Capitolo 1

Statistica

La statistica significa avere una qualche popolazione data da una qualche variabile aleatoria $f(x|\theta)$ significa trovare informazioni su θ. Per studiare il parametro θ servirà un campione statistico.

Definizione 1.0.1 (**Modello statistico**). *Si definisce modello statistico di una famiglia di spazi di probabilità $(\Omega, \mathbb{A}, \mathbb{P}_\theta)$ indicizzati da $\theta \in \Theta$ dove Θ viene detta insieme dei parametri.*
Ω e $\mathcal{A}$ rimangono sempre gli stessi è $\mathbb{P}$ che dipende da θ e quindi varia.

Definizione 1.0.2 (**Campione statistico**). *Un campione statistico è un insieme di v.a. $X_1, \cdots, X_n$ indipendenti e identicamente distribuite aventi tutte legge $f(x|\theta)$.*
Sia $\underline{X} = (X_1, \cdots, X_n)$ esisti di un esperimento sulla popolazione.

$$X_1 \sim f(x|\theta) \qquad \underline{X} \sim f_n(x_1, \cdots, x_n|\theta) = f(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Esempio 1.0.1. *Se $X \sim \exp(\lambda)$ allora*

$$X \sim f(x|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

Osservazione 1.0.1. *Da una popolazione finita ho un campione se le estrazioni sono fatte con reimmissione. Nel caso di estrazioni senza reimmissione ho che le X_i sono identicamente distribuite ma non sono indipendenti. Se n (numerosità del campione) è piccolo rispetto a N (numerosità della popolazione), ad esempio se ho la classe R e la classe B con $R + B = N$ numerosità delle due classi allora*

$$\mathbb{P}(X_2 = r|X_1 = r) = \frac{R-1}{N-1} \sim \frac{R}{N}$$

Che è vicino alla marginale del caso di estrazioni con reimmissione.

Esempio 1.0.2. *Data un'urna di palline numerate $\{1, \cdots, N = 1000\}$, quale è la probabilità che tutti i valori $n = 10$ siano maggiori di 200?*
Nel caso con reimmissione abbiamo che

$$\mathbb{P}(X_1 > 200, \cdots, X_n > 200) = \left(\frac{800}{1000}\right)^{10} \approx 10,74\%$$

Mentre se non c'è reimmissione

$$\mathbb{P}(X_1 > 200, \cdots, X_n > 200) = \frac{\binom{200}{10}\binom{200}{0}}{\binom{1000}{10}} \approx 10,62\%$$

1.1 Campionamento normale (Gaussiano)

Definizione 1.1.1. *Dato un campione statistico $X_1, \cdots, X_n$ allora definisco:*

- ***Media campionaria*** *come*

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

- ***Varianza campionaria*** *come*

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

Definizione 1.1.2. *Sia $\{X_n\}_{n\in\mathbb{N}}$ v.a. iid tali per cui $X_i \sim N(0,1)$ allora posto*

$$Y_n = \sum_{i=1}^{n} X_i$$

Si dice che Y segue una legge χ^2 a n gradi di libertà, indicata con $Y \sim \chi^2(n)$ se la funzione densità è tale per cui

$$f_n(x) = \frac{\lambda^n}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{k}{2}}$$

Osservazione 1.1.1. *Se $\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ inoltre valgono i seguenti fatti:*

$$\mathbb{E}[Y_n] = n$$
$$Var(Y_n) = 2n$$

Se $n > 30$ allora $\chi(n)$ si può approssimare con $N(n, 2n)$.

Lemma 1.1.1. *Se $Z \sim N(0,1)$ allora $Z^2 \sim \chi(1)$.*

Dimostrazione. Per definizione ho che

$$F_{Z^2}(t) = \mathbb{P}(Z^2 \le t) = \mathbb{P}(Z \le \sqrt{t}) = \int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

Dunque essendo $F_{Z^2}(t) \in C^1$ allora ho che

$$f_{z^2}(t) = F'_{Z^2}(t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{t}{2}}$$

Perciò per definizione ho che $Z^2 \sim \chi^2(1)$. □

Lemma 1.1.2. *Se $\{x_1, \cdots, X_n\}$ sono v.a. $\chi^2(k_i)$ indipendenti allora $X_1 + \cdots + X_n \sim \chi^2(k_1 + \cdots + k_n)$*

Lemma 1.1.3. *$\forall i = 1, \cdots, n$ le v.a. $X_i - \overline{X}$ e $\overline{X}$ sono indipendenti.*

Dimostrazione. Si vede che $(X_i - \overline{X}, \overline{X})$ è un v.a. gaussiano e per quanto dimostrato in precedenza ho che

$$\text{indipendenti} \iff \text{non sono correlate}$$

Quindi per dimostrare ciò basta verificare che la loro covarianza è nulla.
Per semplicità posso considerare che

$$X_i \sim N(0,1)$$
$$\overline{X} \sim N\left(0, \frac{1}{n}\right)$$

A questo punto osservo che

$$\overline{X} - X_1 = \frac{1}{n}\sum_{k=1}^{n} X_k - X_1 = \left(\frac{1}{n} - 1\right) X_1 + \underbrace{\frac{1}{n}\sum_{k=2}^{n} X_k}_{=:Z_2 \sim N\left(0, \frac{1}{n^2}(n-1)\right)}$$

Dunque ho che

$$Cov(X_i - \overline{X}, \overline{X}) = \mathbb{E}[(X_i - \overline{X})\overline{X}] - \mathbb{E}[X_i - \overline{X}]\mathbb{E}[\overline{X}]$$

Ma per costruzione risulta che $\mathbb{E}[X_i - \overline{X}]\mathbb{E}[\overline{X}] = 0$ inoltre ho per definizione che

$$\begin{aligned}
\mathbb{E}[(\overline{X} - X_1)\overline{X}] &= \mathbb{E}\left[\left(\frac{1-n}{n}X_1 + Z_2\right)\left(\frac{1}{n}X_1 + Z_2\right)\right] \\
&= \frac{1-n}{n^2}\mathbb{E}\left[X_1^2\right] + \left(\frac{1}{n} + \frac{1-n}{n}\right)\mathbb{E}[X_1 Z_2] + \mathbb{E}\left[Z_2^2\right] \\
&= \frac{1-n}{n^2} + \frac{1}{n^2}(n-1) = 0
\end{aligned}$$

Dunque le due variabili sono scorrelate e quindi indipendenti. □

Esercizio. *Sia $(X_1, \cdots, X_n)$ un campione casuale gaussiano. $\overline{X}_n$ e S_n^2 la media e la varianza campionaria. Sia X_{n+1} un'ulteriore osservazione. Allora*

$$\overline{X}_{n+1} = \frac{n\overline{X}_n + X_{n+1}}{n+1}$$

E che

$$nS_{n+1}^2 = (n-1)S_n^2 + \frac{n}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2$$

Teorema 1.1.1. *Se $X_1, \cdots, X_n$ sono v.a. $N(\mu, \sigma^2)$ indipendenti allora valgono i seguenti punti:*

1. ***Media campionaria***
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$
2. *$\overline{X}$ e S^2 sono v.a. indipendenti.*
3. ***Varianza campionaria***
$$S^2\frac{n-1}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - X}{\sigma}\right)^2 \sim \chi^2(n-1)$$

Dimostrazione. Dimostro procedendo con ordine.

Punto 1
Per definizione ho che

$$M_{\overline{X}}(t) = M_{\frac{1}{n}\sum_i X_i}(t) = \prod_{i=1}^{n} M_{X_i}\left(\frac{1}{n}t\right) = \prod_{i=1}^{n} \exp\left(\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n}\right)$$

Dunque per definizione ho che $\overline{X} \sim X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Punto 2 Dato che S^2 è una funzione di $(X_1 - \overline{X}), \cdots, (X_n - \overline{X})$ e ognuno dei termini è indipendente da $\overline{X}$, si ottiene che S^2 è indipendente da $\overline{X}$.

Punto 3
Il mio obiettivo è dimostrare che

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

Dunque

$n = 2$ In questo caso ho che

$$S_2^2 = \left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2 = 2\left(\frac{X_1 - X_2}{2}\right)^2 = \left(\frac{X_1 - X_2}{\sqrt{2}}\right)^2$$

Allora $X_1, X_2 \sim N(0,1)$ indipendenti allora

$$\frac{X_1 - X_2}{\sqrt{2}} \sim N(0,1)$$

$n + 1$ Dal lemma precedente ho che

$$nS_{n+1}^2 = (n-1)S_n^2 + \frac{n}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2$$

Ma per ipotesi induttiva ho che $(n-1)S_n^2 \sim \chi^2(n-1)$ inoltre la v.a. $\overline{X}_n - X_{n+1}$ è indipendente da S_n^2 e

$$\mathbb{E}\left[\left(\overline{X}_n - X_{n+1}\right)^2\right] = V\left(\overline{X}_n - X_{n+1}\right) = \frac{1}{n} + 1 = \frac{n+1}{n}$$

Perciò

$$\mathbb{E}\left[\left(\frac{\overline{X}_n - X_{n+1}}{\sqrt{\frac{n+1}{n}}}\right)^2\right] = 1$$

Dunque $\overline{X}_n - X_{n+1} \sim \chi^2(n)$.

□

Osservazione 1.1.2. *Se $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ allora posso stimare la distanza tra la media campionaria e la media*

$$\mathbb{P}\left(|\overline{X} - \mu| > \epsilon\right) = \mathbb{P}\left(\left|\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}\right| > \frac{\epsilon\sqrt{n}}{\sigma}\right) = 2\left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\right)$$

Se il valore di σ non è noto possiamo sostituire σ^2 con S^2 che è il suo stimatore

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Dunque bisogna chiedersi la distribuzione di

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}}$$

Possiamo banalmente riscrivere T come

$$T = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{(n-1)\frac{S^2}{\sigma^2}\frac{1}{n-1}}}$$

Per quanto dimostrato in precedenza so che $\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ e $(n-1)\frac{S^2}{\sigma^2}$ sono indipendenti quindi il tutto si riporta a studiare la distribuzione di una gaussiana standard Z in questo modo

$$\frac{Z}{\sqrt{\frac{V}{k}}} \qquad Z \sim N(0,1) \quad V \sim \chi^2(k)$$

Definizione 1.1.3. *Data una variabile aleatoria T definita come*

$$T := \frac{Z}{\sqrt{\frac{V}{k}}} \qquad Z \sim N(0,1) \quad V \sim \chi^2(k)$$

Con Z e V indipendenti. Si definisce distribuzione t di Student a k gradi di libertà, $T \sim t_k$, se ha come densità

$$f(t|k) = \frac{\Gamma\left(\frac{k-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sqrt{\pi k}} \frac{1}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}}$$

Proposizione 1.1.1. *Siano $\overline{X}_n$ e S_n^2 la media e la varianza campionaria per una successione $\{X_n\}_{n\in\mathbb{N}}$ di v.a. iid allora*

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$$

Dimostrazione. Per quanto visto in precedenza sappiamo che

$$Z_n = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0.1)$$

Mentre

$$V_n = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi(n-1)$$

Dunque valendo l'identità seguente:

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{S_n^2}{n}}} = \frac{Z_n}{\sqrt{\frac{V_n}{n-1}}}$$

Allora per definizione ho che $T \sim t(n-1)$. □

Osservazione 1.1.3. *Se $k = 1$ allora T ha una legge di Cauchy, cioè*

$$f(t) = \frac{1}{\pi}\frac{1}{1+t^2}, \quad t \in \mathbb{R}$$

Non ha media finita (quindi non ha momenti di nessun ordine).
In generale se $T \sim t_n$ ha momenti finiti fino all'ordine $n-1$, inoltre se:

- $n > 1$ *allora*

$$\mathbb{E}[T] = 0$$

- $n > 2$ *allora*

$$Var(T) = \frac{n}{n-1}$$

Proposizione 1.1.2. *Sia*

$$T = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{(n-1)\frac{S^2}{\sigma^2}\frac{1}{n-1}}}$$

Allora

- *Per il teorema del limite centrale vale che*

$$\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{D} N(0,1)$$

- *Per la legge dei grandi numeri vale che*

$$\frac{S^2}{\sigma^2} \xrightarrow{\mathbb{P}} 1$$

Teorema 1.1.2 (**Slutsky**). *Se* $X_n \xrightarrow{D} X$ *e* $Y_n \xrightarrow{\mathbb{P}} a$ *allora*

$$X_n Y_n \xrightarrow{D} aX$$

Corollario 1.1.1. *Applicato questo teorema al nostro caso, ottengo che*

$$T_n \xrightarrow{D} N(0,1)$$

1.2 Inferenza statistica

Vuol dire passare dal campione alla popolazione e usare le info sul campione per determinare il valore del parametro θ.

Definizione 1.2.1 (**Stimatore**). *Dato un campione statistico $(X_1, \cdots, X_n)$ lo stimatore è una funzione $F(X_1, \cdots, X_n)$ e pertanto una variabile aleatoria mentre la stima è definita dalla funzione funzione $F(x_1, \cdots, x_n)$ è un numero $\psi(\hat{\theta})$ come stima per $\psi(\theta)$.*

Definizione 1.2.2 (**Statistica**). *Dato un campione $X = (X_1, \cdots, X_n)$ e una funzione $T : \mathbb{R}^n \to \mathbb{R}$ allora definiamo $T(X)$ statistica ed è una v.a. la cui legge dipende da θ*

Definizione 1.2.3. *Data una statistica $T(X)$ la chiamerò stimatore del parametro $\psi(\theta)$ se date le osservazioni $x = (x_1, \cdots, x_n)$ definisco*

$$T(\underline{x}) = T(x_1, \cdots, x_n) = \hat{\psi}(\theta)$$

Stima di $\psi(\theta)$.

Esempio 1.2.1. *Riprendendo l'esempio delle monete ho che*

- $T_1 = T_1(X_1, \cdots, x_n) = X_1$ *questa rappresenta una statistica e uno stimatore se pongo*

  $$x_1 = \hat{p}$$

  (Stimo il tutto con l'esito della prima uscita)

- $T_2 = T_2(X_1, \cdots, x_n) = X_1 + \cdots + X_n$ *questa rappresenta una statistica e uno stimatore se pongo*

  $$\hat{p} = x_1 + \cdots + x_n$$

  (Stimo il tutto con l'esito di n lanci)

- $T_3 = T_3(X_1, \cdots, x_n) = \overline{X}$ *questa rappresenta una statistica e uno stimatore se pongo*

  $$\hat{p} = \frac{x_1 + \cdots + x_n}{n}$$

Esempio 1.2.2. *Consideriamo una moneta di parità p incognita. Il campione in questo caso rappresenta un'insieme di lanci $(X_1, \cdots, X_n)$ v.a. iid $B(p)$. Come stimare p?*

Definizione 1.2.4 (**Principio di sufficienza**). *Data una statistica $T(X)$ dico che è una statistica sufficiente per θ se la distribuzione di X dato $T(X)$ non dipende da θ.*

Osservazione 1.2.1. *Quello che significa è che se T è una statistica sufficiente allora tutta l'informazione relativa a θ contenuta nel campione è contenuta nel valore $t = T(x)$ oppure $t = T(x)$ contiene la stessa quantità di informazione su θ dell'intero campione $x = (x_1, \cdots, x_n)$.*

Teorema 1.2.1. *Se $f(x|\theta)$ è la legge della popolazione e se $q(t|\theta)$ è la legge della statistica $T(X)$ allora $T(X)$ è una statistica sufficiente se $\forall x$ nello spazio campionario ho che $\theta \mapsto \frac{f(x|\theta)}{g(x|\theta)}$ è una costante.*

Esempio 1.2.3. *$T_2 = T_2(X_1, \cdots, X_n) = X_1 + \cdots + X_n$ è una statistica sufficiente per la parità p della moneta. Si vede che*

$$T_2 \sim B(n, p)$$

Dunque per definizione ho che

$$f(x|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$q(t|p) = \binom{n}{t} p^t (1-p)^{n-1}$$

Perciò posto $t = \sum_{i=1}^{n} x_i$ *ho che*

$$\frac{\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-1}} = \frac{p^{\sum x_i}(1-p)^{\sum(1-x_i)}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

Non dipende dal parametro p *e quindi per definizione è una statistica sufficiente.*

Esempio 1.2.4. *Popolazione tale per cui* $f(x|\theta) \sim Unif\{1, \cdots, \theta\}, \theta \in \mathbb{N}, \theta \geq 1$ *allora una statistica sufficiente per stimare* θ *è*

$$T(X) = \max_{1 \leq i \leq n} X_i$$

In quanto per definizione ho che

$$f(\underline{x}|\theta) = \prod_{i=1}^{n} \left(\frac{X_i}{\theta} \right) = \frac{\prod_{i=1}^{n} X_i}{\theta^n}$$

E la legge della statistica è data da

$$q(\underline{x}|\theta) = \mathbb{P}^{\theta} \left(\max_{1 \leq i \leq n} X_i \leq t \right) = \prod_{i=1}^{n} \mathbb{P}^{\theta}(X_i \leq t) = \prod_{i=1}^{n} \left(\frac{X_i}{\theta} \right) = \left(\frac{X_i}{\theta} \right)^n$$

Dunque

$$q(T(\underline{X})|\theta) = \left(\frac{\max_{1 \leq i \leq n} X_i}{\theta} \right)^n$$

Allora osservo che

$$\theta \mapsto \frac{f(\underline{X}|\theta)}{q(T(\underline{X})|\theta)} = \frac{\prod_{i=1}^{n} X_i}{\left(\max_{1 \leq i \leq n} X_i\right)^n}$$

Dunque dato che non dipende da θ *allora per definizione è una statistica sufficiente.*

1.3 Stimatori

Siamo interessati a determinare il valore corretto di θ oppure $\psi(\theta)$.
$T(X)$ è uno stimatore per $\psi(\theta)$ se decidiamo che $T(x) = \hat{\psi}(\theta)$

Definizione 1.3.1. *$T(X)$ è uno **stimatore corretto** (non distorto) se vale che*

$$\mathbb{E}^\theta[T] = \psi(\theta), \qquad \forall\theta$$

Esempio 1.3.1. *Posto $\phi(\theta) = \theta$ allora vedo che*

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(X_i) = \theta$$

Dunque la media campionaria è un buono stimatore, come lo è anche $T(X) = X_1$ è uno stimatore corretto per μ ma non va bene per studiare la media.

Teorema 1.3.1 (**Cramer-Rao**). *(Sotto ipotesi opportune) Vale che la varianza di uno stimatore non può essere inferiore a un certo livello*

$$V^\theta(T) \geq V(\theta) = \frac{1}{n}\frac{(\psi'(\theta))}{\mathbb{E}^\theta\left[\frac{d}{d\theta}\log(f(X_1(\theta)))^2\right]}$$

Definizione 1.3.2. *Uno stimatore per cui $V^\theta(T) = V(\theta)$ si dice efficiente.*

Definizione 1.3.3. *T è uno stimatore asintoticamente corretto se*

$$T = T(X) = T_n(X)$$

Questa rappresenta una statistica su un campione di popolosità n.
Deve valere che

$$\mathbb{E}^\theta\left[T_n\right] \xrightarrow{n\to+\infty} \psi(\theta)$$

Definizione 1.3.4. *T è uno stimatore consistente se*

$$T_n \xrightarrow{\mathbb{P}^\theta} \psi(\theta) \qquad \forall\theta$$

Esempio 1.3.2.

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

Questo è dipendente da n. Dunque per la legge dei grandi numeri so che

$$\mathbb{P}^\mu(|\overline{X}_n - \mu| > \epsilon) \leq \frac{1}{n}\frac{\sigma^2}{\epsilon^2} \xrightarrow{n\to+\infty} 0$$

Definizione 1.3.5. *T è uno stimatore asintoticamente normale se $\forall\theta, \exists\sigma = \sigma(\theta) > 0$ tale per cui*

$$\frac{T - \psi(\theta)}{\frac{\sigma(\theta)}{\sqrt{n}}} \xrightarrow{D} N(0,1)$$

Uno stimatore asintoticamente normale è consistente.

Esistono 2 metodi per determinare gli stimatori:

1. Metodo dei momenti.
2. Metodo di massima somiglianza

1.3.1 Metodo dei momenti

Sia $X = (X_1, \cdots, X_n)$ iid, $\theta = (\theta_1, \cdots, \theta_k)$ e $f(x|\theta)$ ho che

$$\mathbb{E}^{\theta}[X^j] = m_j(\theta_1, \cdots, \theta_k) \qquad \forall j = 1, \cdots, k$$

D'altra parte chiamerò

$$\overline{m}_j = \frac{1}{n}\sum_{l=1}^{n} X_l^j \qquad \forall j = 1, \cdots, k$$

$\forall j, \overline{m}_j$ è una statistica.

$$\begin{cases} \overline{m}_j = m_j(\theta_1, \cdots, \theta_k) \\ j = 1, \cdots, k \end{cases}$$

Risolvo questo sistema nelle incognite $\theta_1, \cdots, \theta_k$. La soluzione $\hat{\theta} = \left(\hat{\theta}_1, \cdots, \hat{\theta}_k\right)$ è la stima con il metodo dei momenti del parametro θ.

Esempio 1.3.3. *Dato $X \sim N(\mu, \sigma^2)$ allora i primi due parametri sono $\theta_1 = \mu$ e $\theta_2 = \sigma^2$ con*

$$\begin{aligned} m_1 &= \mu \\ m_2 &= \sigma^2 + \mu^2 \end{aligned}$$

Perciò ho che

$$\begin{cases} \hat{\theta}_1 = \dfrac{1}{n}\displaystyle\sum_{l=1}^{n} x_l \xrightarrow{\text{Stiamtore corrispondente}} & \overline{X} = \dfrac{1}{n}\displaystyle\sum_{l=1}^{n} X_l \\ \hat{\theta}_2 = m_2 - \hat{\theta}_1^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2 \to & T_2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{cases}$$

Lemma 1.3.1. *T_2 è uno stimatore asintoticamente corretto*

Dimostrazione.

$$T_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \qquad \mathbb{E}[T_2] = \frac{1}{n} n \mathbb{E}\left[X_i^2\right] - \frac{1}{n^2}\mathbb{E}\left[\left(\sum_k X_k\right)^2\right]$$

Dunque abbiamo che

$$\begin{aligned} \mathbb{E}[T_2] &= (\sigma^2 + \mu^2) - \frac{1}{n^2}\mathbb{E}\left[\sum_k^n X_k^2 + \sum_{j \neq k} X_k X_j\right] \\ &= \sigma^2 + \mu^2 - \frac{1}{n^2}(n(\sigma^2 + \mu^2) + n(n-1)\mu^2) \\ &= \sigma^2\left(1 - \frac{1}{n}\right) + \mu^2\left(1 - \frac{1}{n} - \frac{n-1}{n}\right) \\ &= \frac{n-1}{n}\sigma^2 \xrightarrow{n \to +\infty} \sigma^2 \end{aligned}$$

□

Lemma 1.3.2. *Lo stimatore*

$$S^2 = \frac{n}{n-1} T_2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2$$

È uno stimatore corretto per la varianza

1.3.2 Metodo di massima verosimiglianza

Con questo metodo si stima il parametro θ attraverso il valore che rende massima la funzione di verosimiglianza ovvero si trova il valore del parametro per cui è massima la probabilità di ottenere il campione osservato.

Definizione 1.3.6 (Funzione di verosimiglianza). *Dato* $X \sim \prod_{i=1}^{n} f(x_i|\theta)$ *dunque chiamiamo la funzione di verosimiglianza*

$$L(\cdot|x)\colon \mathbb{R}^n \to \mathbb{R}$$

$$\theta \mapsto L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

Definizione 1.3.7. *Fissato* $\underline{x}$ *chiamo stima di massima verosimiglianza (MLE) il valore*

$$\hat{\theta} = MLE(\theta)$$

Che se esiste è l'unico punto di massimo per la funzione $L(\theta|x)$.

Ci sono 2 problemi fondamentalmente con questa situazione:

1. Trovare il massimo e dimostrare che è unico.
2. Verificare la sensibilità rispetto ai dato.

Se $L(\theta|x)$ è derivabile in θ allora posso cercare il punto di massimo tra i punti che annullano la derivata, dunque

$$\frac{d}{d\theta} L(\theta|x) = \frac{d}{d\theta} \prod_{k=1}^{n} f(x_k|\theta) = 0$$

Se θ è ristretta in un dominio, devo anche considerare anche i punti di bordo.
Alla fine devo trovare quale è il massimo globale.

Si può usare la funzione di log-verosimiglianza indicata con

$$l(\theta|x) = \log(F(\theta|x)) = \sum_{k=1}^{n} \log(f(x_k|\theta))$$

Siccome log è una funzione monotona strettamente crescente allora θ risolve

$$\frac{d}{d\theta} L(\theta|x) = 0 \iff \frac{d}{d\theta} l(\theta|x) = 0$$

Equazione di log-verosimiglianza allora

$$\frac{d}{d\theta} l(\theta|x) = \sum_{k=1}^{n} \frac{d}{d\theta} \log(f(x_k|\theta))$$

Esempio 1.3.4. *Posto* $f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \lambda > 0$ *dunque*

$$\log\left(e^{-\lambda}\frac{\lambda^k}{k!}\right) = -\lambda + x\log\lambda - \log x!$$

Quindi è ovvio ch e

$$\frac{d}{d\lambda}\log f(x|\lambda) = -1 + x\frac{1}{\lambda} \to 0 = \sum_{k=1}^{n}\left(-1 + x_k\frac{1}{\lambda}\right)$$

Dunque ho che

$$n = \frac{1}{\lambda}\sum x_k$$

Quindi ho che

$$\hat{\lambda} = \frac{1}{n}\sum x_k = \overline{X}$$

Questo non sorprende in quanto il parametro di Poissont è uguale alla sua media.

Esempio 1.3.5. *Se $f(x|p) \sim Be(p)$ allora*

$$f(x|p) = p^x(1-P)^{1-x}$$

Per $x = 0, 1$ dunque ho che $\sigma = \sqrt{p(1-p)}$ e $MLE(\sigma) = \sqrt{\hat{p}(1-\hat{p})}$

Teorema 1.3.2. *Se $\hat{\theta} = MLE(\theta)$ allora per ogni funzione ψ vale che*

$$\hat{\psi}(\theta) = MLE(\psi(\theta)) = \psi\left(\hat{\theta}\right) = \psi(MLE(\theta))$$

Teorema 1.3.3. *Se $\theta = \mathbb{E}^{\theta}[X]$ allora $MLE(\theta) = \overline{X}$ e lo stimatore $\overline{X}$ è corretto, efficiente, consistente e asintoticamente normale, quindi ha tutte le buone proprietà che posso volere da uno stimatore.*

1.4 Test di ipotesi

Definizione 1.4.1. *Una ipotesi è un'affermazione su un parametro della distribuzione della v.a. di interesse nella popolazione.*

Definizione 1.4.2 (Test di ipotesi). *Le due ipotesi complementari in un test sono chiamate:*

1. *H_0: ipotesi nulla*
2. *H_1: ipotesi alternativa*

In generale $H_0 : \theta \in \Theta_0$ $H_1 : \theta \in \Theta_0^c$ dove Θ è lo spazio dove vive θ e $\Theta_0 \subset \Theta$.

Esempio 1.4.1. *Test su un medicinale e diremo come ipotesi nulla $H_0 : \theta = 0$ (la medicina non ha effetto) e $H_1 : \theta \neq 0$ (la medicina ha un qualche effetto).*

Esempio 1.4.2. *$H_0 = \theta \geq \theta_0$ (il numero d pezzi difettosi è alta) mentre $H_1 : \theta < \theta_0$ il test è costruito in modo da provare a dimostrare che il fornitore è efficiente.*

Per risolvere un test di ipotesi:

- Per quali valori del campione scelgo H_0?
- Per quali valori scelgo H_1?

Definisco regione di rifiuto $R := \{x|F(x) \leq c\}$ se il campione è tale per cui $x \in R$ allora rifiuto H_0 e scelgo H_1.

Esempio 1.4.3. *Supponiamo di avere una monete truccata con una faccia avente il doppio delle probabilità dell'altra di uscire ma non so quale sia tale faccia, quindi $H_0 : p = \frac{1}{3}$ oppure $H_1 : p = \frac{2}{3}$. Raccogliamo un campione lanciando n volte la moneta. In questo caso mi aspetto che la faccia truccata sia quella che esca più spesso quindi*

$$R := \left\{ x = (x_1, \cdots, x_n) : \sum_{i=1}^{n} x_i \geq 6 \right\}$$

*Oltre la metà dei lanci è testa quindi rifiuto H_0. Per costruire R uso **LRT**: test del rapporto di verosimiglianza*

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

Se H_0 è vera allora la vero-somiglianza dei dati $\sup_{\theta \in \Theta_0} L(\theta|x)$ la verosimiglianza dei dati su Θ è $\sup_{\theta \in \Theta} L(\theta|x) = L(\hat{\theta}|x)$ con $\hat{\theta} = MLE(\theta)$ dunque chiamerò

$$W(\underline{X}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\underline{X})}{L(\hat{\theta}|\underline{X})}$$

Se questo rapporto è basso i dati sono contro la mia ipotesi, viceversa se il rapporto è grande.

Definizione 1.4.3. *La regione di rifiuto dello spazio campionario per cui decidiamo di rifiutare l'ipotesi nulla nel caso i dati entrino in questa regione ed è indicata come*

$$R = \{x | W(X) \leq c\}$$

Dove $c \in [0, 1]$. *Il suo complementare, invece, è detta regione di accettazione A.*

Osservazione 1.4.1. *Se il rapporto di* $W(X)$ *è basso significa che i dati sono contro la mia ipotesi, viceversa se il rapporto è grande.*

Osservazione 1.4.2. *Se* $H_0 : \theta = \theta_0$ $\Theta_0 := \{\theta_0\}$ *allora*

$$W(X) = \frac{L(\theta_0|x)}{L(\hat{\theta}|x)}$$

Ripredendo l'esempio della moneta ho che $\Theta := \left\{\frac{1}{3}, \frac{2}{3}\right\}$ e $\Theta_0 := \left\{\frac{1}{3}\right\}$ allora

$$L(\theta|x) = \binom{n}{s} p^s (1-p)^{n-s} \qquad s = \sum_{i=1}^{n} x_i$$

Perciò ho che

$$W(X) = \frac{L\left(\frac{1}{3}|x\right)}{\max\left\{L\left(\frac{1}{3}|x\right), L\left(\frac{2}{3}|x\right)\right\}}$$

Ma

$$\max\left\{L\left(\frac{1}{3}|x\right), L\left(\frac{2}{3}|x\right)\right\} = \begin{cases} L\left(\frac{1}{3}|x\right) & \frac{s}{n} < \frac{1}{2} \\ L\left(\frac{2}{3}|x\right) & \frac{s}{n} > \frac{1}{2} \end{cases}$$

Dunque

$$W = \begin{cases} 1 & s < \frac{n}{2} \\ \frac{L\left(\frac{1}{3}|x\right)}{L\left(\frac{2}{3}|x\right)} = 2^{n-2s} & s > \frac{n}{2} \end{cases}$$

Dunque se

$$R := \{x | W(x) \leq c\} = \left\{x | s = \sum x_i \text{ soddisfa } s \geq \frac{n}{2} - \frac{1}{2}\log_2(c)\right\}$$

Allora se

- $c \geq \frac{1}{2} \Longrightarrow R = \{x | s \geq 6\}$
- $\frac{1}{4} \leq c \leq \frac{1}{2} \Longrightarrow R = \{x | s \geq 7\}$

È sempre possibile onestamente rifiutare H_0 anche se è vera o accettare H_0 anche se è falsa.
Nel caso della moneta se ho 11 teste scelgo $H_1 : p = \frac{2}{3}$ ma c'è una probabilità

$$P = \left(\frac{1}{3}\right)^{11} = 5,6 \cdot 1 =^{-6}$$

Definizione 1.4.4. *Diremo*

- ***Errore del I tipo**: H_0 è vero e noi la rifiutiamo.*
- ***Errore del II tipo**: H_0 è falsa e noi la accettiamo.*

C'è da osservare che se $\theta \in \Theta_0$ allora $\mathbb{P}^\theta(X \in R)$ significa la probabilità di rigiutare H_0 anche se poi θ è effettivamente in quelal regione e quindi H_0 è vera e quindi questo rappresenta un errore del I tipo.
Se $\theta \in \Theta^c$ allora $\mathbb{P}^\theta(X \in R)$ questo rappresenta un errore del II tipo.

Definizione 1.4.5. *Definita la funzione potenza del test come*

$$\beta(\theta) = \mathbb{P}^{\theta}(X \in R)$$

Allora diremo che un test ha livello di significatività $1-\alpha$ *se*

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq 1 - \alpha$$

Osservazione 1.4.3. *Questo valore mi rappresenta la probabilità di fare un errore nello scartare l'ipotesi nulla e prendere l'ipotesi alternativa.*

Nel caso della moneta ho che

$$\beta(\theta) = \mathbb{P}^{\theta}\left(S_n \geq \frac{n}{2} - \frac{1}{2}\log_2 c\right)$$

Definizione 1.4.6. *Sia* $0 \leq \alpha \leq 1$ *allora diremo che un test di ipotesi ha livello di significatività* α *se*

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$$

1.4.1 Test di ipotesi

Supponiamo che la popolazione sia normale di varianza σ^2 nota e quindi posso porre $\sigma^2 = 1$ quindi

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

Quindi prendiamo con ipotesi nulla che $H_0 : \mu = \mu_0$ e come ipotesi alternativa $H_1 : \mu \neq \mu_0$ inoltre poniamo $\Theta = \mathbb{R}$ e $\Theta_0 = \{\mu_0\}$ dunque

$$W(x) = \frac{L(\theta_0|x)}{L(\hat{\theta}|x)}$$

Con $\hat{\mu} = MLE(\mu) = \overline{x}$ perciò

$$R := \left\{ x \,\middle|\, \frac{(2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum (x_1 - \mu_0)^2\right)}{(2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum (x_1 - \overline{x})^2\right)} \leq c \right\}$$

Dunque mediante passaggi algebrici ho che

$$R := \left\{ x \middle| \exp\left(-\frac{n}{2}(\mu_0 - \overline{x})^2\right) \leq c \right\}$$

Dunque

$$R := \left\{ x \middle| (\mu_0 - \overline{x})^2 \geq -\frac{2}{n}\log(c) \right\}$$

Perciò la ragione di rifiuto può essere riscritta come

$$R := \left\{ x \,\middle|\, |\overline{x} - \mu_0| \geq \sqrt{\frac{2}{n}\log\left(\frac{1}{c}\right)} \right\}$$

Adesso devo determinare la potenza del test, dunque

$$\beta(\mu) = \mathbb{P}^{\mu}(X \in R) = \mathbb{P}^{\mu}(|X - \mu_0| \geq \epsilon)$$

La probabilità di errore del I tipo è

$$\beta(\mu_0) = \mathbb{P}^{\mu}(|X - \mu_0| \geq \epsilon)$$

$$X \sim N\left(\mu_0, \frac{1}{n}\right)$$

Perciò

$$\beta(\mu_0) = \mathbb{P}^{\mu_0}\left(\left|\frac{X-\mu_0}{\sqrt{\frac{1}{n}}}\right| \geq \frac{\epsilon}{\sqrt{\frac{1}{n}}}\right) = \mathbb{P}(|z| \geq \epsilon\sqrt{n})$$

Dunque davo calcolare l'area della gaussiana delle due code che essendo simmetriche posso scrivere in maniera del tutto equivalente come

$$2(1-\Phi(\epsilon\sqrt{n}))$$

Il livello di significatività del test α è per definizione

$$\beta(\mu_0) = 1 - \frac{\alpha}{2} = \Phi(\epsilon\sqrt{n})$$

Dunque posso dire che il test di ipotesi per

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

Popolazione normale di varianza nota con $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$ con livello di significatività α ha come regione di rifiuto

$$R := \{x | |\overline{x} - \mu_0| \geq \frac{1}{\sqrt{n}} \Phi\left(1 - \frac{\alpha}{2}\right)\}$$

Osservazione 1.4.4. *Popolazione normale di varianza nota σ^2. Supponiamo che $H_0 : \mu \leq \mu_0$ e $H_1 : \mu > \mu_0$ inoltre per definizione ho che*

$$W(X) = \frac{\sup_{\Theta_0} L(\mu|\underline{x})}{L(\hat{\mu}|\underline{u})} \qquad \sup_{\Theta_0} L(\mu|\underline{x}) = \begin{cases} L(\hat{\mu}|\underline{x}) & \hat{\mu} = \overline{x} \in \Theta_0 \\ W(\underline{x}) = \exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)\right)^2 & \\ L(\mu|\underline{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu\right)^2 & \hat{\mu} = \overline{x} \notin \Theta_0 \iff \overline{x} > \mu_0 \end{cases}$$

Dunque ho che

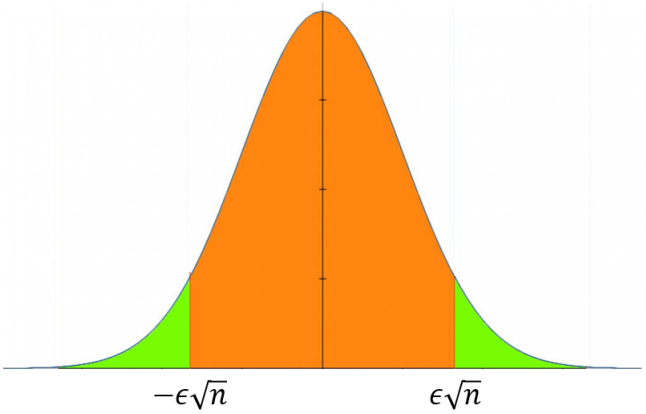
$$W(\underline{x}) = \begin{cases} 1 & \overline{x} \leq \mu_0 \\ \exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)\right)^2 & \overline{x} > \mu_0 \end{cases}$$

Adesso dobbiamo trovare la regione di rifiuto che per definizione è

$$R := \{W(x) \leq c\} = \left\{\overline{x} > \mu_0 \,\middle|\, |\overline{x} - \mu_0| \geq \sqrt{\frac{2}{n}\log\left(\frac{1}{c}\right)}\right\}$$

Dunque pongo $\epsilon = \sqrt{\frac{2}{n}\log\left(\frac{1}{c}\right)}$

$$\mathbb{P}^{\mu}(X \in R) = \mathbb{P}^{\mu}(X \geq \mu_0 + \epsilon)$$



Per studiare la significatività del test devo studiare $\sup_{\mu\leq\mu_0}\mathbb{P}^{\mu}(X>\mu_0+\epsilon)$ *quindi*

$$\alpha=\beta(\mu_0)=\mathbb{P}^{\mu_0}(|\underline{X}-\mu_0|>\epsilon)=\mathbb{P}^{\mu_0}\left(Z>\epsilon\sqrt{n}\right)$$

Perciò fissato il livello di significatività al test α *ho che*

$$\mathbb{P}^{\mu_0}\left(Z>\epsilon\sqrt{n}\right)=2(1-\Phi(\epsilon\sqrt{n}))=\alpha$$

Dunque è ovvio affermare che

$$\epsilon=\frac{1}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}$$

Quindi ricapitolando il tutto, posso scrivere che

$$R=\left\{X\,\middle|\,|\overline{X}-\mu_0|\geq\frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right\}$$

Più alto è il campione e meno mi devo spostare.

Esempio 1.4.4. *Popolazione normale di varianza incognita.*

$$L(\mu,\sigma^2|\underline{x})=\prod_{i=1}^{n}f(x_i|\mu,\sigma^2)\qquad\left(\frac{\partial L}{\partial\mu},\frac{\partial L}{\partial\sigma^2}\right)=(0,0)$$

Ricordando che gli stimatori sono $\hat{\mu}=MLE(\mu)=\overline{x}$ *e* $\hat{\sigma}^2=MLE(\sigma^2)=s^2$*. Supponiamo che* $H_0:\mu\leq\mu_0$ *e* $H_1:\mu>\mu_0$ *il test di ipotesi solo sulla media dunque la regione di rifiuto è*

$$R=\left\{\overline{x}>\mu_0+\sqrt{\frac{s^2}{n}}t_{1-\alpha}^{(n-1)}\right\}$$

Esempio 1.4.5. *Data* $f(x|\theta)=e^{-(x-\theta)}\mathbb{1}_{(x\geq\theta)}$ *dunque supponiamo* $H_0:\theta\leq\theta_0$ *e* $H_1:\theta>\theta_0$ *allora*

$$L(\theta|\underline{x})=\begin{cases}\exp\left(-\sum x_i+n\theta\right) & x_{\min}\geq\theta\\ 0 & x_{\min}>\theta\end{cases}$$

Tra l'altro è facile vedere che $T=X_{\min}=\min\{X_1,\cdots,X_n\}$ *è una statistica sufficiente per* θ*, dunque*

$$\mathbb{P}^{\theta}(T>t)=e^{-n(t-\theta)}\mathbb{1}_{t\geq\theta}$$

A questo punto noto che $\theta\mapsto L(\theta|\underline{x})$ *è crescente se* $\theta\leq x_{\min}$ *e* $=0$ *se* $\theta>x_{\min}$ *quindi*

$$W(X)=\frac{\sup_{\theta\leq\theta_0}\exp\left(-\sum x_i+n\theta\right)}{\sup_{\theta}\left(-\sum x_i+n\theta\right)}=\begin{cases}1 & x_{\min}\leq\theta_0\\ e^{-n(x_{\min}-\theta_0)} & x_{\min}>\theta_0\end{cases}$$

Perciò la regione di rifiuto

$$R = \{x|w(x) \leq c\} = \left\{c|e^{-n(x_{\min}-\theta_0)} \leq c\right\} = \left\{x|x_{\min} - \theta_0 \geq \frac{1}{n}\log\left(\frac{1}{c}\right)\right\}$$

Inoltre

$$\beta(\theta) = \mathbb{P}^\theta(X \in R) = \mathbb{P}^\theta\left(T > \theta_0 + \frac{1}{n}\log\left(\frac{1}{c}\right)\right) = e^{-n\left(\theta_0-\theta+\frac{1}{n}\log\frac{1}{c}\right)} = c\,e^{-n(\theta_0-\theta)}$$

Il livello di significatività del test $\alpha = \sup_{\theta\leq\theta_0}\beta(\theta) = \beta(\theta_0) = c$

$$R = \{x|w(x) \leq \alpha\} = \left\{x|x_{\min} \leq \theta_0 - \frac{1}{n}\log\alpha\right\}$$

Esempio 1.4.6. *Data una popolazione esponenziale* $f(x|\theta) = \theta e^{-\theta x}$ *media* $\frac{1}{\theta}$ *e varianza* $\frac{1}{\theta^2}$ *e supponiamo che* $H_0 : \theta \leq \theta_0$ *e* $H_1 : \theta > \theta_0$ *inoltre*

$$L(\theta|\underline{x}) = \theta^n e^{-\theta S_n} \qquad S_n = \sum_{i=1}^n x_i$$

$$L(\theta|\underline{x}) = n\log\theta - \theta S_n$$

Il punto di massimo di questa funzione sarà indicato da

$$\frac{\partial}{\partial\theta}l(\theta|\underline{x}) = \frac{n}{\theta} - S_n = 0 \Longleftrightarrow MLE(\theta) = \hat{\theta} = \frac{n}{S_n} = \frac{1}{\overline{x}}$$

Perciò

$$W(\underline{x}) = \frac{\sup_{\theta\leq\theta_0} L(\theta|\underline{x})}{L(\hat{\theta}|\underline{x})} = \begin{cases} 1 & \hat{\theta} < \theta_0 \\ \dfrac{L(\theta_0|\underline{x})}{L(\hat{\theta}|\underline{x})} & \hat{\theta} > \theta_0 \end{cases}$$

Dunque

$$W(\underline{x}) = \begin{cases} 1 & \dfrac{S_n}{n/\theta_0} \geq 1 \\ \left[\dfrac{S_n}{n/\theta_0}\exp\left(1 - \dfrac{S_n}{n/\theta_0}\right)\right]^n & \dfrac{S_n}{n/\theta_0} < 1 \end{cases}$$

La regione di rifiuto è dunque per definizione

$$R = \{W(\underline{x}) \leq c\} = \left\{\left[\frac{S_n}{n/\theta_0}\exp\left(1 - \frac{S_n}{n/\theta_0}\right)\right]^n \leq c\right\}$$

A questo punto rimane da calcolare

$$\mathbb{P}^\theta(X \in R) = \mathbb{P}^\theta(W(\underline{x}) \leq c) = \mathbb{P}^\theta\left(\left[\frac{S_n}{n/\theta_0}\exp\left(1 - \frac{S_n}{n/\theta_0}\right)\right]^n \leq c\right)$$

I metodo

Per definizione posso calcolare che

$$\beta(\theta) = \mathbb{P}^\theta\left(\frac{S_n}{n/\theta_0} \leq \Phi_n^{-1}(x)\right)$$

Ricordando che $S_n = \sum_{i=1}^n x_i \sim \Gamma(n,\theta)$ *perciò* $f_{S_n}(x|\theta) = e^{-\theta x}\frac{\theta^n}{\Gamma(n-1)}x^{n-1}$ *allora la probabilità precedente è equivalente a calcolare*

$$F_{S_n}\left(\Phi^{-1}(c)|\theta\right)$$

Questa funzione non può essere calcolata analiticamente.

II metodo

Dato che sappiamo che abbiamo una funzione monotona lalora

$$R = \left\{ x | \frac{S_n}{n/\theta_0} \leq \epsilon \right\} \qquad S_n = \sum_{o=1}^{n} x_i$$

Se n è sufficientemente grande allora per TLC ho che $S_n^\star \sim N(0,1)$ allora

$$R = \left\{ x | S_n^\star \leq \frac{\epsilon \frac{n}{\theta_0} - \mathbb{E}[S_n]}{\sqrt{V(S_n)}} \right\}$$

Ma per come ho definito S_n ho che $\mathbb{E}[S_n] = \frac{n}{\theta}$ e $V(S_n) = \frac{n}{\theta^2}$ allora

$$R = \left\{ x | S_n^\star \leq \frac{\epsilon \frac{n}{\theta_0} - \frac{n}{\theta}}{\sqrt{\frac{n}{\theta^2}}} \right\}$$

A questo punto mi interessa

$$\beta(\theta) = \mathbb{P}\left(S_n^\star \leq \frac{\epsilon \frac{n}{\theta_0} - \frac{n}{\theta}}{\sqrt{\frac{n}{\theta^2}}} \right) = \Phi()$$

Dopo di che bisogna osservare che

$$\beta(\theta) \leq \beta(\theta_0) \quad \theta \leq \theta_0$$

Per calcolare la significatività del test: $\beta(\theta_0)$ dunque

$$\Phi\left((\epsilon - 1)\sqrt{n}\right) = \alpha$$

Questo è equivalente a dire che $\Phi((1-\epsilon)\sqrt{n}) = 1 - \alpha$ e quindi $(1-\epsilon)\sqrt{n} = \phi_{1-\alpha}$ e finalmente si ottiene che

$$R = \left\{ x | \frac{S_n}{n/\theta_0} \leq 1 - \frac{1}{\sqrt{n}} \phi_{1-\alpha} \right\}$$

E per calcolare questo valore servono le tavole della gaussiana.

III metodo

Per $\frac{S_n}{n/\theta_0} \leq 1$ sia

$$W(\underline{x}) = \frac{L(\theta_0 | \underline{x})}{L(\hat{\theta} | \underline{x})}$$

Dunque la regione di rifiuto è identifiata da

$$R = \{ -2\log(W(\underline{x})) \geq \underbrace{-2\log(c)}_{\epsilon} \}$$

Per definizione ho che

$$\log W(\underline{x}) = l(\theta_0 | \underline{x}) - l(\hat{\theta} | \underline{x})$$

Perciò se chiamo

$$\Lambda(\underline{x}) = -2\log W(\underline{x}) = 2\left[l(\hat{\theta} | \underline{x}) - l(\theta_0 | \underline{x}) \right]$$

Facendo lo sviluppo di Taylor ottengo che

$$l(\theta_0 | \underline{x}) = l(\hat{\theta} | \underline{x}) + \frac{\partial}{\partial \theta} l(\hat{\theta} | \underline{x})(\theta_0 - \hat{\theta}) + \frac{1}{2} \frac{\partial^2}{\partial \theta^2} l(\hat{\theta} | \underline{x})(\theta_0 - \hat{\theta})^2$$

Dato che $\hat{\theta}$ è un punto di massimo allora posso approssimare

$$\Lambda(\underline{x}) = -\frac{\partial^2}{\partial \theta^2} l(\hat{\theta} | \underline{x})(\theta_0 - \hat{\theta})^2$$

Dunque

$$-l''(\hat{\theta}|\underline{x}) = \frac{n}{\hat{\theta}^2} = \frac{S_n^2}{n}$$

Perciò

$$\Lambda(\underline{x}) = 2\frac{S_n^2}{n}\left(\frac{n}{S_n} - \theta_0\right)^2 = \left(\frac{S_n - n/\theta_0}{\sqrt{n/\theta_0^2}}\right)^2 \xrightarrow{D} \chi^2(1)$$

Dunque ho che

$$\mathbb{P}(\Lambda(\underline{x} \geq \epsilon) = \alpha \iff \mathbb{P}(\Lambda(\underline{x} \leq \epsilon)) = 1 - \alpha \iff \epsilon = \chi^2_{1-\alpha}(1)$$

Allora

$$R = \left\{ \Lambda(\underline{x}) = 2\frac{S_n^2}{n}\left(\frac{n}{S_n} - \theta_0\right)^2 = \left(\frac{S_n - n/\theta_0}{\sqrt{n/\theta_0^2}}\right)^2 \geq \chi^2_{1-\alpha}(1) \right\}$$

Rappresenta la regione di rifiuto al livello di significatività di α.

Osservazione 1.4.5. *Posto $\theta_0 = 1$ allora possiamo osservare che*

	$n = 10$	$n = 50$
I metodo	$R = \{S_n \leq 6.22\}$	$R = \{S_n \leq 41.18\}$
II metodo	$R = \{S_n \leq 5.95\}$	$R = \{S_n \leq 40.95\}$
III metodo	$R = \{S_n \leq 6.65\}$	$R = \{S_n \leq 40.24\}$

1.5 P-value

Definizione 1.5.1 (**p-value**). *Definisco p-value di un test una statistica $p(X)|0 \leq p(\underline{x}) \leq 1$ e*

$$\mathbb{P}^{\theta}(P(\underline{X}) \leq \alpha) = \alpha \qquad \theta \in \Theta_0 \quad 0 < \alpha \leq 1$$

Osservazione 1.5.1. *Il p-value rappresenta la probabilità che i dati ci portino esattamente a rifiutare H_0.*
Per costruire un $p-$value devo

1. *Costruire la regione di rifiuto R.*
2. *Calcolata Z_0 in funzione del mio test statistico posso trovarmi in 3 casi diversi a seconda dell'ipotesi alternativa:*
   - *Nel caso in cui utilizzi un intervallo unilaterale destro, quindi $H_1 : \mu > \mu_0$, ho che*

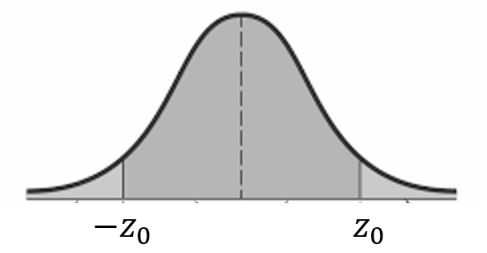
     $$p - value = \mathbb{P}(Z \geq Z_0)$$

   - *Nel caso in cui utilizzi un intervallo unilaterale sinistro, quindi $H_1 : \mu < \mu_0$, ho che*

     $$p - value = \mathbb{P}(Z \leq Z_0)$$

   - *Nel caso in cui utilizzi un intervallo bilaterale, quindi $H_1 : \mu \neq \mu_0$, ho che*

     $$p - value = 2\mathbb{P}(Z \geq Z_0)$$



Inoltre dato il livello di significatività α ho che se $p < \alpha$ allora i dati riscontrati sono sufficienti per scartare H_0 e considerare H_1.

Esempio 1.5.1. *Procedimento di costruzione di pavimenti quindi la solidità della superficie è data da $N(\mu_0, \sigma^2)$ con $\mu_0 = 4.5$ e $\sigma = 1.5$.*
Consideriamo una seconda squadra che viene addestrata e sia vuole vedere se è al livello dello standard precedente supponiamo $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$. Posto

$$W = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Livello di significatività $\alpha = 5\%$. Consideriamo un campione di $n = 25$ e $\overline{x} = 3.75$ posso rifiutare H_0?

$$R := \left\{|W(\underline{x})| > z_{1-\frac{\alpha}{2}}\right\} \qquad z_{0.975} = 1.96$$

Noi abbiamo che

$$|W(\underline{x})| = \left|\frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right| = 2.5$$

Dunque $R = \{x||W(x) > 1.96\}$ quindi siamo in R perciò rifiuto H_0.
In questo caso siccome $|W(x)| = 2.5$ allora il p-value è quel valore α tale per cui $z_{1-\frac{\alpha}{2}} = 2.5$ dunque

$$\alpha = 1.2\%$$

Questo vuol dire che avrei rifiutato H_0 anche al livello $\alpha = 2\%$ ma non al livello $\alpha = 1\%$

$$\mathbb{P}^{\mu_0}(|W(\underline{X})| \geq 2.5) = 2(1 - \Phi(2.5)) = 1.2\%$$

Quindi in questo caso il p-value ci ha dato più sicurezza nella nostra scelta.

Esempio 1.5.2. *Un'azienda di produzione di veicoli compra pneumatici da un fornitore che garantisce una percorrenza media di $\mu_0 = 35000km$, $n = 40$ vengono testati e si ottiene $\overline{x} = 34463km$ con una deviazione standard campionaria di $s = 1348km$, voglio vedere se $H_0 : \mu = \mu_0$ oppure $H_1 : \mu \neq \mu_0$ al livello di significatività $\alpha = 1\%$.*
Per definizione ho che

$$R = \left\{W(\underline{x}) = \frac{\overline{x} - \mu_0}{S/\sqrt{n}} \leq c\right\}$$

Dove

$$W(\underline{X}) = \frac{\overline{x} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \Longrightarrow c = -t_{1-\alpha}(39) = -2.423$$

Mentre

$$W(\underline{x}) = \frac{34463 - 35000}{1348/\sqrt{40}} = -2.52 < -2.423$$

Quindi rifiuto H_0 e accetto che la fornitura di pneumatici non soddisfa lo standard garantito.

1.5.1 Confronto tra medie

Date due popolazioni $X \sim N(\mu, \sigma^2)$ e $Y \sim N(\nu, \sigma^2)$. I parametri in questo caso sono 3 quindi

$$\Theta = \{(\mu, \nu, \sigma^2), \mu, \nu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$$

Per vedere se hanno la stessa media definisco

$$\Theta_0 = \{(\mu, \nu, \sigma^2)|\mu = \nu\}$$

Dunque come ipotesi nulla ho che $H_0 : \mu = \nu \iff H_0 : \theta \in \Theta_0$ mentre come ipotesi alternativa ho che $H_1 : \mu \neq \nu \iff H_1 : \theta \in \Theta_0^c$. Considero un campione di ampiezza m della prima popolazione

$X = (X_1, \cdots, X_m)$ e n della seconda popolazione $Y = (Y_1, \cdots, Y_n)$ dunque la funzione di massima somiglianza è data da

$$L(\theta|\underline{x}, \underline{y}) = (2\pi\sigma^2)^{-\frac{(m+n)}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2 + \sum_{j=1}^{n}(y_j-\nu)^2\right)\right)$$

Perciò posto $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right) = (\hat{\mu}, \hat{\nu}, \hat{\sigma^2}$ dunque si ottiene che

$$\begin{cases} \hat{\mu} = \hat{\theta}_1 = \overline{x} \\ \hat{\nu} = \hat{\theta}_2 = \overline{y} \\ \hat{\sigma^2} = \hat{\theta}_3 = \dfrac{1}{m+n}\left(\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{j=1}^{n}(y_j-\overline{y})^2\right) \end{cases} \implies L(\hat{\theta}|\underline{x}, \underline{y}) = (2\pi e s^2)^{\frac{-(m+n)}{2}}$$

Ora cerco il massimo su Θ_0 in questo caso

$$\hat{\theta}_1^* = \hat{\mu} = \hat{\nu} = \frac{1}{m+n}\left(\sum_{i=1}^{n} x_i + \sum_{j=1}^{n} y_j\right) = \frac{n\overline{x} + m\overline{y}}{n+m}$$

$$\overline{S}^2 = \hat{\theta}_3^* = \frac{1}{m+n}\left(\sum_{i=1}^{n}(x_i-\hat{\theta}_1^*)^2 + \sum_{j=1}^{n}(y_j-\hat{\theta}_1^*\right)$$

Dunque $L(\hat{\theta}^*|\underline{x}, \underline{y}) = (2\pi e\overline{s}^2)^{\frac{-(m+n)}{2}}$ allora

$$W(\underline{x}, \underline{y}) = \left(\frac{S^2}{\overline{S}^2}\right)^{\frac{(m+n)}{2}}$$

Se avessimo che $\mu = \nu$ allora $\overline{x}, \overline{y}$ e $\hat{\theta}_1^*$ devono essere vicini quindi $S^2, \overline{S}^2$ sono vicini quindi $W(\underline{x}, \underline{y}) \approx 1$. Se $\mu < \nu$ allora $\overline{x} < \hat{\theta}_1^* < \overline{y}$ quindi $\overline{S}^2 < \overline{S}^2$ quindi $W(x, y) << 1$ perciò la regione di rifiuto deve essere del tipo $R = \left\{\frac{S^2}{\overline{S}^2} \leq c\right\}$ ma per definizione ho che

$$\begin{aligned} \frac{S^2}{\overline{S}^2} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{j=1}^{n}(y_j-\overline{y}}{\sum_{i=1}^{n}(x_i-\hat{\theta}_1^*)^2 + \sum_{j=1}^{n}(y_j-\hat{\theta}_1^*)} \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{j=1}^{n}(y_j-\overline{y}}{\frac{mn}{m+n}(\overline{x}-\overline{y})^2 + \sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{j=1}^{n}(y_j-\overline{y}} \\ &= \frac{S_{xx}+S_{yy}}{\frac{mn}{m+n}(\overline{x}-\overline{y})^2 + S_{xx} + S_{yy}} \end{aligned}$$

Dunque

$$\frac{S^2}{\overline{S}^2} = \frac{1}{1 + \frac{mn}{m+n}\frac{(\overline{x}-\overline{y})^2}{S_{xx}+S_{yy}}}$$

Dunque la regione di rifiuto è tale per cui

$$R = \left\{\frac{mn}{m+n}\frac{(\overline{x}-\overline{y})^2}{S_{xx}+S_{yy}} \geq \underbrace{\frac{1}{c} - 1}_{=\epsilon^2}\right\} = \left\{\left|\frac{\overline{x}-\overline{y}}{\sqrt{(S_{xx}+S_{yy})\left(\frac{1}{m}+\frac{1}{n}\right)}}\right| \geq \epsilon\right\}$$

Perciò

$$\sup_{\theta\in\Theta_0} \mathbb{P}^\theta\left(\frac{\overline{X}-\overline{Y}}{\sqrt{(S_{xx}+S_{yy})\left(\frac{1}{m}+\frac{1}{n}\right)}} \geq \epsilon\right) \leq \alpha$$

Dato che per ipotesi ho che $\overline{X}, \overline{Y}$ sono gaussiane note

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)}} \sim N(0,1)$$

Per definizione ho che

$$\frac{1}{\sigma^2} S_{xx} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(m-1)$$

$$\frac{1}{\sigma^2} S_{yy} = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2(n-1)$$

Quindi ho che

$$\frac{1}{\sigma^2}(S_{xx} + S_{yy}) \sim \chi^2(m+n-2)$$

E quindi

$$\frac{\overline{X} - \overline{Y}}{\sqrt{(S_{xx} + S_{yy}) \left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t(n+m-2)$$

Allora

$$R = \left\{ \left| \frac{\overline{x} - \overline{y}}{\sqrt{(S_{xx} + S_{yy}) \left(\frac{1}{m} + \frac{1}{n}\right)}} \right| \geq t_{1-\frac{\alpha}{2}}(n+m-2) \right\}$$

1.6 Regressione Lineare

1.6.1 Approccio analitico

Definizione 1.6.1. *La regressione di Y v.a. su X è definita come la media condizionata* $\mathbb{E}[Y|X=x]$

$$\mathbb{E}[Y|X=x] = \alpha + \beta x$$

Servono sempre 2 steps da fare:

1. Descrivere i dati.
2. Fare inferenza sui dati per ottenere l'equazione della retta che meglio approssima i dati.

Uno dei modi per fare ciò è approssimare mediante la minima distanza verticale della retta dai dati. Dunque la retta sarà della forma

$$Y = \beta x + \alpha$$

Perciò il valore che avrei dovuto osservare è $\hat{y}_i = \alpha + \beta x_i$ dunque pongo $\epsilon_i = y_i - \hat{y}_i$.

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Questo mi porta al metodo dei minimi quadrati: procedimento puramente analitico per determinare la retta di regressione.

$$\begin{cases} \dfrac{\partial E}{\partial \alpha} = \displaystyle\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \\ \dfrac{\partial E}{\partial \beta} = \displaystyle\sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0 \end{cases} \Rightarrow \begin{cases} \alpha = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i - \dfrac{1}{n} \beta \sum_{i=1}^{n} x_i = \mathbb{E}[y] - \beta \mathbb{E}[x] = \overline{y} - \beta \overline{x} \\ \displaystyle\sum_{i=1}^{n} x_i (y_i - \overline{y}) - \beta \sum_{i=1}^{n} x_i (x_i - \overline{x}) \end{cases}$$

Dato che $x_i = \overline{x}$ allora

$$0 = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) - \beta \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Posto

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \qquad S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Allora

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

Dunque la retta di regressione è

$$y = \hat{\alpha} + \hat{\beta}x$$

Possiamo anche scartare lo scarto quadratico che è equivalente a

$$E = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Esempio 1.6.1. *Vogliamo vedere se e quanto la produzione del mais è dipendente dall'uso della quantità di fertilizzante usato. Creiamo n esperimenti in cui:*

- *Scegliamo la quantità di fertilizzante x_i.*
- *Misuriamo la quantità di mais prodotto y_i.*

Supponiamo che $n = 20$ con questi risultati

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	169	57	6	171	64	11	177	66	16	181	78
2	170	62	7	173	66	12	177	70	17	181	79
3	170	59	8	173	66	13	178	65	18	183	78
4	171	60	9	174	70	14	180	72	19	184	80
5	171	58	10	174	65	15	181	71	20	185	80

$$\mathbb{E}[Y|x] = \alpha + \beta x$$

Noi ci aspettiamo di avere una famiglia di v.a. $Y_i = \alpha + \beta x_i + \epsilon_i$ sotto questi ipotesi

- *ϵ_i v.a. tra loro indipendenti.*
- *La media di ϵ_i sia nulla $\forall i$.*
- *$V(\epsilon_i) = \sigma^2$.*

Dunque per quanto visto fino ad ora abbiamo che

$$\begin{cases} \overline{x} = 176.15 \\ \overline{y} = 63.3 \\ S_{xx} = 512.55 \\ S_{yy} = 1096.2 \\ S_{xy} = 696,1 \end{cases}$$

Perciò $\hat{\beta} = 1,3581$ e $\hat{\alpha} = -170.9313$ inoltre 150.82.

1.6.2 Approccio statistico

Siano

$$\underline{x} = (x_1, \cdots, x_n) \longrightarrow \text{ valori noti}$$
$$\underline{y} = (y_1, \cdots, y_n) \longrightarrow \text{ valori osservati}$$

Allora posto $\underline{Y} = (Y_1, \cdots, Y_n)$ v.a. le cui componenti sono indipendenti e vale che

$$\mathbb{E}[Y_i] = \alpha + \beta x_i$$

Supporremo inoltre che $V(Y_i) = \sigma^2, \forall i$. Definite delle v.a. (supposte indipendenti ed identicamente distribuite)

$$\epsilon_i = Y_i - \alpha - \beta x_i$$

Tali per cui $\mathbb{E}[\epsilon_i] = 0, \quad V(\epsilon_i) = \sigma^2$.

Definizione 1.6.2. *Uno stimatore si definisce lineare se*

$$T = \sum_{i=1}^{n} d_i Y_i$$

Dove $d_i \in \mathbb{R}$ fissati.

A questo punto dovremo cercare uno stimatore lineare corretto (quindi non distorto) per α, β. Per quel che riguarda β ho che

$$\mathbb{E}[T] = \sum_{i=1}^{n} d_i[\alpha + \beta x_i] = \beta$$

Questa richiesta viene dal fatto che lo abbiamo richiesto corretto, dunque il tutto può essere visto in maniera equivalente a

$$\alpha \left(\sum_{i=1}^{n} d_i \right) + \beta \left(\sum_{i=1}^{n} d_i x_i \right) = \beta$$

Quindi cercherò per β uno stimatore lineare $T = \sum_{i=1}^{n} d_i Y_i$ con le condizioni $\sum d_i = 0$ e $\sum_{i=1}^{n} d_i x_i = 1$.

Teorema 1.6.1. *Esiste un unico stimatore lineare coretto per β di varianza minima. Inoltre questo stimatore è*

$$T = \sum_{i=1}^{n} d_i Y_i \qquad d_i = \frac{x_i - \overline{x}}{S_{xx}}$$

Lo stimatore lineare di β è

$$b = \frac{S_{xy}}{S_{xx}} = \frac{1}{\sum (x_i - \overline{x})^2} \sum_{i=1}^{n} Y_i (x_i - \overline{x})$$

A questo punto posso calcolare

$$V(b) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Teorema 1.6.2. *Lo stimatore lineare corretto di varianza minima per α è*

$$a = \overline{Y} - b\overline{x}$$

Consideriamo $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$ è un vettore gaussiano centrato con $Cov(\epsilon)\sigma^2 Id_{n \times n}$ tale per cui ϵ_i sono v.a. gaussiane centrate identicamente distribuite indipendenti.

$$Y_i = \alpha + \beta x_i + \epsilon_i \sim N(\alpha + \beta x_i, \sigma^2)$$

Teorema 1.6.3. *Sotto queste ipotesi, b lo stimatore per β*

$$b = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) Y_i$$

Ha distribuzione normale di media β e varianza $\sigma^2 \frac{1}{S_{xx}}$.
Dato a lo stimatore di α

$$a = \overline{Y} b \overline{x}$$

Ha distribuzione normale di media α e varianza $\sigma^2 \frac{\overline{x^2}}{S_{xx}}$ dove $\overline{x^2} = \frac{1}{n} \sum x_i^2$.

Da questo possiamo dire che

$$MLE(\sigma^2) = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Teorema 1.6.4. *Lo stimatore $\hat{\sigma}^2$ è distorto (asintoticamente corretto) e vale che*

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - a - bx_i)^2\right] = \frac{n-2}{n}\sigma^2$$

Dimostrazione. **Fare i conti per dimostrare** □

Prenderemo quindi come stimatore di σ^2 la statistica

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - a - bx_i)^2$$

Osservazione 1.6.1.

$$\sum_{i=1}^{n}\frac{\epsilon_i^2}{\sigma^2} \sim \chi^2(n)$$

Teorema 1.6.5. *Gli stimatori (a, b) e S^2 sono indipendenti e $\frac{n-2}{\sigma^2}S^2 \sim \chi^2(n-2)$.*

Osservazione 1.6.2. *Posto $a = \sum_{i=1}^{n} d_i' Y_i$ e $b = \sum_{i=1}^{n} d_i Y_i$ allora*

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - a - bx_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n} d_j' Y_j - \sum_{k=1}^{n} d_k x_i Y_k\right)$$

Teorema 1.6.6. *Vale che*

$$\frac{a-\alpha}{\sqrt{\frac{S^2\overline{x^2}}{nS_{xx}}}} \sim t(n-2)$$

$$\frac{b-\beta}{\sqrt{\frac{S^2}{nS_{xx}}}} \sim t(n-2)$$

Con t di Student.

Osservazione 1.6.3. *È più importante β ed è diverso da 0 in quanto se non lo fosse avrei che*

$$\mathbb{E}[Y_i] = \alpha$$

E quindi non c'è più la relazione tra le variabili x e le variabili y. La questione fondamentale è poter fare un test di ipotesi su $H_0 : \beta = 0$ contro $H_1 : \beta \neq 0$. A questo punto noi sappiamo che b è uno stimatore per β e che segue una relazione t di Student e quindi la regione di rifiuto è identificata come

$$R := \left\{ y \middle| \left| \frac{b(y) - \beta}{\sqrt{\frac{S^2(y)}{S_{xx}}}} \right| \geq c \right\}$$

Sotto H_0 la funzione potenza vale

$$\mathbb{P}^{\beta=0}\left(\left|\frac{b}{\sqrt{\frac{S^2(y)}{S_{xx}}}}\right| \geq c\right) = \alpha$$

Quindi quello che io voglio sarà

$$R := \left\{ y \middle| \left| \frac{b(y) - \beta}{\sqrt{\frac{S^2(y)}{S_{xx}}}} \right| \geq t_{1-\frac{\alpha}{2}}(n-2) \right\}$$

Per il test al livello di significatività di α. In pratica per accettare l'esistenza di una relazione lineare tra x e Y a livello di significatività α devo verificare che

$$\left|\hat{\beta}\right| > \sqrt{\frac{S^2}{S_{xx}}} t_{-\frac{\alpha}{2}}(n-2)$$

Esempio 1.6.2. *Dato il seguente set di dati*

i	x_i	y_i	i	x_i	y_i
1	20	0.18	6	220	0.75
2	60	0.37	7	260	1.18
3	100	0.35	8	300	1.36
4	140	0.78	9	340	1.71
5	180	0.58	10	380	1.65

Con

- *x_i la velocità dell'aria in un motore.*
- *y_i coefficiente di evaporazione per il carburante.*

Si può calcolare che

$$\begin{cases} \overline{x} = 200 \\ \overline{y} = 0.891 \\ S_{xx} = 132000 \\ S_{yy} = 2.7489 \\ S_{xy} = 580.6 \end{cases}$$

Perciò ho che $\hat{\beta} = 0.0044$ e $\hat{\alpha} = 0.0113$. A questo punto calcolo

$$S^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 0.0244$$

$$\frac{\hat{\beta}}{\sqrt{\frac{S^2}{S_{xx}}}} = 10.2323$$

Dunque $\alpha = 5\%$ e $t_{1-\frac{\alpha}{2}}(8) = 2.306$ perciò rifiuto H_0 e accetto l'esistenza di una relazione lineare tra x e Y.

1.7 Intervalli di confidenza

Al variare del campione statistico la stima del parametro cambio dunque quello che dobbiamo fare è inferenza statistica su θ significa dire qualcosa su θ partendo dai dati. Dato $\underline{X} = (X_1, \cdots, X_N 9$ v.a. campione.
Il nostro obiettivo è vedere $\theta \in C$ con $C = C(\underline{x})$ quando faccio $C(\underline{X}) = [\theta_L(\underline{X}), \theta_U(\underline{X})]$.

Definizione 1.7.1. *Dato un modello statistico $(\Omega, \mathcal{A}, \mathbb{P}^\theta)$ un intervallo di confidenza di un parametro reale θ è definito da queste due funzioni $\theta_L(\underline{x}), \theta_U(\underline{x})$ che soddisfano*

$$\theta_L(\underline{x}) \leq \theta_U(\underline{x}),\ \forall \underline{x}$$

Nello spazio campionario.
Se $\underline{x}$ è l'osservazione faremo l'inferenza

$$\theta_L(\underline{x}) \leq \theta \leq \theta_U(\underline{x})$$

$\theta_L(\underline{X}), \theta_U(\underline{X})$ sono statistiche campionarie e $C(\underline{X}) = [\theta_L(\underline{X}), \theta_U(\underline{X})]$ è uno stimatore per intervalli di θ.

Definizione 1.7.2. *Sia $X = (X_1, \cdots, X_n)$ un campione statistico e θ il parametro da stimare, si definisce $1-\alpha$ come la confidenza dell'intervallo C, il seguente valore*

$$1 - \alpha := \inf_{\theta \in \Theta} \underbrace{\mathbb{P}^{\theta}(\theta \in C = [\theta_L, \theta_U])}_{probabilità\ di\ copertura\ di\ \theta}$$

Osservazione 1.7.1. *Avere una confidenza di $1-\alpha$ significa che $(1-\alpha)\%$ delle volte che stimo il mio parametro esso si trova all'interno di quell'intervallo.*

1.7.1 Intervalli di confidenza per Gaussiana e t-student

Quando studiamo un campione statistico che supponiamo $X \sim N(\mu, \sigma^2)$ nel momento in cui studio il suo intervallo di confidenza per μ ci possono essere 3 casi:

1. Conosco μ, σ^2.
2. Conosco σ^2.
3. Non conosco alcun termine.

1° caso

Non abbiamo nulla da trovare dato che sappiamo tutto quello che ci interessa.

2° caso

Supponiamo che $\hat{\mu} = \overline{x}$ (questa rappresenta una stima puntuale) dunque possiamo supporre che la forma dell'intervallo di confidenza sia del tipo:

$$C(\underline{x}) = [\overline{x} - \epsilon, \overline{x} + \epsilon]$$

Riuscire a determinare a priori μ è impossibile in quanto X v.a. continua e

$$\mathbb{P}^{\mu,\sigma^2}(\overline{X} = \mu) = 0$$

Dunque per quanto visto in precedenza posso affermare che se vogliamo una confidenza di $1 - \alpha$ allora per definizione ho che

$$\mathbb{P}^{\mu}\left(\left|\overline{X} - \mu\right| \leq \epsilon\right) = 1 - \alpha$$

Dunque

$$\epsilon = \sigma^2 \phi_{1-\frac{\alpha}{2}}$$

Dunque mediante le tabelle ho la soluzione.

Esempio 1.7.1. *Prendiamo in considerazione una popolazione descritta da una funzione*

$$f(x|\theta) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$$

Mentre $\Theta = \mathbb{R}_+$ e come stimatore di θ prendiamo $Y = \max\{X_1, \cdots, X_n\}$. Prendiamo in considerazione due intervalli di confidenza:

1. *$C = [aY, bY]$ con $1 \leq a \leq b$.*
2. *$C' = [Y + c, Y + d]$*

I intervallo

In questo caso vedo che

$$\mathbb{P}^{\theta}(\theta \in C) = \mathbb{P}^{\theta}(aY \leq \theta \leq bY) = \mathbb{P}^{\theta}\left(\frac{\theta}{b} \leq Y \leq \frac{\theta}{a}\right)$$

Allora possiamo calcolare la funzione di ripartizione di x come

$$\mathbb{P}^{\theta}(Y \leq x) = \begin{cases} \left(\frac{x}{\theta}\right)^n & x \leq \theta \\ 1 & x \geq \theta \end{cases}$$

Quindi

$$\mathbb{P}^{\theta}(\theta \in C) = \frac{1}{a^n} - \frac{1}{b^n}$$

Questo non dipende da θ e quindi se vogliamo $(1 - \alpha) = 95\%$ allora ho che $\frac{1}{a^n} - \frac{1}{b^n} = 95\%$ dunque una possibile soluzione è $\alpha = 1, b = \alpha^{-\frac{1}{n}}$ dunque ottengo

$$X = \left[Y, \frac{Y}{\alpha^{\frac{1}{n}}}\right]$$

II intervallo

In questo caso vedo che

$$\mathbb{P}^\theta\left(\theta \in C' = [Y+c, Y+d]\right) = \mathbb{P}^\theta(\theta - d \leq Y \leq \theta - c) = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n$$

Da questo si nota che

$$\lim_{\theta \to +\infty} \mathbb{P}^\theta\left(\theta \in C' = [Y+c, Y+d]\right) = 0$$

Quindi la confidenza di questo intervallo è nulla.

Osservazione 1.7.2. *Questo procedimento è equivalente a fare l'inverso del test di ipotesi dato che se prendiamo una popolazione descritta da*

$$f(x|\mu) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Supponiamo che $H_0 : \mu = \mu_0$ e che $H_1 : \mu \neq \mu_0$ quindi definiamo la regione di rifiuto come

$$R = \left\{\underline{x} | |\overline{x} - \mu_0| > \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right\}$$

Dunque posso considerare il suo complementare: la regione di accettazione come

$$A = \left\{\underline{x} | |\overline{x} - \mu_0| \leq \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right\}$$

Perciò ho che

$$\mathbb{P}^{\mu_0}(\underline{X} \in A) = 1-\alpha = \mathbb{P}^{\mu_0}\left(\overline{X} \in \left[\mu_0 - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \mu + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right]\right) = \mathbb{P}^{\mu_0}\left(\mu_0 \in \left[\overline{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right]\right)$$

Quindi se io definisco $C = [\theta_L(\underline{X}), \theta_U(\underline{X})] = \left[\overline{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right]$ questo è un intervallo aleatorio quindi è un intervallo di confidenza al livello $1-\alpha$.

I parte: costruzione teorica

Sia $\underline{X}$ una v.a. campione allora $C := [\overline{X} - \epsilon, \overline{X} + \epsilon]$ è un intervallo aleatorio quindi posto μ il valore vero del parametro è un numero reale incognito ed ha senso chiederci $\mathbb{P}^\mu(\mu \in C) \geq 1 - \alpha$.

II parte: campionamento

$\underline{x}$ è un vettore e $[\overline{x} - \epsilon, \overline{x} + \epsilon]$ è un intervallo reale che comprende oppure no il valore vero μ quindi nel momento in cui faccio campionamento non posso più fare probabilità.

III parte

Se facessi tanti campionamenti, la legge dei grandi numeri mi direbbe che la frequenza di intervalli che comprendono il valore vero μ tenderebbe a $1 - \alpha$.

$$A(\mu_0) = \left\{\underline{x} | \mu_0 - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \leq \overline{x} \leq \mu_0 + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right\}$$

Rappresenta il sottoinsieme dello spazio campionario e

$$C(\underline{x}) = \left[\mu_0 | \overline{x} - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \leq \mu_0 \leq \overline{x} + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right\}$$

Rappresenta un sottoinsieme dello spazio dei parametri. Perciò $\underline{x} \in A(\mu_0) \iff \mu_0 \in C(\underline{x})$.

3° **caso**

Nel caso in cui non si conosca né μ né σ^2 allora bisogna ricorrere alla t-student in quanto per definizione sappiamo che

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{S_n^2}{n}}} \sim t(n-1)$$

Dunque se vogliamo avere un intervallo di confidenza del tipo

$$C(\underline{X}) = \left[\overline{X} - \epsilon, \overline{X} + \epsilon\right)$$

Con grado di confidenza di $1-\alpha$ allora per definizione ho che

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left(\overline{X} - \epsilon\sqrt{\frac{S_n^2}{n}} \leq \mu \leq \overline{X} + \epsilon\sqrt{\frac{S_n^2}{n}}\right) \\ &= \mathbb{P}(|T| \leq \epsilon) \\ &= 2(1 - t(\epsilon)) \end{aligned}$$

Dunque ho che

$$\epsilon = t_{1-\frac{\alpha}{2}}(n-1)$$

E questi sono valori tabulati. 2

Teorema 1.7.1. *$\forall \theta_0 \in \Theta$ indico con $A(\theta_0)$ la regione di accettazione al livello di significatività α nel test di $h_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$. Se $\forall \underline{x}$ indico con*

$$C(\underline{x}) = \{\theta_0 | \underline{x} \in A(\theta_0)\}$$

Allora $C(\underline{x})$ è un intervallo di confidenza al livello $1-\alpha$.
Viceversa, se $\forall \underline{x}$ indico con $C(\underline{x})$ intervallo di confidenza al livello $1-\alpha$, allora ponendo $A(\theta_0) = \{\underline{x} | \theta_0 \in C(\underline{x})\}$, $A(\theta_0)$ è la regione di accettazione di un test d'ipotesi al livello α per $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$.

Esempio 1.7.2. *Data una popolazione con $f(x|\lambda) = \lambda e^{-\lambda x}\mathbb{1}_{x>0}(x)$, trovare un intervallo di confidenza per la media $\mu = \frac{1}{\lambda}$ dunque la funzione di verosomiglianza è*

$$\frac{L(\mu_0|\underline{x})}{\sup_\mu L(\mu|\underline{x})} = \frac{\left(\frac{1}{\mu_0}\right)^n \exp\left(-\frac{1}{\mu_0}\sum_{i=1}^n x_i\right)}{\sup_{\mu>0}\left(\frac{1}{\mu}\right)^n \exp\left(-\frac{1}{\mu}\sum_{i=1}^n x_i\right)} = \frac{\left(\frac{1}{\mu_0}\right)^n \exp\left(-\frac{n\overline{x}}{\mu_0}\right)}{\left(\frac{1}{\overline{x}}\right)^n \exp\left(-\frac{n\overline{x}}{\overline{x}}\right)} = \left(\frac{\overline{x}}{\mu_0}\right)^n \exp\left(n\left(1 - \frac{\overline{x}}{\mu_0}\right)\right)$$

Con

$$A(\mu_0) = \left\{\underline{x} | \left(\frac{\overline{x}}{\mu_0}\right)^n \exp\left(n\left(1 - \frac{\overline{x}}{\mu_0}\right)\right) \geq k\right\} = \left\{\underline{x} | \frac{\underline{x}}{\mu_0} e^{1-\frac{x}{\mu_0}} \geq k^*\right\}$$

Dove k^* è tale per cui $\mathbb{P}^{\mu_0}(\underline{X} \in A(\mu_0)) = 1-\alpha$. Inoltre

$$\begin{aligned} &ae^{1-a} = be^{1-b} \\ &\mathbb{P}^{\mu_0}\left(a \leq \frac{\overline{x}}{\mu_0} \leq b\right) = 1 - \alpha \end{aligned}$$

Nel momento in cui posso scrivere

$$A(\mu_0) = \left\{\underline{x} | a \leq \frac{\overline{x}}{\mu_0} \leq b\right\} \Rightarrow C(\underline{x}) = \left\{\mu | \frac{\overline{x}}{b} \leq \mu \leq \frac{\overline{x}}{a}\right\}$$

Posto

$$S_n = \sum_{i=1}^n X_i \sim \Gamma(n, \lambda) = \Gamma(n, 1/\mu)$$

Allora

$$\mathbb{P}^{\mu_0}\left(a \leq \frac{S_n}{n\mu_0} \leq b\right) = \int_{n\mu_0 a}^{n\mu_0 b} \frac{1}{(n-1)!} e^{-\frac{x}{\mu_0}} \left(\frac{1}{\mu_0}\right)^n x^{n-1}\, dx = \int_{na}^{nb} \frac{1}{(n-1)!} e^{-x} x^{n-1}\, dx$$

Perciò il sistema diventa

$$ae^{1-a} = be^{1-b}$$
$$\int_{na}^{nb} \frac{1}{(n-1)!} e^{-x} x^{n-1}\, dx = 1-\alpha$$

Sia T stimatore di θ (una statistica sufficiente), T v.a. continua $F_T(t|\theta)$ funzione di ripartizione $\mathbb{P}^\theta(T \leq t) = F_T(t|\theta)$. Da ricordare che $\forall \theta, F_T(T|\theta) \sim Unif(0,1)$ sotto $\mathbb{P}^\theta$, a questo punto prendiamo $\alpha = \alpha_1 + \alpha_2$ con $\alpha = 5\%$ dato che non ci sono motivi per scegliere qualcosa al posto di un'altra allora potrei fare sì che

$$\alpha_1 = \alpha_2 = \frac{\alpha}{2} \qquad \alpha_1 = 0, \alpha_2 = \alpha \qquad \alpha_1 = \alpha, \alpha_2 = 0$$

Supponiamo che $H_0 : \theta = \theta_0$ mentre $H_1 : \theta \neq \theta_0$. Una possibile regione di accettazione per H_0 è

$$A(\theta_0) = \{t | \alpha_1 \leq F_T(t|\theta) \leq 1 - \alpha_2\}$$

Mentre

$$C(t(\underline{x})) = \{\theta_0 | \alpha_1 \leq F_T(t|\theta) \leq 1 - \alpha_2$$

A questo punto ci sono delle condizioni per cui $C(t)$ è un intervallo, quello che conviene richiedere è che $F_T(t|\theta)$ sia monotona (crescente o decrescente) in θ per ogni t fissato.

Teorema 1.7.2. *Valgono i seguenti fatti:*

1. *Se $F_T(t|\theta)$ è monotona crescente in θ per ogni t allora definisco $\theta_L(t)$ e $\theta_U(t)$ dalle relazioni $F_T(t|\theta_L(t)) = \alpha_1$ e $F_T(t|\theta_U(t)) = 1 - \alpha_2$.*
2. *Se $F_T(t|\theta)$ è monotona decrescente in θ per ogni t allora definisco $\theta_L(t)$ e $\theta_U(t)$ dalle relazioni $F_T(t|\theta_L(t)) = 1 - \alpha_2$ e $F_T(t|\theta_U(t)) = 1 - \alpha_1$.*

Allora $C(t(\underline{x})) = [\theta_L(t), \theta_U(t)]$ è un intervallo di confidenza al livello $1 - \alpha$ per θ.

Esempio 1.7.3. *Data una funzione $f(x|\theta) = \frac{1}{\theta}\mathbb{1}_{(0,\theta)}(x)$, $\underline{X} = (X_1, \cdots, X_n)$ e $Y = \max_{i=1,\cdots,n} X_i$*

$$F_Y(t|\theta) = \mathbb{P}^\theta(Y \leq t) = \prod_{i=1}^{n} \mathbb{P}^\theta(x_i \leq t) = \left(\frac{t}{\theta}\right)^n \mathbb{1}_{t \leq \theta} + \mathbb{1}_{t > \theta}$$

Ponendo $\alpha = 5\%$ e $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ allora

$$F_Y(t|\theta_L) = \left(\frac{t}{a}\right)^n \mathbb{1}_{t \leq \theta_L} = 1 - \frac{\alpha}{2} \qquad F_Y(t|\theta_U) = \left(\frac{t}{a}\right)^n \mathbb{1}_{t < a} = \frac{\alpha}{2}$$

Quindi

$$\theta_L(t) = \frac{t}{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{n}}} \qquad \theta_U(t) = \frac{t}{\left(\frac{\alpha}{2}\right)^{\frac{1}{n}}}$$

E

$$C(t(\underline{x})) = \left[\frac{\max x_i}{\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{n}}}, \frac{\max x_i}{\left(\frac{\alpha}{2}\right)^{\frac{1}{n}}}\right]$$