$, $Z_{p,q}$ ha legge $t(p+q-2)$.

All'esempio successivo premettiamo una

Definizione 18.2 Assegnati i due numeri interi positivi n_1 e n_2, si chiama *legge di Fisher-Snedecor* $F(n_1, n_2)$ la legge della v.a.

$$\frac{\frac{Z_1}{n_1}}{\frac{Z_2}{n_2}},$$

dove Z_1 e Z_2 sono due v.a. indipendenti e di leggi rispettive $\chi^2(n_1)$ e $\chi^2(n_2)$.

L'espressione della densità è piuttosto complicata; sono state compilate delle tavole della funzione di ripartizione. Alcuni autori chiamano legge di Fisher-Snedecor la legge di $\frac{Z_1}{Z_2}$, altri ancora quella di $\frac{\sqrt{Z_1}}{\sqrt{Z_2}}$. Si tratta naturalmente di convenzioni, a cui bisogna fare attenzione al momento di consultare le tavole.

L'esempio che segue è molto generale, ma servirà per trattare in maniera concisa l'esempio annunciato.

Esempio 18.3 Sia $X = (X_1, \ldots, X_n)$ una v.a. vettoriale di legge $\mathcal{N}_n(m, \sigma^2 I_n)$, dove $m \in E$ (E è un sottospazio di $\mathbb{R}^n$ di dimensione $k < n$) e σ qualunque. Consideriamo un sottospazio H di E di dimensione $r < k$ e il test di ipotesi $H_0 : m \in H$ contro $H_1 : m \in E \setminus H$. Vogliamo trovare la "buona" regione critica.

È conveniente rappresentare X nella forma $X = m + Y$, dove Y ha legge $\mathcal{N}_n(0, \sigma^2 I_n)$. Per il Teorema di Cochran 9.3, i vettori aleatori $Y - Y_E$, $Y_E - Y_H$ e Y_H (dove Y_E= proiezione di Y su E e Y_H= proiezione di Y su H) sono indipendenti, ed inoltre le v.a.

$$\frac{\|Y - Y_E\|^2}{\sigma^2}, \qquad \frac{\|Y_E - Y_H\|^2}{\sigma^2}$$

hanno legge $\chi^2(n-k)$ e $\chi^2(k-r)$ rispettivamente. Dunque la v.a.

$$Z_Y := \frac{\frac{\|Y_E - Y_H\|^2}{k-r}}{\frac{\|Y - Y_E\|^2}{n-k}}$$

ha legge $F(k-r, n-k)$.

Tornando al test, osserviamo che, dato che $m \in E$, si ha $m_E = m$, e quindi

$$X - X_E = (Y + m) - (Y_E + m) = Y - Y_E$$

e

$$X_E - X_H = (Y_E + m) - (Y_H + m_H) = (Y_E - Y_H) + (m - m_H).$$

La seconda di queste relazioni suggerisce che, sotto l'alternativa $H_1 : m \in E \setminus H$, bisognerà attendersi che $\|X_E - X_H\|^2$ sia tanto più grande quanto più grande è $\|m - m_H\|^2$. Questo suggerisce una regione critica del tipo

$$\left\{ Z_X := \frac{\frac{\|X_E - X_H\|^2}{k-r}}{\frac{\|X - X_E\|^2}{n-k}} > C \right\}.$$

La costruzione è solo intuitiva, ma si potrebbe renderla rigorosa dimostrando che le leggi delle v.a.

$$Z_X := \frac{\frac{\|X_E - X_H\|^2}{k-r}}{\frac{\|X - X_E\|^2}{n-k}}$$

(che dipendono dal parametro $\|m - m_H\|$) sono a rapporto di verosimiglianza crescente rispetto a $\|m - m_H\|$.
Per calcolare il numero C in funzione del livello desiderato, si tiene presente che, sotto l'ipotesi $H_0 : m \in H$ (che implica $m - m_H = 0$), Z_X coincide con Z_Y, e quindi ha legeg $F(k - r, n - k)$.

Le considerazioni precedenti ci serviranno per costruire il cosiddetto *test di omogeneità*, che è la generalizzazione del problema di Behrens-Fisher al caso di più campioni (ma il caso di due soli campioni si sa risolvere anche senza supporre uguali le varianze).

Esempio 18.4 Il test di omogeneità. Siano $(X_{1,1}, \dots, X_{1,n_1})$, ... ,$(X_{k,1}, \dots, X_{k,n_k})$ k campioni indipendenti di legge rispettivamente $\mathcal{N}(m_1, \sigma^2), \dots\dots, \mathcal{N}(m_k, \sigma^2)$ (σ sconosciuta). Consideriamo il test $H_0 : m_1 = m_2 = \cdots = m_k$. Indichiamo con X il campione globale

$$X = (X_{1,1}, \dots, X_{1,n_1}, \dots\dots, X_{k,1}, \dots, X_{k,n_k})$$

(di numerosità $n = n_1 + \cdots + n_k$) e poniamo

$$\overline{X} = \frac{\sum_{i,j} X_{i,j}}{n}; \quad \overline{X}_i = \frac{\sum_{j=1}^{n_i} X_{i,j}}{n_i}, \quad i = 1, 2, \dots, k.$$

La v.a. $\overline{X}$ è la media campionaria globale (media campionaria di X), mentre la $\overline{X}_i$ è la media campionaria all'interno del gruppo i-esimo (media campionaria di $X_i = (X_{i,1}, \dots, X_{i,n_i})$).
La media del vettore X è $(\underbrace{m_1, \dots, m_1}_{n_1 volte}, \underbrace{m_2, \dots, m_2}_{n_2 volte}, \dots, \underbrace{m_k, \dots, m_k}_{n_k volte})$, e quindi appartiene al sottospazio E di dimensione k generato da $\eta_1, \eta_2, \dots, \eta_k$ dove

$$\eta_1 = \Big(\underbrace{\frac{1}{\sqrt{n_1}}, \frac{1}{\sqrt{n_1}}, \dots, \frac{1}{\sqrt{n_1}}}_{n_1 volte}, 0, 0, \dots, 0, 0 \Big); \quad \eta_2 = \Big(\underbrace{0, 0, \dots, 0}_{n_1 volte} \underbrace{\frac{1}{\sqrt{n_2}}, \frac{1}{\sqrt{n_2}}, \dots, \frac{1}{\sqrt{n_2}}}_{n_2 volte}, 0, \dots, 0 \Big)$$

e così via.
Se l'ipotesi H_0 è vera, la media di X è $(\underbrace{m_1, \dots, m_1}_{n\, volte})$, e quindi appartiene al sottospazio 1-dimensionale H di E generato dal vettore

$$\eta = \Big(\underbrace{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}}_{n\, volte} \Big).$$

Si verifica poi facilmente che

$$X_E = \overline{X}_1 \sqrt{n_1} \eta_1 + \cdots + \overline{X}_k \sqrt{n_k} \eta_k, \qquad X_H = \overline{X} \sqrt{n} \eta. \tag{29}$$

Infatti la prima uguaglianza segue da

$$X_E = \sum_{i=1}^{k} \langle X, \eta_i \rangle \eta_i = \sum_{i=1}^{k} \frac{\langle X, \eta_i \rangle}{\sqrt{n_i}} \sqrt{n_i} \eta_i = \sum_{i=1}^{k} \overline{X}_i \sqrt{n_i} \eta_i$$

perché, per ogni $i = 1, 2, \dots, k$

$$\frac{\langle X, \eta_i \rangle}{\sqrt{n_i}} = \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \frac{X_{i,j}}{\sqrt{n_i}} = \sum_{j=1}^{n_i} \frac{X_{i,j}}{n_i} = \overline{X}_i.$$

La seconda uguaglianza si dimostra in modo analogo.
Notiamo che le due equazioni (29) in modo esplicito significano rispettivamente

$$X_E = (\underbrace{\overline{X}_1, \dots, \overline{X}_1}_{n_1 volte}, \dots\dots, \underbrace{\overline{X}_k, \dots, \overline{X}_k}_{n_k volte}); \qquad X_H = (\underbrace{\overline{X}, \dots, \overline{X}}_{n\, volte}).$$

Quindi

$$X - X_E = (X_{1,1} - \overline{X}_1, \dots, X_{1,n_1} - \overline{X}_1, \dots\dots, X_{k,1} - \overline{X}_k, X_{k,n_k} - \overline{X}_k);$$
$$X_E - X_H = (\underbrace{\overline{X}_1 - \overline{X}, \dots, \overline{X}_1 - \overline{X}}_{n_1 volte}, \dots\dots, \underbrace{\overline{X}_k - \overline{X}, \dots, \overline{X}_k - \overline{X}}_{n_k volte}).$$

Pertanto

$$\|X - X_E\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_i)^2; \qquad \|X_E - X_H\|^2 = \sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2.$$

Osservazione 18.5 La quantità $\|X - X_E\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_i)^2$ si chiama anche *variazione interna*; si tratta cioè della somma delle variazioni di ogni gruppo attorno alla propria media campionaria; $\|X_E - X_H\|^2 = \sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2$ è invece la cosiddetta *variazione esterna* (cioè la somma delle variazioni delle medie campionarie dei vari gruppi rispetto alla media campionaria del campione globale). È naturale considerare anche la *variazione totale*, che è definita come

$$\|X - X_H\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X})^2.$$

Ora, i vettori $X - X_E$ e $X_E - X_H$ sono tra loro perpendicolari (i vettori X_E e X_H appartengono a E, mentre $X - X_E$ è perpendicolare a E per definizione di proiezione su E); dunque per i Teorema di Pitagora abbiamo

$$\|X - X_H\|^2 = \|(X - X_E) + (X_E - X_H)\|^2 = \|X - X_E\|^2 + \|X_E - X_H\|^2.$$

In altre parole

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_i)^2 + \sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2.$$

Questa formula, nota in fisica come *formula di Huygens*, dice che la variazione totale è la somma delle variazioni interna e esterna.

Tornando alla descrizione del test, consideriamo la v.a. (ved. il Lemma precedente)

$$Z = \frac{\frac{\|X_E - X_H\|^2}{k-1}}{\frac{\|X - X_E\|^2}{n-k}} = \frac{\frac{\sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2}{k-1}}{\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_i)^2}{n-k}},$$

che, sotto l'ipotesi, ha legge $F(k-1, n-k)$ per il Teorema di Cochran 9.3. Si sceglierà un test di regione critica $\{Z > C\}$, e si determina C con i soliti calcoli.

19 Il modello bayesiano

Cominciamo ricordando la nozione di "nucleo di transizione" tra due spazi misurabili.

Siano $(E, \mathcal{E})$ e $(F, \mathcal{F})$ due spazi misurabili.

Definizione 19.1 Si chiama *probabilità* (o *nucleo*)*di transizione* di $(E, \mathcal{E})$ su $(F, \mathcal{F})$ una funzione $N : E \times \mathcal{F} \to [0, 1]$ tale che

(i) per ogni fissato $A \in \mathcal{F}$, la funzione $x \mapsto N(x, A)$ (da $(E, \mathcal{E})$ in $[0, 1]$) è $\mathcal{E}$- misurabile;

(ii) per ogni fissato $x \in E$, la funzione $A \mapsto N(x, A)$ (da $\mathcal{F}$ in $[0, 1]$) è una probabilità su $(F, \mathcal{F})$.

Richiamiamo inoltre (senza dimostrazione, che è una semplice estensione del Teorema di Fubini classico) il

Teorema 19.2 (di Fubini generalizzato). *Sia N una probabilità di transizione di $(E, \mathcal{E})$ su $(F, \mathcal{F})$, e sia P una misura di probabilità su $(E, \mathcal{E})$. Allora*

(a) per ogni funzione $f : E \times F \to \mathbb{R}$ che sia $\mathcal{E} \otimes \mathcal{F}$- misurabile e limitata, la funzione

$$x \mapsto \int_F f(x, y) N(x, \mathrm{d}y)$$

è $\mathcal{E}$-misurabile;

(b) esiste una e una sola probabilità Q su $\mathcal{E} \otimes \mathcal{F}$ tale che, se f è limitata, valga la formula

$$\iint_{E \times F} f(x, y) Q(\mathrm{d}x, \mathrm{d}y) = \int_E P(\mathrm{d}x) \int_F f(x, y) N(x, \mathrm{d}y);$$

(c) la proprietà (a) e la formula in (b) rimangono vere per f positiva, non necessariamente limitata;

(d) sia f misurabile di segno qualunque; allora f è Q-integrabile se e solo se

$$\iint_{E \times F} |f(x, y)| Q(\mathrm{d}x, \mathrm{d}y) = \int_E P(\mathrm{d}x) \int_F |f(x, y)| N(x, \mathrm{d}y) < +\infty.$$

In tal caso restano valide per f la proprietà (a) e la formula in (b).

Possiamo ora dare la

Definizione 19.3 Un *modello statistico bayesiano* è formato da

(a) un modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in (\Theta, \mathcal{T})\}\big)$, nel quale si suppone che l'insieme dei parametri Θ sia munito di una σ-algebra $\mathcal{T}$ e che, fissato $A \in \mathcal{F}$, l'applicazione $\theta \mapsto P^\theta(A)$ sia misurabile;

(b) una misura di probabilità ν su $(\Theta, \mathcal{T})$, chiamata *legge a priori* del parametro.

Osservazione 19.4 La probabilità ν va interpretata come la conoscenza che si ha della situazione prima di fare l'indagine statistica.

Esempio 19.5 Un segnale proviene il 40% delle volte da un'apparecchiatura A_1 e per il restante 60% delle volte da una seconda apparecchiatura A_2. Esso può essere di due tipi: "lungo" (L) oppure "breve"(B). È noto che A_1 (risp. A_2) trasmette un segnale breve il 48% (risp. il 63%) delle volte. In un certo istante viene ricevuto un segnale breve; qual è la probabilità che esso provenga da A_1? e da A_2?

Soluzione. Consideriamo i tre eventi

$$\begin{aligned} A_1 &= \{\text{il segnale proviene da } A_1\};\\ A_2 &= \{\text{il segnale proviene da } A_2\};\\ B &= \{\text{il segnale risulta breve}\}. \end{aligned}$$

Per essi si ha

$$P(A_1) = 0,40; \quad P(A_2) = 0,60; \quad P(B|A_1) = 0,48; \quad P(B|A_2) = 0,63.$$

Ci interessa calcolare $P(A_1|B)$ e $P(A_2|B)$; per la formula di Bayes si ha

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} = \frac{0,48 \times 0,40}{0,48 \times 0,40 + 0,63 \times 0,60} = 0,3368,$$

e di conseguenza $P(A_2|B) = 1 - P(A_1|B) = 0,6632$.

Per riformulare la situazione nell'ambito dei modelli bayesiani, osserviamo che la nostra conoscenza "a priori" (cioè prima dell'invio del segnale) delle probabilità di A_1 e di A_2 era $P(A_1) = 0,40$ e $P(A_2) = 0,60$ (conoscenza ricavata in base ad esperienze precedenti, "a priori" appunto); dopo l'invio del segnale abbiamo rivalutato queste probabilità in $P(A_1|B) = 0,3368$ e $P(A_2|B) = 0,6632$ (probabilità "a posteriori ").
Con le notazioni della Definizione 19.3, diremo allora che

(i) $\Omega = \{B, L\}$;

(ii) il parametro θ appartiene a $\Theta = \{1, 2\}$;

(iii) $P^1(\cdot) = P(\cdot|A_1)$, , $P^2(\cdot) = P(\cdot|A_2)$;

(iv) $\nu(\{1\}) = 0,40$, $\nu(\{2\}) = 0,60$.

La teoria bayesiana si occupa, fra l'altro, dei metodi per calcolare la legge "a posteriori"; nel nostro esempio si tratta della legge che assegna a $\theta = 1$ il peso $0,3368$ e a $\theta = 2$ il peso $0,6632$, dati ottenuti dalla formula di Bayes (di qui il nome della teoria) in base al risultato dell'esperimento. Vedremo fra poco (Esempio 19.10) come si formalizza il calcolo fatto sopra.

Torniamo alla definizione generale di modello bayesiano. La funzione $(\theta, A) \mapsto P^\theta(A)$ è una probabilità di transizione da $(\Theta, \mathcal{T})$ in $(\Omega, \mathcal{F})$, e perciò in questo contesto useremo la notazione $P(\theta, A)$ piuttosto che quella solita $P^\theta(A)$. Consideriamo poi la probabilità Q su $(\Theta \times \Omega, \mathcal{T} \otimes \mathcal{F})$ associata, secondo il Teorema di Fubini generalizzato, alla legge a priori ν e alla probabilità di transizione $P(\theta, A)$; in particolare vale la relazione

$$\begin{aligned} Q(T \times A) &= \iint_{\Theta\times\Omega} 1_{T\times A}(\theta,\omega) Q(\mathrm{d}\theta, \mathrm{d}\omega) = \int_\Theta \nu(\mathrm{d}\theta) \int_\Omega 1_T(\theta) 1_A(\omega) P(\theta, \mathrm{d}\omega)\\ &= \int_T \nu(\mathrm{d}\theta) \int_A P(\theta, \mathrm{d}\omega) = \int_T \nu(\mathrm{d}\theta) P(\theta, A), \qquad \forall\, T \in \mathcal{T}, \forall\, A \in \mathcal{F}. \end{aligned} \tag{30}$$

Indichiamo con $\tilde{\mathcal{T}}$ la σ-algebra su $\Theta\times\Omega$ formata dagli insiemi del tipo $T\times\Omega$, con $T\in\mathcal{T}$. Osserviamo che una variabile $X:\Theta\times\Omega\to\mathbb{R}$ è $\tilde{\mathcal{T}}$-misurabile se e solo se $X(\theta,\omega)=V(\theta)$ (cioè X dipende solo da θ) e $V:(\Theta,\mathcal{T})\to\mathbb{R}$ è misurabile (dimostrazione per esercizio).

Analogamente si definisce la σ-algebra $\tilde{\mathcal{F}}$ su $\Theta\times\Omega$ come la σ-algebra formata dagli insiemi del tipo $\Theta\times A$, con $A\in\mathcal{F}$, e una variabile $X:\Theta\times\Omega\to\mathbb{R}$ è $\tilde{\mathcal{F}}$-misurabile se e solo se $X(\theta,\omega)=W(\omega)$ con $W:(\Omega,\mathcal{F})\to\mathbb{R}$ misurabile.

Teorema 19.6 *Sia X una v.a. limitata definita su $(\Theta\times\Omega,\mathcal{T}\otimes\mathcal{F},Q)$. Vale allora l'uguaglianza*

$$E[X|\tilde{\mathcal{T}}](\theta)=\int_\Omega X(\theta,\omega)P(\theta,\mathrm{d}\omega),\qquad Q\text{-q.c.}$$

(Ovviamente, in questa formula, la speranza condizionale è calcolata rispetto all probabilità Q su $\Theta\times\Omega$).

Dimostrazione. La variabile $(\theta,\omega)\mapsto Y(\theta,\omega)=\int_\Omega X(\theta,\omega)P(\theta,\mathrm{d}\omega)$ è $\tilde{\mathcal{T}}$-misurabile (perché dipende solo da θ e, per il Teorema di Fubini generalizzato, la funzione $\theta\mapsto\int_\Omega X(\theta,\omega)P(\theta,\mathrm{d}\omega)$ è $\tilde{\mathcal{T}}$-misurabile). Dunque, per definizione di probabilità condizionale, basta provare che, su ogni insieme della forma $T\times\Omega$ si ha

$$\iint_{T\times\Omega}X\,\mathrm{d}Q=\iint_{T\times\Omega}Y\,\mathrm{d}Q.$$

Infatti, per il Teorema di Fubini,

$$\iint_{T\times\Omega}X\,\mathrm{d}Q=\int_T\nu(\mathrm{d}\theta)\int_\Omega X(\theta,\omega)P(\theta,\mathrm{d}\omega);$$

$$\begin{aligned}\iint_{T\times\Omega}Y\,\mathrm{d}Q&=\int_T\nu(\mathrm{d}\theta)\int_\Omega Y(\theta,\omega)P(\theta,\mathrm{d}\omega)=\int_T\nu(\mathrm{d}\theta)\int_\Omega\Big(\int_\Omega X(\theta,\omega)P(\theta,\mathrm{d}\omega)\Big)P(\theta,\mathrm{d}\omega)\\&=\int_T\nu(\mathrm{d}\theta)\Big(\int_\Omega X(\theta,\omega)P(\theta,\mathrm{d}\omega)\Big)\underbrace{\Big(\int_\Omega P(\theta,\mathrm{d}\omega)\Big)}_{=1}=\int_T\nu(\mathrm{d}\theta)\int_\Omega X(\theta,\omega)P(\theta,\mathrm{d}\omega),\end{aligned}$$

e si ha l'uguaglianza cercata.

□

La relazione (30) permette di interpretare la probabilità P^θ (definita su $(\Omega,\mathcal{F})$) come la legge condizionale (su $\Theta\times\Omega$), noto θ: infatti, se per un certo valore θ_0 si ha $\{\theta_0\}\in\mathcal{T}$ e $\nu(\{\theta_0\})>0$, allora, preso $A\in\mathcal{F}$, $P^{\theta_0}(A)=P(\theta_0,A)$ è effettivamente una probabilità condizionale:

$$P^{\theta_0}(A)=P(\theta_0,A)=Q\big(\Theta\times A|\{\theta_0\}\times\Omega\big),$$

o, con abuso di scrittura,

$$P^{\theta_0}(A)=P(\theta_0,A)=Q(A|\theta_0).$$

Infatti

$$Q\big(\Theta \times A | \{\theta_0\} \times \Omega\big) = \frac{Q\big((\Theta \times A) \cap (\{\theta_0\} \times \Omega)\big)}{Q\big((\{\theta_0\} \times \Omega)\big)} = \frac{Q(\{\theta_0\} \times A)}{Q(\{\theta_0\} \times \Omega)}$$
$$= \frac{\int_{\{\theta_0\}} \nu(\mathrm{d}\theta) P(\theta, A)}{\int_{\{\theta_0\}} \nu(\mathrm{d}\theta) P(\theta, \Omega)} = \frac{P(\theta_0, A)\nu(\{\theta_0\})}{P(\theta_0, \Omega)\nu(\{\theta_0\})} = P(\theta_0, A).$$

Quindi si può dire che la probabilità di transizione $P(\theta, A)$ di $(\Theta, \mathcal{T})$ su $(\Omega, \mathcal{F})$ dà la probabilità condizionale di A, "noto il valore di θ".

Se vogliamo invece "aggiornare" la nostra conoscenza della probabilità a priori ν (cioè del fenomeno che stiamo osservando), "noto il valore di ω"(cioè dopo aver effettuato l'esperimento), abbiamo bisogno di procedere "al contrario", cioè dobbiamo trovare una probabilità di transizione $N(\omega, T)$ di $(\Omega, \mathcal{F})$ su $(\Theta, \mathcal{T})$. Diamo allora la seguente

Definizione 19.7 Supponiamo che esista una probabilità di transizione $N(\omega, T)$ di $(\Omega, \mathcal{F})$ su $(\Theta, \mathcal{T})$ tale che, per ogni v.a. X limitata definita su $(\Theta \times \Omega, \mathcal{T} \otimes \mathcal{F})$, risulti

$$E[X|\tilde{\mathcal{F}}](\omega) = \int_\Theta X(\theta, \omega) N(\omega, \mathrm{d}\theta), \qquad Q\text{-}q.c.$$

Allora la probabilità $N(\omega, \cdot)$ si chiama *legge a posteriori su* $(\Theta, \mathcal{T})$ *condizionale a* ω.

Osservazione 19.8 Dunque il meccanismo bayesiano è il seguente:

(a) prima dell'esperimento la legge a priori ν rappresenta la nostra conoscenza del fenomeno;

(b) si effettua l'esperimento, che dà esito ω;

(c) la conoscenza del fenomeno è aggiornata passando alla probabilità $N(\omega, \cdot)$.

L'esistenza della legge a posteriori in ipotesi generali è un teorema piuttosto complicato; noi ci limiteremo al caso di un modello dominato, che copre la maggior parte delle applicazioni.

Supponiamo dunque che il modello sia dominato da una misura μ su $(\Omega, \mathcal{F})$ e che esista una versione della verosimiglianza $L(\theta, \omega) = \frac{\mathrm{d}P^\theta}{\mathrm{d}\mu}(\omega)$ che sia $\mathcal{T} \otimes \mathcal{F}$- misurabile (questo garantisce in particolare che $\theta \mapsto P(\theta, A)$ sia $\mathcal{T}$-misurabile).

Teorema 19.9 *Nelle ipotesi sopra enunciate, poniamo*

$$G = \Big\{\omega \in \Omega : \int_\Theta L(\theta, \omega)\nu(\mathrm{d}\theta) = 0\Big\};$$

$$M = \Big\{\omega \in \Omega : \int_\Theta L(\theta, \omega)\nu(\mathrm{d}\theta) = +\infty\Big\};$$

$$g(\theta, \omega) = \begin{cases} \dfrac{L(\theta, \omega)}{\int_\Theta L(\tau, \omega)\nu(\mathrm{d}\tau)} & \text{se } \omega \in G^c \\ \\ 1 & \text{se } \omega \in G \end{cases}$$

(con la convenzione $\frac{c}{+\infty} = 0$*). Allora*

(a) $\Theta \times G$ *è* Q*-trascurabile;*

(b) M *è* μ*-trascurabile;*

(c) fissato $\omega \in M^c$*, la funzione* $\theta \mapsto g(\theta, \omega)$ *è una densità di probabilità su* $(\Theta, \mathcal{T})$ *rispetto a* ν*;*

(d) fissato $\omega \in M^c$*, la probabilità di transizione definita da*

$$T \mapsto N(\omega, T) = \int_T g(\theta, \omega)\nu(\mathrm{d}\theta)$$

è la legge a posteriori su $(\Theta, \mathcal{T})$ *condizionale a* ω*.*

Dimostrazione. (a) Si ha, per la definizione di verosimiglianza e per il Teorema di Fubini,

$$Q(\Theta \times G) = \int_\Theta \nu(\mathrm{d}\theta) \int_G P(\theta, \mathrm{d}\omega) = \int_\Theta \nu(\mathrm{d}\theta) \int_G L(\theta, \omega)\mu(\mathrm{d}\omega) = \int_G \mu(\mathrm{d}\omega) \int_\Theta L(\theta, \omega)\nu(\mathrm{d}\theta) = 0,$$

per la definizione di G.
(b) È ovvio perché

$$\int_\Omega \mu(\mathrm{d}\omega) \int_\Theta L(\theta, \omega)\nu(\mathrm{d}\theta) = \int_\Theta \nu(\mathrm{d}\theta) \int_\Omega L(\theta, \omega)\mu(\mathrm{d}\omega) = \int_\Theta \nu(\mathrm{d}\theta) \underbrace{\int_\Omega P(\theta, \mathrm{d}\omega)}_{=1}$$

$$= \int_\Theta \nu(\mathrm{d}\theta) = 1 < +\infty.$$

(c) Dobbiamo provare che, fissato $\omega \in M^c$, si ha $g(\theta, \omega) \geq 0$ (questo è ovvio) ed inoltre risulta

$$\int_\Theta g(\theta, \omega)\nu(\mathrm{d}\theta) = 1.$$

Ora, se $\omega \in G^c$, si ha

$$\int_\Theta g(\theta, \omega)\nu(\mathrm{d}\theta) = \int_\Theta \frac{L(\theta, \omega)}{\int_\Theta L(\tau, \omega)\nu(\mathrm{d}\tau)}\nu(\mathrm{d}\theta) = 1;$$

se $\omega \in G$

$$\int_\Theta g(\theta, \omega)\nu(\mathrm{d}\theta) = \int_\Theta 1\nu(\mathrm{d}\theta) = 1.$$

(d) Osserviamo prima di tutto che $\omega \mapsto N(\omega, T)$ è misurabile (per il Teorema di Fubini, dato che è l'integrale di una funzione misurabile). Inoltre $T \mapsto N(\omega, T)$ è una probabilità ($N(\omega, \Theta) = \int_\Theta g(\theta, \omega)\nu(\mathrm{d}\theta) = 1$, come abbiamo appena visto nel punto (c)). Dunque $N(\omega, T)$ è effettivamente una probabilità di transizione (di $(\Omega, \mathcal{F})$ su $(\Theta, \mathcal{T})$).
Per vedere che si tratta effettivamente della legge a posteriori, in base alla Definizione 19.7 dobbiamo far vedere che, se X è una v.a. limitata definita su $(\Theta \times \Omega, \mathcal{T} \otimes \mathcal{F})$ e posto

$$V(\omega) = \int_\Theta X(\theta, \omega)g(\theta, \omega)\nu(\mathrm{d}\theta) = \begin{cases} \int_\Theta X(\theta, \omega)\dfrac{L(\theta, \omega)}{\int_\Theta L(\tau, \omega)\nu(\mathrm{d}\tau)}\nu(\mathrm{d}\theta) & \text{se } \omega \in G^c \\ \\ \int_\Theta X(\theta, \omega)\nu(\mathrm{d}\theta) & \text{se } \omega \in G, \end{cases}$$

su ogni insieme della forma $\Theta \times B$ si ha

$$\iint_{\Theta \times B} X \,\mathrm{d}Q = \iint_{\Theta \times B} V \,\mathrm{d}Q.$$

Infatti, per la definizione di Q e per il punto (a), si ha

$$\iint_{\Theta\times B} X\,\mathrm{d}Q = \iint_{\Theta\times(B\cap G^c)} X\,\mathrm{d}Q = \int_\Theta \nu(\mathrm{d}\theta)\int_{B\cap G^c} X(\theta,\omega)P(\theta,\mathrm{d}\omega)$$
$$= \int_\Theta \nu(\mathrm{d}\theta)\int_{B\cap G^c} X(\theta,\omega)L(\theta,\omega)\mu(\mathrm{d}\omega);$$

d'altra parte

$$\iint_{\Theta\times B} V\,\mathrm{d}Q = \int_\Theta \nu(\mathrm{d}\theta)\int_{B\cap G^c} V(\theta,\omega)L(\theta,\omega)\mu(\mathrm{d}\omega)$$
$$= \int_\Theta \nu(\mathrm{d}\theta)\int_{B\cap G^c}\Big(\int_\Theta X(s,\omega)\frac{L(s,\omega)}{\int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)}\nu(\mathrm{d}s)\Big)L(\theta,\omega)\mu(\mathrm{d}\omega)$$
$$= \int_\Theta \nu(\mathrm{d}s)\int_{B\cap G^c} X(s,\omega)L(s,\omega)\mu(\mathrm{d}\omega)\frac{\int_\Theta L(\theta,\omega)\nu(\mathrm{d}\theta)}{\int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)} = \int_\Theta \nu(\mathrm{d}s)\int_{B\cap G^c} X(s,\omega)L(s,\omega)\mu(\mathrm{d}\omega).$$

□

D'ora in avanti indicheremo con ν^ω la legge a posteriori su $(\Theta,\mathcal{T})$ condizionale a ω; come abbiamo appena visto, nelle ipotesi in cui ci siamo posti (modello dominato) essa è assolutamente continua rispetto a ν, di densità

$$\frac{\mathrm{d}\nu^\omega}{\mathrm{d}\nu} = g(\theta,\omega) = \begin{cases} \dfrac{L(\theta,\omega)}{\int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)} & \text{se } \int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)\neq 0 \\ \\ 1 & \text{se } \int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)=0. \end{cases}$$

Esempio 19.10 Riprendiamo l'Esempio 19.5. Abbiamo già detto che $\Theta=\{1,2\}$, $\nu(\{1\})=0,40$ e $\nu(\{2\})=0,60$; inoltre avevamo $\Omega=\{B,L\}$; per motivi di notazione scriviamo ω' al posto di B e ω'' al posto di L, quindi $\Omega=\{\omega',\omega''\}$. Le quantità $P(\omega'|A_1)$, $P(\omega'|A_2)$, $P(\omega''|A_1)$,$P(\omega''|A_2)$ sono da interpretare come i valori della verosimiglianza L: precisamente, se prendiamo come misura dominante su Ω la misura μ che conta i punti, si ha

$$L(1,\omega') = \frac{\mathrm{d}P^1}{\mathrm{d}\mu}(\omega') = P(\omega'|A_1) = 0,48; \qquad L(2,\omega') = \frac{\mathrm{d}P^2}{\mathrm{d}\mu}(\omega') = P(\omega'|A_2) = 0,63;$$

$$L(1,\omega'') = \frac{\mathrm{d}P^1}{\mathrm{d}\mu}(\omega'') = P(\omega''|A_1) = 0,52; \qquad L(2,\omega'') = \frac{\mathrm{d}P^2}{\mathrm{d}\mu}(\omega'') = P(\omega''|A_2) = 0,37.$$

La legge a posteriori è assolutamente continua rispetto a ν con densità

$$g(\theta,\omega) = \frac{L(\theta,\omega)}{L(1,\omega)\nu(\{1\}) + L(2,\omega)\nu(\{2\})}, \qquad (\theta,\omega)\in\{1,2\}\times\{\omega',\omega''\}.$$

In altre parole

$$N(\omega',\{1\}) = \frac{L(1,\omega')}{L(1,\omega')\nu(\{1\})+L(2,\omega')\nu(\{2\})}\cdot\nu(\{1\}) = \frac{0,48}{0,48\times 0,40 + 0,63\times 0,60}\times 0,40 = 0,3368,$$

e questo è il calcolo che abbiamo fatto in precedenza per ottenere $P(A_1|B)$. In modo simile per gli altri valori.

20 Il formalismo decisionale; decisione bayesiana

Anche se non rappresenta tutti gli aspetti della statistica, il formalismo decisionale è abbastanza intuitivo ed ha permesso degli sviluppi matematici rigorosi. Diciamo subito che le definizioni usuali della teoria degli stimatori e dei test possono essere ricondotte al formalismo decisionale.

Lo statistico osserva un fenomeno la cui legge dipende da un parametro $\theta \in \Theta$, con lo scopo di intraprendere un'azione $a \in A$. A, insieme delle azioni possibili, è un insieme di oggetti, che di solito sono numeri ma teoricamente possono essere anche altro, se necessario. La scelta dell'azione a porta come conseguenza un *costo* (o *perdita*) $C(\theta, a) \geq 0$, dipendente dal parametro θ.
Lo statistico effettua un esperimento, formalizzato con un modello statistico $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in \Theta\}\big)$, e la sua decisione sull'azione da intraprendere dipenderà naturalmente dal risultato ω, cioè

Definizione 20.1 Una *regola decisionale* è una funzione $\delta : \Omega \to A$. Si chiama *funzione costo* (*della regola* δ) la funzione $\omega \mapsto C(\theta, \delta(\omega))$.

Supporremo sempre che A sia munito di una σ-algebra $\mathcal{A}$, e che le funzioni $\delta : \omega \mapsto \delta(\omega)$ e $a \mapsto C(\theta, a)$ siano misurabili. In queste ipotesi, la funzione costo è una variabile aleatoria non negativa.

Esempio 20.2 Uno stimatore, così come lo abbiamo definito a suo tempo, è una regola decisionale. In questo caso l'insieme delle azioni è l'aperto D che compare nella Definizione 3.1. Abbiamo notato a suo tempo che un test non è altro che uno stimatore (Osservazione 12.5). Dunque anche la nozione di test può essere ricondotta al formalismo decisionale.

Nella statistica classica si definisce *rischio* della regola decisionale δ la speranza della sua funzione costo, e precisamente

$$R_\delta(\theta) = E^\theta[C(\theta, \delta)]$$

(in realtà noi abbiamo dato questa definizione solo per gli stimatori, e di conseguenza per i test, ma, come si vede subito, la definizione è identica per ogni regola decisionale).
Lo scopo è poi quello di minimizzare il rischio, come accadeva per gli stimatori, ma, dovendo tener conto del fatto che su Θ ora abbiamo una probabilità ν, avremo bisogno di minimizzare rispetto a δ non $R_\delta(\theta)$ a θ fissato, ma il suo integrale rispetto a ν; in altre parole dovremo trovare una regola decisonale δ_0 tale che

$$\int_\Theta R_{\delta_0}(\theta)\nu(\mathrm{d}\theta) \leq \int_\Theta R_\delta(\theta)\nu(\mathrm{d}\theta), \qquad \forall\, \delta.$$

Nel contesto bayesiano si procede dunque come segue. Sia ρ una generica misura di probabilità su $(\Theta, \mathcal{T})$. Cerchiamo prima di tutto di minimizzare la *perdita media*, definita da

$$\int_\Theta C(\theta, a)\rho(\mathrm{d}\theta)$$

rispetto al parametro a. Si dà cioè la seguente

Definizione 20.3 Si chiama *rischio bayesiano* (relativo a ρ) il numero

$$\inf_{a \in A} \int_\Theta C(\theta, a)\rho(\mathrm{d}\theta).$$

Inoltre

Definizione 20.4 Se esiste $a_0 \in A$ tale che

$$\int_\Theta C(\theta, a_0)\rho(\mathrm{d}\theta) = \inf_{a\in A} \int_\Theta C(\theta, a)\rho(\mathrm{d}\theta),$$

a_0 si chiama *decisione bayesiana* relativa a ρ. In altre parole, in questo modo si definisce una funzione $d : \rho \mapsto a_0 = d(\rho)$.

Il meccanismo decisionale bayesiano consiste allora nel considerare come regola decisionale δ_0 la regola $\omega \mapsto d(\nu^\omega)$ (ammesso che esista), dove ν^ω è la legge a posteriori condizionale a ω di cui abbiamo parlato nel paragrafo precedente.

La bontà di questa procedura è espressa dal risultato seguente:

Teorema 20.5 *Sia $\big(\Omega, \mathcal{F}, \{P^\theta, \theta \in (\Theta, \mathcal{T})\}\big)$ un modello statistico dominato, con verosimiglianza $L(\theta, \omega)$, e sia ν la legge a priori su $(\Theta, \mathcal{T})$. Supponiamo che, per ogni $\omega \in \Omega$, esista una decisione bayesiana $d(\nu^\omega)$ e che la funzione $\omega \mapsto d(\nu^\omega)$ sia misurabile. Allora, per ogni regola di decisione δ, si ha*

$$\int_\Theta R_{d(\nu^\diamond)}(\theta)\nu(\mathrm{d}\theta) \le \int_\Theta R_\delta(\theta)\nu(\mathrm{d}\theta).$$

Dimostrazione. Sia G l'insieme definito nel Teorema 19.9; ricordiamo che $\Theta \times G$ è Q- trascurabile. Per la definizione di Q si ha allora

$$\begin{aligned}
&\int_\Theta R_\delta(\theta)\nu(\mathrm{d}\theta) = \int_\Theta \nu(\mathrm{d}\theta)\Big(\int_\Omega C(\theta, \delta(\omega))P(\theta, \mathrm{d}\omega)\Big) = \iint_{\Theta\times\Omega} C(\theta, \delta(\omega))Q(\mathrm{d}\theta, \mathrm{d}\omega)\\
&= \iint_{\Theta\times(\Omega\cap G^c)} C(\theta, \delta(\omega))Q(\mathrm{d}\theta, \mathrm{d}\omega) = \int_\Theta \nu(\mathrm{d}\theta)\Big(\int_{\Omega\cap G^c} C(\theta, \delta(\omega))P(\theta, \mathrm{d}\omega)\Big)\\
&= \int_\Theta \nu(\mathrm{d}\theta)\Big(\int_{\Omega\cap G^c} C(\theta, \delta(\omega))L(\theta, \omega)\mu(\mathrm{d}\omega)\Big)\\
&= \int_{\Omega\cap G^c} \mu(\mathrm{d}\omega)\Big(\int_\Theta \nu(\mathrm{d}\theta)C(\theta, \delta(\omega))L(\theta, \omega)\Big)\\
&= \int_{\Omega\cap G^c} \mu(\mathrm{d}\omega)\Big(\int_\Theta L(\tau, \omega)\nu(\mathrm{d}\tau)\Big)\Big(\int_\Theta \nu(\mathrm{d}\theta)C(\theta, \delta(\omega))\frac{L(\theta, \omega)}{\int_\Theta L(\tau, \omega)\nu(\mathrm{d}\tau)}\Big)\\
&= \int_{\Omega\cap G^c} \mu(\mathrm{d}\omega)\Big(\int_\Theta L(\tau, \omega)\nu(\mathrm{d}\tau)\Big)\Big(\int_\Theta C(\theta, \delta(\omega))\nu^\omega(\mathrm{d}\theta)\Big)\\
&\ge \int_{\Omega\cap G^c} \mu(\mathrm{d}\omega)\Big(\int_\Theta L(\tau, \omega)\nu(\mathrm{d}\tau)\Big)\Big(\int_\Theta C(\theta, d(\nu^\omega))\nu^\omega(\mathrm{d}\theta)\Big),
\end{aligned}$$

e, con gli stessi calcoli, si vede che quest'ultima quantità è uguale a

$$\int_\Theta R_{d(\nu^\diamond)}(\theta)\nu(\mathrm{d}\theta).$$

Osservazione 20.6 Per le regole di decisione si possono dare le definizioni di preferibile, strettamente preferibile e ammissibile esattamente come per gli stimatori. La decisione bayesiana è ammissibile (in senso classico), sotto diverse ipotesi facili da verificare, per esempio:

(i) se Θ è un aperto di $\mathbb{R}^k$, il supporto di ν è tutto Θ (cioè non esistono sottoinsiemi aperti non vuoti di Θ che siano trascurabili rispetto a ν) e, per ogni regola di decisione δ, la funzione $\theta \mapsto R(\theta, \delta)$ è continua.

Infatti, se $d(\nu^\diamond)$ non fosse ammissibile, esisterebbe una regola di decisione δ strettamente preferibile a $d(\nu^\diamond)$, cioè tale che

$$\begin{aligned} &R_\delta(\theta) \le R_{d(\nu^\diamond)}(\theta), \qquad \forall\, \theta; \\ &R_\delta(\theta_0) < R_{d(\nu^\diamond)}(\theta_0), \qquad \text{per almeno un } \theta_0. \end{aligned}$$

Ma in tal caso, per motivi di continuità, la diseguaglianza stretta sarebbe verificata per ogni θ in un intorno U di θ_0, con $\nu(U) > 0$, e quindi la tesi del Teorema precedente, e cioè

$$\int_\Theta R_\delta(\theta)\nu(\mathrm{d}\theta) \ge \int_\Theta R_{d(\nu^\diamond)}(\theta)\nu(\mathrm{d}\theta)$$

sarebbe falsa.

(ii) Se Θ è numerabile e nessun punto di Θ è ν-trascurabile (verifica per esercizio, è simile a quella del punto precedente).

Osservazione 20.7 Se T è una statistica esaustiva, come sappiamo dal Teorema di fattorizzazione 2.6 si ha

$$L(\theta, \omega) = h(\omega) g^\theta(T(\omega))$$

e di conseguenza

$$\frac{\mathrm{d}\nu^\omega}{\mathrm{d}\nu} = g(\theta, \omega) = \begin{cases} \dfrac{g^\theta(T(\omega))}{\int_\Theta g^\tau(T(\omega))\nu(\mathrm{d}\tau)} & \text{se } \int_\Theta g^\tau(T(\omega))\nu(\mathrm{d}\tau) \neq 0 \\ \\ 1 & \text{se } \int_\Theta g^\tau(T(\omega))\nu(\mathrm{d}\tau) = 0 \end{cases}$$

è $\sigma(T)$-misurabile. Dunque si può sostituire il risultato dell'esperimento (cioè ω) con l'osservazione della statistica $T(\omega)$. Inoltre, in genere $d(\nu^\diamond)$ è $\sigma(T)$-misurabile.

Esercizio 20.8 Sia $\theta = (0, 1)$ e $C(\theta, a) = \theta(1 - a) + (1 - \theta)a$. Qual è la decisione bayesiana per una generica legge ρ? Verificare che il rischio bayesiano è sempre inferiore a $\frac{1}{2}$.

21 Stimatori bayesiani

Sia $g : \Theta \to \mathbb{R}$ misurabile e limitata. Vogliamo stimare la quantità $g(\theta)$; per questo porremo $A = \mathbb{R}$ e utilizzeremo il costo $C(\theta, a) = \big(g(\theta) - a\big)^2$. La procedura bayesiana illustrata precedentemente richiede per prima cosa che, per ogni fissata probabilità ρ su Θ, si cerchi (se esiste) $a_0 = d(\rho)$ che minimizza la quantità

$$\int_\Theta \big(g(\theta) - a\big)^2 \rho(\mathrm{d}\theta).$$

Lemma 21.1 *Se $\int_\Theta g^2(\theta)\rho(\mathrm{d}\theta) < +\infty$, allora*

$$\int_\Theta \big(g(\theta) - a\big)^2 \rho(\mathrm{d}\theta) \ge \int_\Theta \big(g(\theta) - a_0\big)^2 \rho(\mathrm{d}\theta),$$

dove

$$a_0 = \int_\Theta g(\theta)\rho(\mathrm{d}\theta) =: d(\rho).$$

DIMOSTRAZIONE. Infatti

$$\begin{aligned}
&\int_\Theta \big(g(\theta)-a\big)^2\rho(\mathrm{d}\theta) = \int_\Theta \big((g(\theta)-a_0)+(a_0-a)\big)^2\rho(\mathrm{d}\theta)\\
&= \int_\Theta \big(g(\theta)-a_0\big)^2\rho(\mathrm{d}\theta) + (a_0-a)^2 + 2(a_0-a)\underbrace{\int_\Theta \big(g(\theta)-a_0\big)\rho(\mathrm{d}\theta)}_{=0}\\
&= \int_\Theta \big(g(\theta)-a_0\big)^2\rho(\mathrm{d}\theta) + (a_0-a)^2 \geq \int_\Theta \big(g(\theta)-a_0\big)^2\rho(\mathrm{d}\theta).
\end{aligned}$$

□

Il Lemma precedente dice dunque che lo stimatore bayesiano della quantità $g(\theta)$ è

$$T(\omega) := d(\nu^\omega) = \int_\Theta g(\theta)\nu^\omega(\mathrm{d}\theta) = \begin{cases} \displaystyle\int_\Theta g(\theta)\frac{L(\theta,\omega)}{\int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)}\nu(\mathrm{d}\theta) & \text{se } \displaystyle\int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)\neq 0\\[2ex] \displaystyle\int_\Theta g(\theta)\nu(\mathrm{d}\theta) & \text{se } \displaystyle\int_\Theta L(\tau,\omega)\nu(\mathrm{d}\tau)= 0. \end{cases}$$

Notiamo che, per il Teorema 19.9, si ha l'uguaglianza

$$T(\omega) = E[g|\tilde{\mathcal{F}}], \qquad Q-\text{q.c.}$$

Infatti

$$E[g|\tilde{\mathcal{F}}] \underbrace{=}_{Def.19.7} \int_\Theta g(\theta)N(\omega,\mathrm{d}\theta) \underbrace{=}_{Teor.19.9} \int_\Theta g(\theta)\nu^\omega(\mathrm{d}\theta) = T(\omega).$$

Questo fatto permette quindi di interpretare lo stimatore di $g(\theta)$ in termini di speranza condizionale; questa interpretazione è utile nel risultato che segue, che dice che, tranne che in casi banali, gli stimatori bayesiani non sono mai corretti (ved. Osservazione 21.3).

Teorema 21.2 *Sia T lo stimatore bayesiano di $g(\theta)$. Se T è corretto, allora*

$$\iint_{\Theta\times\Omega} \big|T(\omega)-g(\theta)\big|^2 Q(\mathrm{d}\theta,\mathrm{d}\omega) = 0.$$

DIMOSTRAZIONE. Consideriamo la v.a. $X(\theta,\omega) = T(\omega)-g(\theta)$ come elemento di $L^2(\Theta\times\Omega,\mathcal{T}\otimes\mathcal{F},Q)$. Si ha

$$E[X|\tilde{\mathcal{F}}] = E[T|\tilde{\mathcal{F}}] - E[g|\tilde{\mathcal{F}}] = 0,$$

perché $E[g|\tilde{\mathcal{F}}] = T$, come detto sopra, e dunque T è $\tilde{\mathcal{F}}$-misurabile, quindi

$$E[T|\tilde{\mathcal{F}}] = T = E[g|\tilde{\mathcal{F}}].$$

Pertanto X è ortogonale (in $L^2(\Theta\times\Omega,\mathcal{T}\otimes\mathcal{F},Q)$) ad ogni funzione V che sia $\tilde{\mathcal{F}}$-misurabile. Infatti

$$E[XV] = E\big[E[XV|\tilde{\mathcal{F}}]\big] = E\big[VE[X|\tilde{\mathcal{F}}]\big] = 0.$$

D'altra parte, si ha anche

$$E[X|\tilde{\mathcal{T}}] = 0,$$

perché l'applicazione $(\theta, \omega) \mapsto g(\theta)$ è $\tilde{\mathcal{T}}$- misurabile (per definizione della σ-algebra $\tilde{\mathcal{T}}$) e dunque $E[g|\tilde{\mathcal{T}}] = g$; inoltre, per il Teorema 19.6, risulta

$$E[T|\tilde{\mathcal{T}}](\theta) = \int_\Omega T(\omega) P(\theta, \mathrm{d}\omega) = E^\theta[T] = g(\theta),$$

perché T è corretto per ipotesi. Dunque

$$E[X|\tilde{\mathcal{T}}] = E[T|\tilde{\mathcal{T}}] - E[g|\tilde{\mathcal{T}}] = g - g = 0.$$

Ne segue che X è ortogonale (in $L^2(\Theta \times \Omega, \mathcal{T} \otimes \mathcal{F}, Q)$) ad ogni funzione V che sia $\tilde{\mathcal{T}}$-misurabile. Allora X è ortogonale anche a se stesso, perché è somma di $(\theta, \omega) \mapsto g(\theta)$ (che è $\tilde{\mathcal{T}}$-misurabile) e di $(\theta, \omega) \mapsto T(\omega)$ (che è $\tilde{\mathcal{F}}$-misurabile). Questo significa che X ha norma nulla in $L^2(\Theta \times \Omega, \mathcal{T} \otimes \mathcal{F}, Q)$, che è esattamente l'affermazione del Teorema.

□

Osservazione 21.3 Notiamo che

$$\iint_{\Theta \times \Omega} \big|T(\omega) - g(\theta)\big|^2 Q(\mathrm{d}\theta, \mathrm{d}\omega) = \int_\Theta \nu(\mathrm{d}\theta) \int_\Omega \big|T(\omega) - g(\theta)\big|^2 P^\theta(\mathrm{d}\omega).$$

Quindi, se T è corretto, per il risultato precedente esiste un $N \in \mathcal{T}$ con $\nu(N) = 0$ e tale che, per ogni $\theta \in N^c$, l'insieme

$$B_\theta = \{\omega \in \Omega : T(\omega) \neq g(\theta)\}$$

è P^θ-trascurabile. Ora, se

(i) le probabilità P^θ sono tutte equivalenti;

(ii) N^c contiene almeno due elementi;

(iii) g non è costante su N^c

questo non è possibile. Infatti, siano θ_0 e θ_1 due elementi di N^c tali che $g(\theta_0) \neq g(\theta_1)$. Allora, dato che

$$\{\omega \in \Omega : T(\omega) = g(\theta_1)\} \subseteq B_{\theta_0},$$

e B_{θ_0} è P^{θ_0}-trascurabile, si ha anche

$$P^{\theta_0}\Big(\{\omega \in \Omega : T(\omega) = g(\theta_1)\}\Big) = 0;$$

di conseguenza, dato che le P^θ sono tutte equivalenti, si ha anche

$$0 = P^{\theta_1}\Big(\{\omega \in \Omega : T(\omega) = g(\theta_1)\}\Big) = 1 - P^{\theta_1}\big(B_{\theta_1}\big) = 1,$$

(ricordando che $P^{\theta_1}\big(B_{\theta_1}\big) = 0$), e questo è assurdo.

Osservazione 21.4 I risultati enunciati restano validi se g, anziché essere limitata, è di quadrato integrabile rispetto ad ogni legge a posteriori ν^ω, e anche se g è a valori vettoriali (cioè $g : \Theta \to \mathbb{R}^k$) e si impone $C(\theta, a) = \|g(\theta) - a\|^2$.

Esempio 21.5 Sia $(X_1, \ldots, X_n)$ un campione di legge geometrica di parametro $\theta \in \Theta = (0, 1)$. Come legge a priori su Θ prendiamo la misura di Lebesgue. La verosimiglianza è

$$L(\theta; k_1, \ldots, k_n) = \prod_{i=1}^n \theta(1-\theta)^{k_i - 1} = \theta^n (1-\theta)^{\sum_{i=1}^n k_i - n},$$

e dunque lo stimatore bayesiano di $g(\theta) = \theta$ è

$$T(k_1, \dots, k_n) = \frac{\int_0^1 \theta^{n+1}(1-\theta)^{\sum_{i=1}^n k_i - n}\,\mathrm{d}\theta}{\int_0^1 \theta^{n}(1-\theta)^{\sum_{i=1}^n k_i - n}\,\mathrm{d}\theta}.$$

Ricordiamo che $\Gamma(n) = (n-1)!$ per n intero e che

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \alpha > 0, \beta > 0.$$

Dunque

$$\int_0^1 \theta^n(1-\theta)^m\,\mathrm{d}\theta = \frac{\Gamma(n+1)\Gamma(m+1)}{\Gamma(n+m+2)} = \frac{n!m!}{(n+m+1)!};$$

ricaviamo pertanto l'espressione

$$T = T(X_1, \dots, X_n) = \frac{(n+1)!\big(\sum_{i=1}^n X_i - n\big)!}{(\sum_{i=1}^n X_i - n + n + 1 + 1)!} \cdot \frac{(\sum_{i=1}^n X_i - n + n + 1)!}{n!\big(\sum_{i=1}^n X_i - n\big)!} = \frac{n+1}{\sum_{i=1}^n X_i + 2}.$$

Questo stimatore è molto simile allo stimatore di massima verosimiglianza, che risulta essere

$$\hat\theta = \frac{n}{\sum_{i=1}^n X_i}$$

(verifica per esercizio).

Esempio 21.6 Sia $(X_1, \dots, X_n)$ un campione di legge $\mathcal{E}(\theta)$, $\theta \in \Theta = (0, +\infty)$, e prendiamo come legge a priori su Θ la probabilità

$$\nu(\mathrm{d}\theta) = a\mathrm{e}^{-a\theta}\mathrm{d}\theta, \qquad a > 0$$

(prendiamo questa legge essenzialmente per rendere possibili i conti). Lo stimatore bayesiano di $g(\theta) = \theta$ è

$$T = \frac{\int_0^{+\infty} \theta^{n+1}\mathrm{e}^{-\theta(\sum_{i=1}^n X_i + a)}\mathrm{d}\theta}{\int_0^{+\infty} \theta^{n}\mathrm{e}^{-\theta(\sum_{i=1}^n X_i + a)}\mathrm{d}\theta}.$$

Dato che

$$\int_0^{+\infty} \theta^n\mathrm{e}^{-\theta b}\mathrm{d}\theta = \frac{\Gamma(n+1)}{b^{n+1}} \underbrace{\int_0^{+\infty} \frac{b^{n+1}}{\Gamma(n+1)}\theta^n\mathrm{e}^{-\theta b}\mathrm{d}\theta}_{=1} = \frac{n!}{b^{n+1}},$$

si ottiene

$$T = T(X_1, \dots, X_n) = \frac{n+1}{\sum_{i=1}^n X_i + a}.$$

Di nuovo, questo stimatore è molto simile allo stimatore di massima verosimiglianza, che è

$$\hat\theta = \frac{n}{\sum_{i=1}^n X_i}$$

(ved. Esempio 8.4).

———————————————o———————————————

Nel seguito ci interesserà il concetto di *mediana*, di cui ora parleremo brevemente.

Sia μ una misura di probabilità su $\mathbb{R}$. Alcuni test definiscono la mediana di μ come un numero reale $\mathfrak{m}$ tale che

$$\mu\big((-\infty,\mathfrak{m}]\big) = \mu\big([\mathfrak{m},+\infty\big) = \frac{1}{2}.$$

Esprimendoci in termini della funzione di ripartizione F di μ, con questa definizione una mediana è un valore $\mathfrak{m}$ tale che

$$F(\mathfrak{m}) = F(\mathfrak{m}^-)(:= \lim_{t\to\mathfrak{m}^-} F(t)) = \frac{1}{2}.$$

Da queste relazioni segue che nel punto $\mathfrak{m}$ la funzione F deve essere continua (e deve valere l'uguaglianza $F(\mathfrak{m}) = \frac{1}{2}$). Dunque è facile capire che, con questa definizione, una mediana può non esistere, e questo non è soddisfacente; un esempio semplice è quello della legge μ tale che $\mu(\{0\}) = \mu(\{1\}) = \mu(\{2\}) = \frac{1}{3}$. Sarebbe naturale dire che la mediana di questa legge è il numero $\mathfrak{m} = 1$, ma questo valore non rispetta la definizione data sopra.

Allora una migliore definizione è la seguente

Definizione 21.7 Si chiama *mediana* della legge di probabilità μ ogni valore $\mathfrak{m}$ tale che

$$\mu\big((-\infty,\mathfrak{m}]\big) \geq \frac{1}{2} \qquad \text{e} \qquad \mu\big([\mathfrak{m},+\infty)\big) \geq \frac{1}{2}.$$

In termini di F, una mediana m deve dunque verificare le condizioni

$$F(\mathfrak{m}^-) \leq \frac{1}{2} \leq F(\mathfrak{m}). \tag{31}$$

È facile dimostrare che, con questa definizione, una mediana esiste sempre. Infatti le relazioni (2) enunciate dopo la Proposizione 11.3, applicate con $\alpha = \frac{1}{2}$, diventano esattamente le (31), dunque $F^{\leftarrow}(\frac{1}{2})$ (quantile di ordine $\frac{1}{2}$ di F) è una mediana per μ (per qualsiasi μ). Tuttavia la mediana non sempre è unica. Ad esempio, la legge di densità

$$f(x) = \begin{cases} 1 & \text{per } 0 \leq x \leq \frac{1}{2} \text{ oppure per } 1 \leq x \leq \frac{3}{2} \\ 0 & \text{altrove} \end{cases}$$

ha come mediana ogni valore $\mathfrak{m}$ con $\frac{1}{2} \leq \mathfrak{m} \leq 1$.

In generale, non è difficile vedere che l'insieme delle mediane di una legge μ è un intervallo chiuso non vuoto $[\mathfrak{m}_0, \mathfrak{m}_1]$. (SUGGERIMENTO: porre $\mathfrak{m}_0 = \sup\{x \in \mathbb{R} : F(x) < \frac{1}{2}\}$; $\mathfrak{m}_1 = \sup\{x \in \mathbb{R} : F(x) \leq \frac{1}{2}\}$).

Si capisce che ogni mediana $\mathfrak{m}$ è un "indice di centralità" della legge μ, come la media m. Dunque è naturale chiedersi quali relazioni intercorrano tra m e $\mathfrak{m}$. Nel caso di una legge ν diffusa e simmetrica rispetto ad un dato valore m_0, media e mediana coincidono con m_0; in generale, però, m e $\mathfrak{m}$ sono due valori differenti, (trovare esempi), se non altro per il fatto che la media è unica, mentre le mediane possono essere in numero infinito. Tuttavia

Esercizio 21.8 Dimostrare che la media m, ogni mediana $\mathfrak{m}$ e la varianza σ^2 (supposta esistente) di una legge ν verificano la diseguaglianza

$$\big|m - \mathfrak{m}\big| \leq \sigma.$$

SOLUZIONE (di Dario Trevisan). Sia X una v.a. definita sullo spazio di probabilità $(\Omega, \mathcal{F}, P)$ e avente legge ν. Facciamo vedere innanzitutto che, per ogni mediana $\mathfrak{m}$, si ha

$$\inf_{a\in\mathbb{R}} E\big[|X-a|\big] = E\big[|X-\mathfrak{m}|\big]. \tag{32}$$

Senza perdere in generalità , si può supporre che $\mathfrak{m} = 0$: ponendo $Y = X - \mathfrak{m}$ e $a' = a - \mathfrak{m}$, Y ha mediana 0; basta allora dimostrare che

$$\inf_{a' \in \mathbb{R}} E\big[|Y - a'|\big] = E\big[|Y|\big].$$

Basta inoltre dimostrare che

$$\inf_{a<0} E\big[|Y - a|\big] = E\big[|Y|\big],$$

dato che, se vale la relazione precedente, si ha

$$\inf_{a>0} E\big[|Y - a|\big] = \inf_{a>0} E\big[|(-Y) - (-a)|\big] = \inf_{a<0} E\big[|(-Y) - a|\big] = E\big[|-Y|\big] = E\big[|Y|\big].$$

Per $a < 0$ e per $x \in \mathbb{R}$ vale la diseguaglianza (facile da dimostrare)

$$|x| \le |x - a| + a\big(1_{[0,+\infty)}(x) - 1_{(-\infty,0)}(x)\big).$$

Sostituiamo $Y(\omega)$ al posto di x; osservando che

$$1_{[0,+\infty)}(Y(\omega)) = 1_{\{Y \ge 0\}}(\omega), \qquad 1_{(-\infty,0)}(Y(\omega)) = 1_{\{Y<0\}}(\omega),$$

si ottiene

$$|Y| \le |Y - a| + a\big(1_{\{Y \ge 0\}} - 1_{\{Y<0\}}\big);$$

passando infine alla speranza si trova

$$E[|Y|] \le E[|Y - a|] + a\big(P(Y \ge 0) - P(Y < 0)\big) \le E[|Y - a|],$$

perché $a < 0$ e, per la definizione di mediana,

$$P(Y \ge 0) - P(Y < 0) = 1 - 2P(Y < 0) \ge 1 - 2 \cdot \frac{1}{2} = 0.$$

Possiamo ora dimostrare la diseguaglianza richiesta. L'asserto risulta dalla catena di relazioni

$$|m - \mathfrak{m}| = \big|E[(X - \mathfrak{m})]\big| \le E[|X - \mathfrak{m}|] \le E[|X - m|] \le \sqrt{E[|X - m|^2]} = \sigma.$$

□

———————————o———————————

Torniamo agli stimatori bayesiani.

Osservazione 21.9 Talvolta ha interesse prendere in considerazione funzioni costo differenti dal costo quadratico. Ad esempio, se si pone $C(\theta, a) = |\theta - a|$, si può mostrare che la decisione bayesiana relativa alla legge di probabilità ν è la mediana di ν. Precisamente

Proposition 21.10 *Se $\mathfrak{m}$ è una mediana per la legge di probabilità ν, allora, per ogni $a \in \mathbb{R}$,*

$$\int_{\mathbb{R}} |\theta - \mathfrak{m}| \nu(\mathrm{d}\theta) \le \int_{\mathbb{R}} |\theta - a| \nu(\mathrm{d}\theta)$$

in particolare, se $\mathfrak{m}_1$ e $\mathfrak{m}_2$ sono due mediane per ν, si ha

$$\int_{\mathbb{R}} |\theta - \mathfrak{m}_1| \nu(\mathrm{d}\theta) = \int_{\mathbb{R}} |\theta - \mathfrak{m}_2| \nu(\mathrm{d}\theta).$$

La dimostrazione di questo risultato è identica a quella della relazione (32): basta sostituire in quest'ultima lo spazio $(\Omega, \mathcal{F}, P)$ con lo spazio $(\Theta, \mathcal{T}, \nu)$ e la v.a. $\omega \mapsto X(\omega)$ con la fuzione misurabile $\theta \mapsto \theta$.

Esercizio 21.11 Sia X (unica osservazione) una v.a. avente legge $\mathcal{U}\big([0,\theta]\big)$, con $\theta \in \Theta = \mathbb{R}^+$. Scegliamo come legge a priori su $(\mathbb{R}^+, \mathcal{B}\big(\mathbb{R}^+\big)$

$$\nu(\mathrm{d}\theta) = \theta \mathrm{e}^{-\theta} \mathrm{d}\theta.$$

Trovare lo stimatore bayesiano di $g(\theta) = \theta$ prendendo come costo

(i) $C(\theta, a) = |\theta - a|^2$;

(ii) $C(\theta, a) = |\theta - a|$.

Soluzione. (i) Si ha

$$L(\theta, x) = \frac{1}{\theta} 1_{[0,\theta]}(x) = \frac{1}{\theta} 1_{[x,+\infty)}(\theta).$$

Quindi

$$\int_0^{+\infty} L(\tau, X)\nu(\mathrm{d}\tau) = \int_0^{+\infty} \frac{1}{\tau} 1_{[X,+\infty)}(\tau)\tau \mathrm{e}^{-\tau}\mathrm{d}\tau = \int_X^{+\infty} \mathrm{e}^{-\tau}\mathrm{d}\tau = \mathrm{e}^{-X};$$

$$\int_0^{+\infty} g(\theta)L(\theta, X)\nu(\mathrm{d}\theta) = \int_0^{+\infty} \theta \cdot \frac{1}{\theta} 1_{[X,+\infty)}(\theta)\theta \mathrm{e}^{-\theta}\mathrm{d}\theta = \int_X^{+\infty} \theta \mathrm{e}^{-\theta}\mathrm{d}\theta = \mathrm{e}^{-X}(X+1).$$

Ricordiamo che, nel caso che $\int_0^{+\infty} L(\tau, X)\nu(\mathrm{d}\tau) \neq 0$, lo stimatore bayesiano di $g(\theta)$ è dato dalla formula

$$T = \frac{\int_0^{+\infty} g(\theta)L(\theta, X)\nu(\mathrm{d}\theta)}{\int_0^{+\infty} L(\tau, X)\nu(\mathrm{d}\tau)};$$

Dunque in questo caso si ottiene

$$T = \frac{\mathrm{e}^{-X}(X+1)}{\mathrm{e}^{-X}} = X + 1.$$

(ii) In questo secondo caso, per la proposizione precedente, sappiamo che T è dato dalla mediana della legge a posteriori ν^X, la quale è assolutamente continua rispetto alla legge a priori ν con densità

$$f(\theta, X) = \frac{L(\theta, X)}{\int_0^{+\infty} L(\tau, X)\nu(\mathrm{d}\tau)} = \frac{\frac{1}{\theta} 1_{[X,+\infty)}(\theta)}{\mathrm{e}^{-X}} = \frac{\mathrm{e}^X}{\theta} 1_{[X,+\infty)}(\theta).$$

Si tratta di una legge assolutamente continua, dunque continua, e pertanto la mediana è l'unico valore $\mathfrak{m}$ tale che

$$\frac{1}{2} = \nu^X\big([\mathfrak{m}, +\infty)\big) = \int_{\mathfrak{m}}^{+\infty} f(\theta, X)\nu(\mathrm{d}\theta) = \int_{\mathfrak{m}}^{+\infty} \frac{\mathrm{e}^X}{\theta} 1_{[X,+\infty)}(\theta) \cdot \theta \mathrm{e}^{-\theta}\mathrm{d}\theta = \mathrm{e}^X \int_{\mathfrak{m}\vee X}^{+\infty} \mathrm{e}^{-\theta}\mathrm{d}\theta.$$

D'altra parte, se $\mathfrak{m} \leq X$, l'equazione precedente non ha soluzioni, perché il suo secondo membro vale 1. Se invece $\mathfrak{m} > X$, l'equazione diventa

$$\frac{1}{2} = \mathrm{e}^X \int_{\mathfrak{m}}^{+\infty} \mathrm{e}^{-\theta}\mathrm{d}\theta = \mathrm{e}^{X-\mathfrak{m}},$$

da cui si ricava

$$\mathfrak{m} = T = X + \log 2.$$

22 Test dal punto di vista bayesiano

In un problema di test dell'ipotesi $H_0 : \theta \in \Theta_0$ contro l'alternativa $H_1 : \theta \in \Theta_1$ in un contesto bayesiano, Θ_0 e Θ_1 sono supposti misurabili, perché ora su Θ c'è una σ-algebra $\mathcal{T}$. L'insieme delle azioni è $A = \{0, 1\}$ (0 = accetto H_0, 1 = respingo H_0). Il costo è definito assegnando due costi differenti: $c_0 > 0$ all'errore di prima specie e $c_1 > 0$ a quello di seconda specie. Precisamente si pone

$$C(\theta, 1) = \begin{cases} c_0 & \text{per} \quad \theta \in \Theta_0 \\ 0 & \text{per} \quad \theta \in \Theta_1; \end{cases}$$

$$C(\theta, 0) = \begin{cases} 0 & \text{per} \quad \theta \in \Theta_0 \\ c_1 & \text{per} \quad \theta \in \Theta_1. \end{cases}$$

Volendo riassumere in una sola formula:

$$C(\theta, a) = \begin{cases} ac_0 & \text{per} \quad \theta \in \Theta_0 \\ (1-a)c_1 & \text{per} \quad \theta \in \Theta_1. \end{cases}$$

Questa formula assegna il costo anche nel caso di test aleatori, in cui l'insieme delle azioni è $A = [0, 1]$.

Mentre in contesto non bayesiano la discrezionalità dello statistico consiste nello scegliere il livello desiderato, in contesto bayesiano consiste nello scegliere, oltre alla legge a priori, i due costi c_0 e c_1, o piuttosto il numero $c = \frac{c_1}{c_0}$, detto *rapporto dei costi.*

Se ρ è una probabilità su $(\Theta, \mathcal{T})$, la perdita media è in questo caso

$$\int_\Theta C(\theta, a)\rho(\mathrm{d}\theta) = ac_0\rho(\Theta_0) + (1-a)c_1\rho(\Theta_1) = a\big\{c_0\rho(\Theta_0) - c_1\rho(\Theta_1)\big\} + c_1\rho(\Theta_1),$$

e la regola di decisione è pertanto

$$d(\rho) = \begin{cases} 0 & \text{se} \quad \rho(\Theta_0) > c\rho(\Theta_1) \\ 1 & \text{se} \quad \rho(\Theta_0) < c\rho(\Theta_1) \\ \text{indifferente} & \text{se} \quad \rho(\Theta_0) = c\rho(\Theta_1). \end{cases}$$

La funzione test bayesiana (cioè la decisione bayesiana relativa alla legge a posteriori ν^ω) è di conseguenza

$$\Phi(\omega) = d(\nu^\omega) = \begin{cases} 0 & \text{se} \quad \nu^\omega(\Theta_0) > c\nu^\omega(\Theta_1) \\ 1 & \text{se} \quad \nu^\omega(\Theta_0) < c\nu^\omega(\Theta_1) \\ \text{indifferente} & \text{se} \quad \nu^\omega(\Theta_0) = c\nu^\omega(\Theta_1), \end{cases}$$

a condizione naturalmente che si tratti di una funzione misurabile.

Notiamo che si prende la decisione 1 (cioè si respinge H_0) nel caso che $\nu^\omega(\Theta_0) < c\nu^\omega(\Theta_1)$, che, supponendo $\int_\Theta L(\theta, \omega)\nu(\mathrm{d}\theta) > 0$, significa

$$\frac{\int_{\Theta_0} L(\theta, \omega)\nu(\mathrm{d}\theta)}{\int_\Theta L(\theta, \omega)\nu(\mathrm{d}\theta)} < c\frac{\int_{\Theta_1} L(\theta, \omega)\nu(\mathrm{d}\theta)}{\int_\Theta L(\theta, \omega)\nu(\mathrm{d}\theta)},$$

o anche

$$\int_{\Theta_0} L(\theta, \omega)\nu(\mathrm{d}\theta) < c\int_{\Theta_1} L(\theta, \omega)\nu(\mathrm{d}\theta).$$

Dunque,se $E = \{\omega \in \Omega : \nu^\omega(\Theta_0) = c\nu^\omega(\Theta_1)\}$ è $\mu-$ trascurabile, si ottiene un test di regione critica

$$D = \Big\{\omega \in \Omega : \int_{\Theta_0} L(\theta,\omega)\nu(\mathrm{d}\theta) < c\int_{\Theta_1} L(\theta,\omega)\nu(\mathrm{d}\theta)\Big\}.$$

Si può notare una vaga rassomiglianza con il test del rapporto di verosimiglianza (se si sostituisce $\int_{\Theta_0} L(\theta,\omega)\nu(\mathrm{d}\theta)$ con $\sup_{\theta\in\Theta_0} L(\theta,\omega)$ e $\int_{\Theta_1} L(\theta,\omega)\nu(\mathrm{d}\theta)$ con $\sup_{\theta\in\Theta_1} L(\theta,\omega)$ si ottiene infatti la regione critica del test del rapporto di verosimiglianza).

Se E non è trascurabile, un test bayesiano è un test con una qualunque regione critica $D \in \mathbb{F}$ tale che

$$\{\omega \in \Omega : \nu^\omega(\Theta_0) < c\nu^\omega(\Theta_1)\} \subseteq D \subseteq \{\omega \in \Omega : \nu^\omega(\Theta_0) \le c\nu^\omega(\Theta_1)\}$$

Esempio 22.1 Se $\Theta = \{0,1\}$, $H_0 : \theta = 0$, $H_1 : \theta = 1$ e $\nu(\{0\})$, $\nu(\{1\})$ è la legge a priori, la funzione test è

$$\Phi(\omega) = \begin{cases} 0 & \text{se} \quad L(0,\omega) > \kappa L(1,\omega) \\ 1 & \text{se} \quad L(0,\omega) < \kappa L(1,\omega) \\ \text{indifferente} & \text{se} \quad L(0,\omega) = \kappa L(1,\omega), \end{cases}$$

dove $\kappa = \frac{c_1\nu(\{1\})}{c_0\nu(\{0\})}$. Si tratta dunque di un test di Neyman-Pearson.

———————————o———————————

Chiudiamo questa parte di trattazione con un

Esempio riassuntivo

Sia μ^θ la legge su $[0,1]$ avente densità $f^\theta(x) = \theta x^{\theta-1}$ (rispetto alla misura di Lebesgue μ su $[0,1]$), con $\theta \in \mathbb{R}^+$. Sia $(X_1,\dots,X_n)$ un campione di legge μ^θ, con $n \ge 2$.

(1) Trovare una statistica esaustiva. È completa?

(2) Trovare lo stimatore di massima verosimiglianza.

(3) Trovare uno stimatore corretto di θ e dire se è ottimale.

(4) Trovare un test unilaterale dell'ipotesi $H_0 : \theta \le 1$ contro $H_1 : \theta > 1$ di livello $\alpha = 0,05$.

(5) Trovare uno stimatore bayesiano di θ per la legge a priori $\nu(\mathrm{d}\theta) = \mathrm{e}^{-\theta}\mathrm{d}\theta$.

(6) Trovare un test bayesiano dell'ipotesi $H_0 : \theta = 1$ contro $H_1 : \theta \ne 1$ con rapporto dei costi c e con legge a priori $\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\mu$, dove δ_1 è la misura di Dirac concentrata in 1 e $\mu(\mathrm{d}\theta) = \mathrm{e}^{-\theta}\mathrm{d}\theta$.

Soluzione. (1) La verosimiglianza del campione è

$$L(\theta; x_1,\dots,x_n) = \theta^n\Big(\prod_{i=1}^n x_i\Big)^{\theta-1} = \exp\Big((\theta-1)\sum_{i=1}^n \log x_i + n\log\theta\Big).$$

Si tratta come si vede di un modello esponenziale (cambio di parametro $\theta \mapsto \theta - 1$), e quindi $T = \sum_{i=1}^n \log X_i$ è una statistica esaustiva completa (confrontare con l'esempio (c) delle statistiche esaustive).

(2) L'equazione di massima verosimiglianza è

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Big((\theta-1)\sum_{i=1}^{n}\log x_i + n\log\theta\Big)\Big|_{\theta=\hat\theta} = 0,$$

e cioè

$$\sum_{i=1}^{n}\log x_i + \frac{n}{\hat\theta} = 0,$$

da cui si ricava

$$\hat\theta = -\frac{n}{\sum_{i=1}^{n}\log X_i},$$

(osservazione: $\hat\theta > 0$, μ-q.c. perché $X_i \in (0,1)$ per ogni i, μ-q.c.).
(3)Cerchiamo la legge di $-\sum_{i=1}^{n}\log X_i$ sotto P^θ. Il singolo addendo $-\log X_i$ ha legge $\mathcal{E}(\theta)$ (verifica per esercizio). Quindi $-\sum_{i=1}^{n}\log X_i$ (somma di n v.a. indipendenti tutte con legge $\mathcal{E}(\theta) = \Gamma(1,\theta)$) ha legge $\Gamma(n,\theta)$. Ispirandoci alla forma dello stimatore di massima verosimiglianza, proviamo a calcolare

$$E\left[-\frac{1}{\sum_{i=1}^{n}\log X_i}\right] = \int_0^{+\infty}\frac{1}{x}\cdot\frac{\theta^n}{\Gamma(n)}x^{n-1}\mathrm{e}^{-\theta x}\mathrm{d}x = \frac{\theta^n}{\Gamma(n)}\int_0^{+\infty}x^{n-2}\mathrm{e}^{-\theta x}\mathrm{d}x$$
$$= \frac{\theta^n}{\Gamma(n)}\cdot\frac{\Gamma(n-1)}{\theta^{n-1}}\underbrace{\int_0^{+\infty}\frac{\theta^{n-1}}{\Gamma(n-1)}x^{n-2}\mathrm{e}^{-\theta x}\mathrm{d}x}_{=1} = \frac{\theta}{n-1},$$

perché la funzione integranda è la densità $\Gamma(n-1,\theta)$ e dunque il suo integrale su $\mathbb{R}$ vale 1.
Se ne deduce che la statistica

$$U = -\frac{n-1}{\sum_{i=1}^{n}\log X_i}$$

è uno stimatore corretto di θ, funzione della statistica esaustiva completa $T = -\sum_{i=1}^{n}\log X_i$. Per il Teorema 4.5, si tratta dunque di uno stimatore ottimale tra gli stimatori corretti.

(4) Il modello è esponenziale, e quindi a rapporto di verosimiglianza crescente. Direttamente

$$\frac{L(\theta_2)}{L(\theta_1)} = \Big(\frac{\theta_2}{\theta_1}\Big)^n\Big(\prod_{i=1}^{n}X_i\Big)^{\theta_2-\theta_1},$$

che è a raporto di verosimiglianza crescente rispetto a $V = \prod_{i=1}^{n}X_i$. In queste condizion i sappiamo che un test unilaterale ha una regione critica della forma

$$D = \big\{V > c\big\} = \Big\{\prod_{i=1}^{n}X_i > c\Big\}$$

(con $c \in (0,1)$ perché $X_i \in (0,1)$ per ogni i) tale che

$$P^1\Big(\prod_{i=1}^{n}X_i > c\Big) = P^1\Big(-\sum_{i=1}^{n}\log X_i < -\log c\Big) = 0,05.$$

Sotto P^1, $-\sum_{i=1}^{n}\log X_i$ ha legge $\Gamma(n,1)$, e quindi dobbiamo trovare c in modo che

$$P^1\Big(-\sum_{i=1}^{n}\log X_i < -\log c\Big) = \int_0^{-\log c}\frac{1}{(n-1)!}x^{n-1}\mathrm{e}^{-x}\mathrm{d}x = 0,05.$$

A questo punto si procede con metodi numerici.

(5) Dalla teoria, sappiamo che lo stimatore bayesiano di θ, con la misura a priori assegnata, è

$$W = \frac{\int_0^{+\infty} \theta^{n+1} \big(\prod_{i=1}^n X_i\big)^{\theta-1} \mathrm{e}^{-\theta}\,\mathrm{d}\theta}{\int_0^{+\infty} \theta^{n} \big(\prod_{i=1}^n X_i\big)^{\theta-1} \mathrm{e}^{-\theta}\,\mathrm{d}\theta}.$$

Si ha in generale, per $0 < b < \mathrm{e}$,

$$\begin{aligned}\int_0^{+\infty} \theta^n b^\theta \mathrm{e}^{-\theta}\,\mathrm{d}\theta &= \int_0^{+\infty} \theta^n \mathrm{e}^{\theta \log b} \mathrm{e}^{-\theta}\,\mathrm{d}\theta = \int_0^{+\infty} \theta^n \mathrm{e}^{-\theta(1-\log b)}\,\mathrm{d}\theta \\ &= \frac{\Gamma(n+1)}{(1-\log b)^{n+1}} \underbrace{\int_0^{+\infty} \frac{(1-\log b)^{n+1}}{\Gamma(n+1)} \theta^n \mathrm{e}^{-\theta(1-\log b)}\,\mathrm{d}\theta}_{=1} = \frac{\Gamma(n+1)}{(1-\log b)^{n+1}} = \frac{n!}{(1-\log b)^{n+1}},\end{aligned}$$

perché la funzione nell'ultimo integrale è la densità $\Gamma(n+1, 1-\log b)$, e quindi il suo integrale vale 1. Dunque

$$W = \frac{\big(1 - \sum_{i=1}^n \log X_i\big)^{n+1}}{n!} \cdot \frac{(n+1)!}{\big(1 - \sum_{i=1}^n \log X_i\big)^{n+2}} = \frac{n+1}{1 - \sum_{i=1}^n \log X_i}.$$

(6) La regione critica è

$$D = \Big\{\omega \in \Omega : \int_{\Theta_0} L(\theta, \omega)\nu(\mathrm{d}\theta) < c \int_{\Theta_1} L(\theta, \omega)\nu(\mathrm{d}\theta)\Big\},$$

che con i nostri dati ($\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\mathrm{e}^{-\theta}\mathrm{d}\theta$, $\Theta_0 = \{1\}$, $\Theta_1 = (0,1) \cup (1, +\infty)$) diventa

$$D = \Big\{\omega \in \Omega : \frac{1}{2} L(1, \omega) < \frac{c}{2} \int_0^{+\infty} L(\theta, \omega)\mathrm{e}^{-\theta}\mathrm{d}\theta\Big\} = \Big\{\omega \in \Omega : 1 < c \int_0^{+\infty} \theta^n \Big(\prod_{i=1}^n X_i\Big)^{\theta-1} \mathrm{e}^{-\theta}\mathrm{d}\theta\Big\}.$$

Dal conto fatt0 sopra si ricava che

$$\int_0^{+\infty} \theta^n b^{\theta-1} \mathrm{e}^{-\theta}\,\mathrm{d}\theta = \frac{n!}{b(1-\log b)^{n+1}},$$

e dunque la regione critica è

$$D = \left\{\omega \in \Omega : 1 < c \cdot \frac{n!}{\big(\prod_{i=1}^n X_i\big)\big(1 - \sum_{i=1}^n \log X_i\big)^{n+1}}\right\}$$

Per vedere come è fatta la regione critica, bisogna risolvere la diseguaglianza

$$b(1-\log b)n + 1 < \frac{n!}{c}, \qquad 0 < b < 1,$$

che si può studiare qualitattivamente disegnando il garfico di $b \mapsto b(1-\log b)n + 1$.

Esercizio 22.2 Studiare il test $H_0 : \theta = 1$ contro $\theta \neq 1$ per un campione di taglia n e legge $\mathcal{E}(\theta)$, $\theta \in \Theta = \mathbb{R}^+$, prendendo come legge a priori $\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\mathrm{e}^{-\theta}\mathrm{d}\theta$.

Soluzione. Si ha

$$L(\theta; X_1, \ldots, X_n) = \theta^n \mathrm{e}^{-\theta(\sum_{i=1}^n X_i)}; \qquad \int_{\Theta_0} L(\theta; X_1, \ldots, X_n)\nu(\mathrm{d}\theta) = \frac{1}{2}\mathrm{e}^{-(\sum_{i=1}^n X_i)};$$

$$\int_{\Theta_1} L(\theta; X_1, \dots, X_n)\nu(\mathrm{d}\theta) = \frac{1}{2}\int_0^{+\infty} \theta^n \mathrm{e}^{-(\sum_{i=1}^n X_i+1)}\,\mathrm{d}\theta = \frac{1}{2}\cdot\frac{(n+1)!}{(\sum_{i=1}^n X_i+1)^{n+1}};$$

la regione critica è

$$\left\{\mathrm{e}^{-(\sum_{i=1}^n X_i)}\big(\sum_{i=1}^n X_i+1\big)^{n+1} < c(n+1)!\right\};$$

studiando la funzione

$$t \mapsto (t+1)^{n+1}\mathrm{e}^{-t}$$

si vede facilmente che essa è del tipo

$$D = \left\{\sum_{i=1}^n X_i \le a\right\}\bigcup\left\{\sum_{i=1}^n X_i \ge b\right\},$$

doev $a < b$ soddisfano l'equazione

$$(a+1)^{n+1}\mathrm{e}^{-a} = (b+1)^{n+1}\mathrm{e}^{-b}.$$

Inoltre, sotto H_0, $\sum_{i=1}^n X_i \sim \Gamma(n,1)$, quindi, una volta assegnato il livello α, a (e di conseguenza b per la relazione precedente) si calcola risolvendo l'equazione

$$\int_0^a \frac{1}{\Gamma(n)}x^{n-1}\mathrm{e}^{-x}\,\mathrm{d}x + \int_b^{+\infty}\frac{1}{\Gamma(n)}x^{n-1}\mathrm{e}^{-x}\,\mathrm{d}x = \alpha.$$

Vedere l'Esercizio 17.7 per dei risultati analoghi con i metodi classici (non bayesiani).

23 La funzione di ripartizione empirica e il Teorema di Glivenko-Cantelli

Sia $(X_1, \dots, X_n)$ un campione di legge μ (da pensare come l'incognita del problema). Con F indicheremo la f.d.r. di μ.

Definizione 23.1 Si chiama *di ripartizione empirica* la variabile aleatoria ($x \in \mathbb{R}$ fissato)

$$\omega \mapsto F_n(x,\omega) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \le x\}}(\omega).$$

Osservazione 23.2 Sia $\omega \in \Omega$ fissato, e supponiamo che $X_1(\omega) < X_2(\omega) < \cdots < X_n(\omega)$ (in realtà non sarebbe necessario, basterebbe considerare le statistiche ordinate).
Allora, per $X_1(\omega) \le x < X_2(\omega)$

$$F_n(x,\omega) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \le x\}}(\omega) = \frac{1}{n},$$

perché $1_{\{X_1 \le x\}} = 1$ mentre $1_{\{X_i \le x\}}(\omega) = 0$ per $i \ge 2$; per $X_2(\omega) \le x < X_3(\omega)$ si ha

$$F_n(x,\omega) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \le x\}}(\omega) = \frac{2}{n},$$

perché $1_{\{X_1\le x\}} = 1_{\{X_2\le x\}} = 1$, mentre $1_{\{X_i\le x\}}(\omega) = 0$ per $i \ge 3$, e così via. Dunque, la funzione $x \mapsto F_n(x,\omega)$ è la funzione di ripartizione di una legge di probabilità discreta che assegna massa $\frac{1}{n}$ ai punti di ascissa $X_1(\omega), X_2(\omega), \dots, X_n(\omega)$, cioè della legge (aleatoria)

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)},$$

(δ_t = misura di Dirac nel punto t).

Sia x fissato. Le v.a. $1_{\{X_i\le x\}}$ sono indipendenti e hanno legge $\mathcal{B}(1, F(x))$ perché

$$P(1_{\{X_i\le x\}} = 1) = P(X_i \le x) = F(x).$$

Dunque hanno media $F(x)$ e varianza $F(x)(1-F(x))$. Quindi, per la Legge dei Grandi Numeri, per ogni x, $F_n(x,\cdot)$ converge $\mu^{\otimes\mathbb{N}}$-q.c. a $F(x)$ (dunque ne è un'approssimazione, di qui il nome di funzione di ripartizione empirica); inoltre, per il Teorema Limite Centrale, $\sqrt{n}\big(F_n(x,\cdot) - F(x)\big)$ converge in legge ad una $\mathcal{N}\big(0, F(x)(1-F(x))\big)$.

Il concetto di funzione di ripartizione empirica trova applicazione nel cosiddetto *metodo dei momenti* per trovare stimatori.

Sia p un intero positivo e sia

$$m_p(F) := \int x^p \mathrm{d}F(x)$$

il *p-esimo momento teorico* di F (definito per $F \in \big\{F : \int |x|^p \mathrm{d}F(x) < +\infty\big\}$). Il metodo dei momenti consiste nello stimare $m_p(F)$ con

$$m_p(F_n)(\omega) = \int x^p \mathrm{d}F_n(x,\omega) = \frac{\sum_{i=1}^n X_i^p(\omega)}{n}$$

(ricordare che la legge $F_n(\cdot,\omega)$ è la legge che assegna massa $\frac{1}{n}$ ai punti $X_i(\omega)$). La quantità $\frac{\sum_{i=1}^n X_i^p(\omega)}{n}$ si chiama anche *p-esimo momento empirico* di F.

Più in generale, il metodo dei momenti stima una funzione $f(m_1(F), \dots, m_p(F))$ con $f(m_1(F_n), \dots, m_p(F_n))$. Per esempio, la media $m_1(F)$ si stima con la media campionaria (o media empirica)

$$m_1(F_n) = \overline{X} = \frac{\sum_{i=1}^n X_i}{n},$$

la varianza $\sigma^2 = m_2(F) - m_1^2(F)$ si stima con la *varianza empirica*

$$\sigma_n^2 = m_2(F_n) - m_1^2(F_n) = \frac{\sum_{i=1}^n X_i^2}{n} - \Big(\frac{\sum_{i=1}^n X_i}{n}\Big)^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}$$

Questo stimatore somiglia alla varianza campionaria, che abbiamo già incontrato in varie occasioni, e cioè

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

S^2 ha il vantaggio di esser uno stimatore corretto di σ^2, come abbiamo visto nell'Esercizio 3.9 (b).

Torniamo alla funzione di ripartizione empirica. La Legge dei Grandi Numeri dice che

(i) La convergenza è puntuale in x (cioè il risultato vale per ogni fissato x);

(ii) L'evento

$$A_x = \{\omega \in \Omega : F_n(x,\omega) \to F(x)\}$$

è tale che $P(A_x) = 1$, ma dipende da x. Quindi a priori l'insieme $\cap_x A_x$ (che è l'evento su cui si ha la convergenza di $F_n(x,\omega) \to F(x)$ per ogni x), potrebbe non avere probabilità uguale a 1.

La convergenza verso $F(x)$ è però rafforzata dal seguente risultato:

Teorema 23.3 (di Glivenko-Cantelli). *Per ogni $n \in \mathbb{N}^*$ poniamo*

$$D_n(\omega) := \sup_{x\in\mathbb{R}} \big|F_n(x,\omega) - F(x)\big|.$$

Allora:

(i) Per ogni n, D_n è una variabile aleatoria;

(ii) La successione $(D_n)_{n\geq 1}$ converge q.c. a 0 per $n \to \infty$.

Osservazione 23.4 Il Teorema dice dunque che

(i) la convergenza è uniforme in x;

(ii) Esiste un evento E, tale che $P(E) = 1$, sul quale si ha convergenza uniforme. Dunque, se $\omega \in E$, si ha anche la convergenza puntuale per ogni fissato x di $F_n(x,\omega)$ verso $F(x)$.

Dimostrazione. (i) D_n è misurabile perché il sup può essere fatto sui soli razionali.

(ii) Indichiamo con $F(x^-) = \lim_{t\uparrow x} F(t)$; notazione analoga per $F_n(x,\cdot)$. Si ha

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i<x\}}(\omega) = \lim_{t\uparrow x} \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i\leq t\}}(\omega) = F_n(x^-,\omega);$$

inoltre, ancora per la Legge dei Grandi Numeri, per ogni x esiste un evento B_x con $P(B_x) = 1$, tale che $F_n(x^-,\omega) \to F(x^-)$, per ogni x.

Ricordiamo che la funzione quantile $\phi = F^{\leftarrow}$ è così definita

$$\phi(u) = F^{\leftarrow}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, \qquad u \in (0,1).$$

Essa ha, fra le altre, le due seguenti proprietà (dimostrate dopo la Proposizione 11.3)

(i) Se $u \leq v$ allora $\phi(u) \leq \phi(v)$;

(ii) $F\big(\phi(u)^-\big) \leq u \leq F\big(\phi(u)\big)$.

Sia ora k un intero ≥ 1 fissato. Poniamo $x_0 = -\infty$, $x_k = +\infty$ e, per ogni $j = 1, 2, \ldots, k-1$,

$$x_j = \phi\Big(\frac{j}{k}\Big).$$

Osserviamo che dalla prima delle relazioni (ii) (applicata a $u = \frac{j}{k}$) segue che, per $1 \leq j \leq k-1$

$$F(x_j^-) \leq \frac{j}{k},$$

mentre dalla seconda delle (ii) (applicata a $u = \frac{j-1}{k}$) segue, per $2 \le j \le k$

$$F(x_{j-1}) \ge \frac{j-1}{k}.$$

In particolare si ha

$$F(x_1^-) \le \frac{1}{k}, \qquad F(x_{k-1}) \ge 1 - \frac{1}{k} \tag{33}$$

e sottraendo, per $j = 2, \dots, k-1$ si ottiene

$$F(x_j^-) - F(x_{j-1}) \le \frac{1}{k}. \tag{34}$$

Se si pone poi per convenzione $F(x_0^-) = F(x_0) = 0$ e $F(x_k^-) = F(x_k) = 1$, usando le (33) si vede che la (34) vale per ogni $j = 1, \dots, k$.

Poniamo ora

$$R_n(\omega) = \max_{0 \le j \le k} \left\{ |F_n(x_j, \omega) - F(x_j)| \vee |F_n(x_j^-, \omega) - F(x_j^-)| \right\}$$

e

$$E_k := \bigcap_{j=0}^{k} (A_{x_j} \cap B_{x_j}).$$

Allora $P(E_k) = 1$ ed inoltre

$$\lim_{n \to \infty} R_n(\omega) = 0, \qquad \forall\, \omega \in E_k. \tag{35}$$

Per la relazione (i) si ha

$$-\infty = x_0 \le x_1 \le \cdots \le x_k = +\infty,$$

e dunque quelli non vuoti tra gli intervalli $[x_{j-1}, x_j)$, $j = 1, \dots, k$ costituiscono una partizione di $\mathbb{R}$. Sia allora $x \in \mathbb{R}$, e sia $j(= j(x))$ tale che $x_{j-1} \le x < x_j$. Dalla relazione $\{X_i \le x\} \subseteq \{X_i < x_j\}$ segue che, per ogni $\omega \in \Omega$,

$$F_n(x, \omega) \le F_n(x_j^-, \omega) \le F(x_j^-) + R_n(\omega),$$

per la definizione di R_n. Continuando, e usando la (34), si trova

$$F(x_j^-) + R_n(\omega) \le F(x_{j-1}) + \frac{1}{k} + R_n(\omega) \le F(x) + \frac{1}{k} + R_n(\omega),$$

per la non decrescenza di F. Si conclude che

$$F_n(x, \omega) \le F(x) + \frac{1}{k} + R_n(\omega).$$

In modo analogo si hanno le diseguaglianze

$$F_n(x, \omega) \ge F_n(x_{j-1,\omega}) \ge F(x_{j-1}) - R_n(\omega) \ge F(x_j^-) - \frac{1}{k} - R_n(\omega) \ge F(x) - \frac{1}{k} - R_n(\omega),$$

e quindi, in conclusione, <u>per ogni $x \in \mathbb{R}$</u>,

$$|F_n(x, \omega) - F(x)| \le \frac{1}{k} + R_n(\omega), \qquad \forall\, \omega \in \Omega,$$

o, equivalentemente,

$$D_n(\omega) \le \frac{1}{k} + R_n(\omega), \qquad \forall\, \omega \in \Omega.$$

Se allora $\omega \in E_k$, per la (35) si ottiene

$$\limsup_{n\to\infty} D_n(\omega) \leq \frac{1}{k}. \tag{36}$$

Poniamo ora

$$E = \bigcap_{k=1}^{\infty} E_k,$$

e osserviamo che $P(E) = 1$. Per ogni $\omega \in E$ si ha $\omega \in E_k$ per ogni k, e quindi la relazione (36) vale per ogni k. Passando allora in essa al limite per $k \to \infty$ si ottiene, per ogni $\omega \in E$,

$$\lim_{n\to\infty} D_n(\omega) = \limsup_{n\to\infty} D_n(\omega) = 0,$$

e il Teorema è dimostrato.

□

Per terminare, diamo una semplice e utile applicazione del Teorema di Glivenko-Cantelli 23.3. Sia $(X_1, \dots, X_n)$ un campione di legge μ su $\mathbb{R}$. Siamo interessati a stimare la mediana $\mathfrak{m}_\mu$ di μ (supposta unica).

Per ogni $\omega \in \Omega$ consideriamo la legge aleatoria

$$\mu_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

(di cui, come sappiamo $F_n(\cdot, \omega)$ è la f.d.r.); sappiamo (da quel che abbiamo detto a suo tempo) che una sua mediana è il suo quantile di ordine $\frac{1}{2}$. Poniamo dunque

$$\mathfrak{m}_{\mu_n(\omega)} = \inf \Big\{ x \in \mathbb{R} : \mu_n(\omega)\big((-\infty, x]\big) \geq \frac{1}{2} \Big\} = \inf \Big\{ x \in \mathbb{R} : F_n(x, \omega) \geq \frac{1}{2} \Big\}. \tag{37}$$

Per il Teorema di Glivenko-Cantelli 23.3, la funzione di ripartizione di $\mu_n(\omega)$, cioè $F_n(\cdot, \omega)$, converge alla funzione di ripartizione di μ, cioè F. È quindi ragionevole chiedersi se $\mathfrak{m}_{\mu_n(\omega)}$ converga in qualche senso a $\mathfrak{m}_\mu$ (in modo da esserne un buon stimatore).

Teorema 23.5 *(i) La funzione $\omega \mapsto \mathfrak{m}_{\mu_n(\omega)}$ è una variabile aleatoria.*
(ii) Se $\mathfrak{m}_\mu$ è l'unica mediana per μ, allora $P = \mu^{\otimes \mathbb{N}}$-quasi certamente, risulta

$$\mathfrak{m}_{\mu_n(\omega)} \to \mathfrak{m}_\mu, \qquad n \to \infty.$$

Dimostrazione. (i) Segue facilmente dalla (37), tenendo conto che $\omega \mapsto F_n(x, \omega)$ è una variabile aleatoria.

(ii)Poiché $\mathfrak{m}_\mu$ è l'unica mediana di μ, fissato $\epsilon > 0$, esiste $\delta > 0$ tale che valgano entrambe le relazioni

$$\begin{cases} F(\mathfrak{m}_\mu - \epsilon) < \frac{1}{2} - \delta \\ F(\mathfrak{m}_\mu + \epsilon) > \frac{1}{2} + \delta \end{cases} \tag{38}$$

perché questo sistema significa che

$$0 < \delta < \Big\{ \frac{1}{2} - F(\mathfrak{m}_\mu - \epsilon) \Big\} \wedge \Big\{ F(\mathfrak{m}_\mu + \epsilon) - \frac{1}{2} \Big\}.$$

Per il Teorema di Glivenko-Cantelli 23.3, esiste $A \in \mathcal{F}$ con $P(A) = 1$ tale che, per ogni $\omega \in A$

$$\sup_{x \in \mathbb{R}} |F_n(x, \omega) - F(x)| \to 0, \qquad n \to \infty.$$

Fissato $\omega \in A$, esiste un intero $n(\omega)$ tale che, per ogni $n > n(\omega)$, si abbia

$$\sup_{x \in \mathbb{R}} |F_n(x, \omega) - F(x)| < \delta. \tag{39}$$

Per n cosiffatto, si ha

(1) $\mathfrak{m}_{\mu_n(\omega)} > \mathfrak{m}_\mu - \epsilon$. Infatti, se fosse $\mathfrak{m}_{\mu_n(\omega)} \leq \mathfrak{m}_\mu - \epsilon$, allora

$$F(\mathfrak{m}_\mu - \epsilon) \underbrace{>}_{(39)} F_n(\mathfrak{m}_\mu - \epsilon, \omega) - \delta \underbrace{\geq}_{def.\ di\ \mathfrak{m}_{\mu_n}\ (37)} \frac{1}{2} - \delta,$$

assurdo per la prima delle (38);

(2) $\mathfrak{m}_{\mu_n(\omega)} < \mathfrak{m}_\mu + \epsilon$. Infatti, se fosse $\mathfrak{m}_{\mu_n(\omega)} \geq \mathfrak{m}_\mu + \epsilon$, allora

$$F(\mathfrak{m}_\mu + \epsilon) \underbrace{<}_{(39)} F_n(\mathfrak{m}_\mu + \epsilon, \omega) + \delta \underbrace{\leq}_{def.\ di\ \mathfrak{m}_{\mu_n}\,(37)} \frac{1}{2} + \delta,$$

assurdo per la seconda delle (38).

Dalle relazioni (1) e (2) si conclude che, per ogni $n > n(\omega)$, risulta

$$|\mathfrak{m}_{\mu_n}(\omega) - \mathfrak{m}_\mu| < \epsilon,$$

e cioè che

$$\lim_{n\to\infty} \mathfrak{m}_{\mu_n}(\omega) = \mathfrak{m}_\mu.$$

Poiché questo accade per ogni $\omega \in A$ (che ha probabilità uguale a 1), si ha la tesi.

□

24 Il test del χ^2

Sullo spazio $(\Omega, \mathcal{F}, P)$ sia $(X_1, \ldots, X_n)$ un campione di variabili a valori in un insieme finito $\{1, 2, \ldots, k\}$. La legge di ciascuna delle v.a. X_j, $j = 1, \ldots, n$, è data dal vettore $q = (q_1, \ldots, q_k)$, dove $q_i = P(X_j = i)$, $i = 1, \ldots, k$. Si ha ovviamente $q_i \geq 0$ per ogni $i = 1, \ldots, k$, e $\sum_{i=1}^k q_i = 1$.
Il *test del chi-quadro* si usa per verificare se la legge q coincide con una legge assegnata (e ovviamente nota) $p = (p_1, \ldots, p_k)$, tale che $p_i > 0$ per ogni $i = 1, \ldots, k$.
L'ipotesi è $H_0 : q = p$, l'alternativa $H_1 : q \neq p$. Il test si basa sulla statistica

$$T = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}, \qquad \text{dove} \qquad N_i = \sum_{j=1}^n 1_{\{X_j = i\}}.$$

Ovviamente la v.a. N_i indica il numero di osservazioni X_j che hanno dato valore i. N_i viene chiamato anche *effettivo empirico del valore i* . Dalla Legge dei Grandi Numeri segue che $N_i \approx nq_i$; per analogia chiameremo allora *effettivo teorico del valore i* la quantità np_i.

Ci serve un teorema (Teorema di Pearson) che dice cosa accade quando la taglia n del campione tende all'infinito (nelle applicazioni questo significa che n è grande; dunque considereremo una successione infinita $X_1, X_2, X_3, \ldots$ di v.a. i.i.d. e porremo un apice/indice n a N_i /T, cioè scriveremo

$$N_i^n = \sum_{j=1}^n 1_{\{X_j = i\}}, \qquad T_n = \sum_{i=1}^k \frac{(N_i^n - np_i)^2}{np_i}.$$

Teorema 24.1 (di Pearson).
(a) Se la legge comune delle X_j è diversa da p, allora $T_n \to +\infty$, P-q.c.
(b) Se la legge comune delle X_j coincide con p, allora T_n converge in legge ad una $\chi^2(k-1)$.

Dimostrazione. (a) Per ipotesi esiste $i_0 \in \{1, 2, \ldots, k\}$ tale che $q_{i_0} \neq p_{i_0}$. Per la Legge Forte dei Garndi Numeri, al tendere di n all'infinito si ha

$$\frac{N_{i_0}^n}{n} \to q_{i_0}, \qquad P\text{-q.c.};$$

quindi

$$\frac{(N^n_{i_0} - np_{i_0})^2}{np_{i_0}} = n \cdot \frac{\left(\frac{N^n_{i_0}}{n} - p_{i_0}\right)^2}{p_{i_0}} \to +\infty, \qquad P\text{-q.c.}$$

(b) Ci servono alcuni preliminari.

Lemma 24.2 *Sia Z una v.a. vettoriale k-dimensionale avente legge $\mathcal{N}_k(\mathbf{0}, A)$, dove $A = (a_{i,j})_{i,j=1,\dots,k}$ è la matrice $k \times k$ con $a_{i,j} = \delta_{i,j} - \sqrt{p_i}\sqrt{p_j}$. Allora la v.a. $\|Z\|^2$ ha legge $\chi^2(k-1)$.*

Dimostrazione. Sia X un vettore aleatorio k-dimensionale avente legge $\mathcal{N}_k(\mathbf{0}, I_k)$, e sia E il sottospazio di $\mathbb{R}^k$ generato dal versore $\sqrt{p} := (\sqrt{p_1}, \dots, \sqrt{p_n})$. Il Teorema di Cochran 9.3 dice che $(X - X_E)$ e X_E sono indipendenti ed inoltre $\|X_E\|^2$ ha legge $\chi^2(1)$ e $\|X - X_E\|^2$ ha legge $\chi^2(k-1)$. Dunque basterà vedere che $Y = X - X_E$ ha legge $\mathcal{N}_k(\mathbf{0}, A)$. Si ha

$$Y = X - \langle X, \sqrt{p}\rangle \cdot \sqrt{p} = X - \Big(\sum_{i=1}^k X_i\sqrt{p_i}\Big)\sqrt{p},$$

ed è facile vedere che $\langle u, Y\rangle$ è una v.a. gaussiana per ogni $u \in \mathbb{R}^k$. Basta allora calcolare il vettore delle medie e la matrice di covarianza. Per ogni $i, j = 1, \dots, k$ si ha

$$E[Y_j] = E\Big[X_j - \Big(\sum_{i=1}^k X_i\sqrt{p_i}\Big)\sqrt{p_j}\Big] = E[X_j] - \Big(\sum_{i=1}^k E[X_i]\sqrt{p_i}\Big)\sqrt{p_j} = 0,$$

perché X è un vettore centrato. Inoltre

$$\begin{aligned}
E[Y_iY_j] &= E\left[\Big(X_i - \Big(\sum_{h=1}^k X_h\sqrt{p_h}\Big)\sqrt{p_i}\Big) \cdot \Big(X_j - \Big(\sum_{h=1}^k X_h\sqrt{p_h}\Big)\sqrt{p_j}\Big)\right] \\
&= \underbrace{E[X_iX_j]}_{=\delta_{i,j}} - \Big(\sum_{h=1}^k \underbrace{E[X_jX_h]}_{=\delta_{j,h}}\sqrt{p_h}\Big)\sqrt{p_i} - \Big(\sum_{h=1}^k \underbrace{E[X_iX_h]}_{=\delta_{i,h}}\sqrt{p_h}\Big)\sqrt{p_j} \\
&+ \Big(\sum_{h,r=1}^k \underbrace{E[X_hX_r]}_{=\delta_{h,r}}\sqrt{p_h}\sqrt{p_r}\Big)\sqrt{p_i}\sqrt{p_j} \\
&= \delta_{i,j} - \sqrt{p_i}\sqrt{p_j} - \sqrt{p_i}\sqrt{p_j} + \underbrace{\Big(\sum_{h=1}^k p_h\Big)}_{=1}\sqrt{p_i}\sqrt{p_j} = \delta_{i,j} - \sqrt{p_i}\sqrt{p_j}.
\end{aligned}$$

□

Richiamiamo il

Teorema 24.3 (Teorema Limite Centrale Vettoriale, abbreviato TLCV). *Sia $(Y_n)_{n\geq 1}$ una successione di vettori aleatori a valori in $\mathbb{R}^k$, i.i.d. con $E[Y_n] = \mathbf{m}$ ($\in \mathbb{R}^k$) e matrice di covarianza $\Gamma = (\gamma_{i,j})_{i,j=1,\dots,k}$, con $\gamma_{i,j} = Cov(Y_i, Y_j)$. Allora la successione di vettori aleatori*

$$\frac{Y_1 + \cdots + Y_n - n\mathbf{m}}{\sqrt{n}}$$

converge in legge ad una $\mathcal{N}_k(\mathbf{0}, \Gamma)$.

Per la dimostrazione, si veda ad esempio [2].

Infine

Esercizio 24.4 Sia $(Z_n)_{n\geq 1}$ una successione di vettori aleatori convergente in legge ad una densità $\mathcal{N}_k(\mathbf{0}, A)$, dove A è la matrice descritta nel Lemma 24.2. Allora $\|Z_n\|^2$ converge in legge ad una $\chi^2(k-1)$.

Suggerimento. Usare il Lemma 24.2.

Passiamo finalmente alla dimostrazione del punto (b) del Teorema di Pearson 24.1.

Consideriamo i vettori k-dimensionali, indipendenti ed identicamente distribuiti

$$Y_i := \Big(\frac{1}{\sqrt{p_1}}1_{\{X_i=1\}}, \dots, \frac{1}{\sqrt{p_k}}1_{\{X_i=k\}}\Big).$$

Si ha facilmente

$$E[Y_i] = \Big(\frac{p_1}{\sqrt{p_1}}, \dots, \frac{p_k}{\sqrt{p_k}}\Big) = (\sqrt{p_1}, \dots, \sqrt{p_k}) = \sqrt{p};$$

inoltre, per $r, s = 1, \dots, k$, con $r \neq s$ si ha

$$E\big[(Y_i)_r \cdot (Y_i)_s\big] = E\Big[\frac{1}{\sqrt{p_r}}1_{\{X_i=r\}} \cdot \frac{1}{\sqrt{p_s}}1_{\{X_i=s\}}\Big] = \frac{1}{\sqrt{p_r}\sqrt{p_s}}P\big(\{X_i = r\} \cap \{X_i = s\}\big) = 0,$$

dato che $\{X_i = r\} \cap \{X_i = s\} = \emptyset$.

Invece, per $r = s$ si ha

$$E\big[(Y_i)_r^2\big] = E\Big[\frac{1}{p_r}1_{\{X_i=r\}}\Big] = 1.$$

Ne segue che

$$Cov\big((Y_i)_r, (Y_i)_s\big) = E\big[(Y_i)_r \cdot (Y_i)_s\big] - E\big[(Y_i)_r\big]E\big[(Y_i)_s\big] = \begin{cases} 1 - p_r & \text{per} \quad r = s \\ -\sqrt{p_r}\sqrt{p_s} & \text{per} \quad r \neq s \end{cases} = a_{r,s},$$

dove $a_{r,s} = \delta_{r,s} - \sqrt{p_r}\sqrt{p_s}$ come nel Lemma 24.2.

Osserviamo ora che, dato che $N_j^n = \sum_{i=1}^n 1_{\{X_i=j\}}$, $j = 1, \dots, k$, si ha

$$Y_1 + \cdots + Y_n = \left(\frac{N_1^n}{\sqrt{p_1}}, \cdots, \frac{N_k^n}{\sqrt{p_k}}\right),$$

e per il TLCV 24.3, si ha allora che la successione

$$Z_n := \frac{Y_1 + \cdots + Y_n - n\sqrt{p}}{\sqrt{n}} = \left(\frac{\frac{N_1^n}{\sqrt{p_1}} - n\sqrt{p_1}}{\sqrt{n}}, \dots, \frac{\frac{N_k^n}{\sqrt{p_k}} - n\sqrt{p_k}}{\sqrt{n}}\right) = \left(\frac{N_1^n - np_1}{\sqrt{np_1}}, \dots, \frac{N_k^n - np_k}{\sqrt{np_k}}\right)$$

converge in legge ad una $\mathcal{N}_k(\mathbf{0}, A)$ e, di conseguenza, per l'Esercizio 24.4,

$$\|Z_n\|^2 = T_n$$

converge in legge ad una $\chi^2(k-1)$.

□

Torniamo finalmente al test del chi-quadro. Supponiamo che la taglia n del campione di osservazioni $(X_1, \dots, X_n)$ sia grande. Se l'ipotesi è falsa, allora la statistica T assumerà valori grandi, per il punto (a) del Teorema di Pearson 24.1; dunque ci aspettiamo una regione critica del tipo $\{T > c\}$, con c da determinare. Volendo un test di taglia uguale a α, si deve imporre che sia

$$P^{H_0}(T > c) = \alpha.$$

Ma, se H_0 è vera, per il punto (b) del teorema di Pearson 24.1 T ha asintoticamente legge $\chi^2(k-1)$, e quindi

$$P^{H_0}(T > c) \approx 1 - F_{k-1}(c) = \alpha,$$

da cui $c = \chi^2_{1-\alpha}(k-1)$, e la regione critica è dunque

$$\big\{T > \chi^2_{1-\alpha}(k-1)\big\}.$$

Osservazione 24.5 Se la legge μ delle osservazioni non è concentrata su un insieme finito di valori (come abbiamo supposto all'inizio di questo paragrafo), per verificare se il campione ha legge μ il test del chi-quadro viene usato comunque, adattandolo nel modo seguente. Si prende una funzione $\phi : \mathbb{R} \to \{1, \dots, k\}$ e si verifica se il nuovo campione $\big(\phi(X_1), \dots, \phi(X_n)\big)$ ha legge $\phi(\mu)$. Nella pratica spesso si sceglie una partizione di $\mathbb{R}$ del tipo

$$(-\infty = a_0, a_1], (a_1, a_2], \dots, (a_{k-2}, a_{k-1}], (a_{k-1}, a_k = +\infty]$$

e si pone

$$\phi(x) = \sum_{j=1}^{k} j\, 1_{(a_{j-1}, a_j]} = \text{indice dell'elemento della partizione a cui } x \text{ appartiene.}$$

In questo caso si ha, per ogni $j = 1, \dots, k$

$$\phi(\mu)\big(\{j\}\big) = \mu\big(\phi^{-1}(\{j\})\big) = P\big(X \in \phi^{-1}(\{j\})\big) = P\big(X \in (a_{j-1}, a_j]\big).$$

Osservazione 24.6 Una condizione pratica sotto la quale il test del chi-quadro viene considerato attendibile è che la numerosità n del campione sia tale che $np_i \geq 5$, per ogni $i = 1, \dots, k$.

Osservazione 24.7 Il test del chi-quadro può essere usato anche per valutare se un campione segue una legge appartenente ad una famiglia parametrizzata $p^\theta = (p_1^\theta, \dots, p_k^\theta)$, dove Θ è un aperto di $\mathbb{R}^s$, con $s < k - 1$. Si devono fare le ipotesi seguenti:

(i) le funzioni $\theta \mapsto p_i^\theta$ sono strettamente positive e di classe C^2, per ogni $i = 1, \dots, k$;

(ii) la matrice $\left(\frac{\partial}{\partial \theta_j} p_i^\theta\right)$ è di rango massimo, per ogni $i = 1, \dots, k$;

(iii) esiste una successione consistente $\hat{\theta}_n$ di stimatori di massima verosimiglianza di θ.

Notiamo che

$$L(\theta; x_1, \dots, x_n) = \big(p_1^\theta\big)^{N_1^n} \cdots \big(p_k^\theta\big)^{N_k^n}$$

Dunque, passando al logaritmo, lo stimatore di massima verosimiglianza si trova cercando il

$$\max_{\theta \in \Theta} \Big(N_1^n \log\big(p_1^\theta\big) + \cdots + N_k^n \log\big(p_k^\theta\big) \Big).$$

Sia

$$T_n(\theta) = \sum_{i=1}^{k} \frac{(N_i^n - np_i^\theta)^2}{np_i^\theta}.$$

Si può dimostrare che, sotto le ipotesi precedenti, la successione $T_n(\hat{\theta}_n)$ converge in legge verso una $\chi^2(k-s-1)$. In pratica, sostituendo ai parametri non noti $\theta = (\theta_1, \ldots, \theta_s)$ i loro stimatori di massima verosimiglianza, abbiamo stimato s parametri e questo fa diminuire di s il numero dei gradi di libertà.

Questo suggerisce una regione critica della forma $\{T(\hat{\theta}) > c\}$, e per trovare c si procede come nel caso base.

Esempio 24.8 Un caso di test del chi-quadro con parametri stimati: il test del chi-quadro per l'indipendenza.

Attenzione: le notazioni in questo esempio non corrispondono a quelle usate nella teoria.

Su un opportuno spazio $(\Omega, \mathcal{F}, P)$ siano X e Y due variabili aleatorie (a valori numerici o no), che assumono rispettivamente i valori $x_1, \ldots, x_r$ e $y_1, \ldots, y_s$; dunque la variabile aleatoria bivariata (X, Y) assume i valori (x_i, y_j), con $i = 1, \ldots, r$; $j = y_1, \ldots, y_s$.

Si vuole verificare l'ipotesi nulla

$$H_0 : X \text{ e } Y \text{ sono indipendenti.}$$

Poniamo

$$p_i = P(X = x_i), \qquad i = 1, \ldots, r;$$
$$q_j = P(Y = y_j), \qquad j = 1, \ldots, s.$$

L'ipotesi H_0 può essere allora espressa dicendo che la legge congiunta di (X, Y) coincide con il prodotto tensoriale delle leggi marginali, ovvero: per ogni coppia (i, j) si ha

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) = p_i q_j.$$

Queste sono le probabilità teoriche, e non essendo note, devono essere stimate. Per far questo, si eseguono n osservazioni indipendenti della coppia (X, Y) (cioè con la legge di (X, Y)), che indicheremo con $(X_1, Y_1), \ldots, (X_n, Y_n)$; per ogni coppia $(i, j) \in \{1, \ldots, r\} \times \{1, \ldots, s\}$ poniamo

$$O_{i,j} = \text{effettivo empirico di } (x_i, y_j).$$

(Nota. O=*observed*. Con le notazioni della teoria, dovremmo scrivere $O_{i,j} = \sum_{k=1}^{n} 1_{\{X_k = i, Y_k = j\}}$. La notazione usata qui serve a non appesantire le formule).

È evidente che le quantità

$$O_{i,\cdot} := \sum_{j=1}^{s} O_{i,j}, \qquad O_{\cdot,j} := \sum_{i=1}^{r} O_{i,j}$$

non sono altro che gli effettivi empirici di x_i e y_j, rispettivamente.

Dalla teoria sappiamo che le probabilità teoriche possono essere stimate per mezzo degli stimatori di massima verosimiglianza; è poi facile vedere che tali stimatori per p_i e q_j sono rispettivamente

$$\hat{p}_i = \frac{O_{i,\cdot}}{n}; \qquad \hat{q}_j = \frac{O_{\cdot,j}}{n}.$$

Dunque le probabilità teoriche per la coppia (X, Y) saranno stimate da

$$\hat{p_i}\hat{q_j} = \frac{O_{i,\cdot}O_{\cdot,j}}{n^2},$$

e dunque gli effettivi teorici (stimati) saranno

$$E_{i,j} := n\hat{p_i}\hat{q_j} = \frac{O_{i,\cdot}O_{\cdot,j}}{n}.$$

(E=*expected*).

La statistica di Pearson vale dunque

$$T = \sum_{\substack{1\leq i\leq r\\ 1\leq j\leq s}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = \sum_{\substack{1\leq i\leq r\\ 1\leq j\leq s}} \frac{(O_{i,j} - \frac{O_{i,\cdot}O_{\cdot,j}}{n})^2}{\frac{O_{i,\cdot}O_{\cdot,j}}{n}}.$$

Resta ora da stabilire qual è il numero di gradi di libertà della χ^2 da usare, e per questo è necessario contare quanti parametri sono stati stimati. I parametri p_i sono stati stimati per $i = 1, \ldots, r-1$ (p_r in realtà non lo è stato, poiché esso è determinato dalla relazione $p_r = 1 - (p_1 + \cdots + p_{r-1})$). Per lo stesso motivo i parametri q_j stimati sono stati in numero di $s-1$, e quindi si ha un totale di $m = r + s - 2$. Il numero k di valori assunti dalla variabile (X, Y) è evidentemente rs, e quindi la χ^2 avrà un numero di gradi di libertà pari a

$$k - 1 - m = rs - 1 - (r + s - 2) = (r - 1)(s - 1).$$

Riferimenti bibliografici

[1] Jacod, J., Protter, Ph. (2004), *Probability Essentials*, Springer.

[2] Dacunha-Castelle, D., Duflo, M., *Probability and Statistics* Vol. 1, Springer