Statistica

June 15, 2011

Part I

Teoria

Il termine *statistica* indica una *variabile aleatoria* che è semplicemente una funzione dei dati di un campione. I due principali esempi di statistiche sono la media campionaria e la varianza campionaria.

$$\text{statistica} = f\left(\text{VA}, \text{param noti}\right)$$

L'**inferenza** è l'obiettivo finale del corso di statistica. lo **Stimatore** (2 on page 3) la **Varianza Campionaria** (5.2 on page 5) sono i primi strumenti che abbiamo introdotto per fare inferenza sulle **distribuzioni Normali**, perchè a queste tutte possono essere ricollegate grazie al **teorema centrale del limite**.

1 Modello Esponenziale

Utilizzato per individuare modelli di sopravvivenza di macchinari *non* soggetti ad usura e *non* soggetti a rodaggio. I macchinari non di questo tipo hanno modelli riconducibili a quelli esponenziali con qualche trasformazione in più.

$$f(x,\beta) = \left\{ \begin{array}{cc} \frac{1}{\beta} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{altrimenti} \end{array} \right\}$$

$\frac{1}{\beta}$ è il parametro relativo alla vita del macchinario. Dal calcolo del valore atteso si ottiene, infatti, che il tempo medio di vita di un macchinario è β.

$$\begin{aligned}
\mathrm{E}(x) &= \int_0^{+\infty} \underbrace{x}_{g} \underbrace{\frac{1}{\beta} e^{\frac{-x}{\beta}}}_{f'} dx \\
&= x\left[-e^{-\frac{x}{\beta}}\right]_0^{+\infty} - \int_0^{+\infty} 1\left(-e^{-\frac{x}{\beta}}\right) dx \\
&= -\underbrace{\frac{x}{e^{\frac{x}{\beta}}}_{x\to\infty}}_{=0} + \underbrace{\frac{x}{e^{\frac{x}{\beta}}}_{x\to 0}}_{=0} - \int_0^{+\infty} 1\left(-e^{-\frac{x}{\beta}}\right) dx \\
&= -\beta \int_0^{+\infty} \left(-\frac{1}{\beta}\right) e^{-\frac{x}{\beta}} dx \\
&= -\beta \left(\frac{1}{e^{\frac{+\infty}{\beta}}} - \frac{1}{e^{\frac{0}{\beta}}}\right) \\
&= \beta
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) &= \int_0^{+\infty} (x-\beta)^2 \frac{1}{\beta} e^{-\frac{x}{\beta}} dx \\
&= (x-\beta)^2 \left[-e^{-\frac{x}{\beta}}\right]_0^{+\infty} - \int_0^{+\infty} 2(x-\beta)\left(-e^{-\frac{x}{\beta}}\right) dx \\
&= \beta^2 - 2\left(-\int_0^{+\infty} x e^{-\frac{x}{\beta}} dx + \beta \int_0^{+\infty} e^{-\frac{x}{\beta}} dx\right) \\
&= \beta^2 - 2\left(-\beta^2 + \beta^2\right) \\
&= \beta^2
\end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) \\ &= 2\beta^2 - \beta^2 \\ &= \beta^2 \end{aligned}$$

L'importante osservazione da fare a questo punto è che la media del modello esponenziale coincide proprio con il parametro incognito β . Per questa forte correlazione è possibile utilizzare lo stimatore per ottenere tale parametro.

Anche nella Poisson è possibile ragionare allo stesso modo poichè $\mathrm{E}(X \sim \text{Poisson}) = \lambda$.

2 Stimatore

Lo stimatore permette di ottenere una stima a partire da un campione di dati. Si utilizza lo stimatore quando la totalità dei soggetti da misurare è troppo elevata e si preferisce lavorare sui campioni.

La stima, ottenuta come risultato dello stimatore, è pertanto in funzione del campione analizzato.

$\frac{x_1, x_2 \ldots x_n}{n} = \overline{X} \longrightarrow$ Media empirica o campionaria

$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{\sum_i^n x_i}{n}\right) = \frac{1}{n}\mathbb{E}\left(\sum_i^n x_i\right) = \frac{1}{n}\sum_i^n \mathbb{E}\left(x_i\right)$ ogni campione $x_i$ha la stessa media, quindi $\frac{n\mu}{n} = \mu$. Conclusione: la media campionaria è aleatoria, ma mediamente otterrai quel parametro che non conoscevi.

Nel caso dell'esponenziale quel μ è β

$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{\sum_i^n x_i}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_i^n x_i\right)$ (n^2 perchè la varianza è un operatore quadratico)$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

2.1 Stimatore non distorto

Uno stimatore distorto è uno stimatore che per qualche ragione ha valore atteso diverso dalla quantità che stima; uno stimatore non distorto è detto stimatore corretto.[1]

$\overline{X}$ è uno stimatore non distorto quando $\lim_{n\to\infty} \mathrm{Var}\left(\overline{X}\right) = 0$ significa che $\overline{X}$ si concentra a $\mathbb{E}\left(\overline{X}\right)$ con infiniti esperimenti (nota: la distorsione è chiamata anche *Bias*).

[1] Wikipedia

2.1.1 Consistente in media quadratica

È un attributo riferibile alla media aleatoria quando il suo valor medio coincide con quello reale

$$\lim_{n\to\infty} \mathrm{E}\left(X\right) = \theta \text{ , dove } \theta \text{ corrisponde al valor medio reale}$$

2.2 Mean Square Error (MSE)

È uno degli strumenti possibili per misurare la bontà dello stimatore. Prendendo in considerazione un campione T composto da n individui: $T = \{y_1, y_2, ..., y_n\}$ posso valutare che

$$\begin{aligned} \mathrm{E}\left(T\right) = \theta &\rightarrow \text{allora } MSE = \mathrm{Var}\left(T\right) \\ \mathrm{E}\left(T\right) \neq \theta &\rightarrow \text{allora } MSE = f\left(\mathrm{Var}\left(T\right), \mathrm{E}\left(T\right)\right) \end{aligned}$$

MSE è il momento secondo dell'errore, ovvero di $T-\theta$. $\mathrm{MSE} = \mathrm{Var}\left(T\right) + \left(\mathrm{E}\left(T\right) - \theta\right)^2$. Con poche misurazioni è possibile che ci sia un errore, ma con $n \to \infty$ è asintoticamente non distorto.

Dato che è difficile che le misurazioni non siano soggette ad errori (casuali e non sistematiche) è logico calcolare la probabilità che un certo valore sia all'interno di un range. Per proseguire il ragionamento è necessario introdurre il concetto di *precisione* e *accuratezza* della stima.

Stimare $\Longrightarrow$ Approssimare $\Longrightarrow$ Precisione

3 Precisione e Accuratezza della stima

Accuratezza: è la capacità da parte dello stimatore di eseguire misurazioni il cui valor medio corrisponde a quello reale

Precisione: la precisione è il grado di 'convergenza' (o se vogliamo 'dispersione') di dati, ovvero la capacità da parte dello stimatore di eseguire misurazioni prossime al valor medio della serie a cui appartengono (non necessariamente a quello reale).

$$P\left(\left|\bar{X}-\mu\right| \leq \underbrace{\delta}_{\text{precisione}}\right) = \underbrace{\gamma}_{\text{accuratezza}}$$

4 Teorema centrale del limite

Il teorema centrale del limite è importante perchè permette di approssimare una qualsiasi distribuzione di partenza ad una normale sotto certe condizioni. Il motivo di questa trasformazione è dovuta al fatto che la curva rappresenta la probabilità con la quale una VA (un campione) abbia una certa media.

Non si ragiona più sui singoli individui (distribuzione originale) ma sulla media dei campioni selezionati. Indipendentemente dalla distribuzione originale, quindi, i campioni che avranno una media lontana da quella reale saranno meno rispetto a quelli che ne avranno una media vicino a quella reale.

5 Modello Gaussiano

$X \sim N\left(\mu, \sigma^2\right)$ se $f\left(x, \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \left\{ \begin{array}{l} \mu \in \mathbb{R} \\ \sigma^2 > 0 \end{array} \right\}$

5.1 Teorema

Se $X_1, ..., X_n$ sono n variabili casuali Normali tra loro indipendenti, ciascuna con valore atteso μ_i e varianza σ_i^2

allora la variabile casuale $Y = a_1X_1 + a_2X_2 + ... + a_nX_n$ è a sua volta una variabile casuale Normale con valore atteso $\mu = a_1\mu_1 + a_2\mu_2 + ... + a_n\mu_n$ e varianza $\sigma^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + ... + a_n^2\sigma_n^2$.

5.2 Caso in cui σ non è noto

$X_1, X_2, ..., X_n$ iid con μ, σ^2 incognite

Obiettivo: stimare $\sigma^2 \rightarrow$ calcolo un indice quadratico di dispersione della variabile $\bar{X}$.

Devo calcolare la *Varianza Campionaria*: $S^2 = \frac{\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^2}{n-1}$; come denominatore c'è $n-1$ perchè devo avere almeno un'osservazione, altimenti non ho dispersione.

Se n grande non importa la forma dela densità, allora approssimativamente $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{\mu}}} \sim N(0,1) \rightarrow x \mp \bar{z_{\frac{1+\gamma}{\sqrt{n}}}} \frac{s}{\sqrt{n}}$ il pedice di z è un *intervallo di confidenza IC(μ)* asintotico di livello approssimato γ

$X_1, X_2, ..., X_n$ iid $\sim N(0,1)$ con μ e σ^2 incognite.

Quantità di partenza per costruire IC(μ) è $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} = Q$ maggiore variabilità di Q $\rightarrow$code più pesanti nella distribuzione

6 Significato Distribuzioni

6.1 Bernoulliana

Qual'è la probabilità che il risultato sia x_0 ?

6.2 Binomiale

Qual'è la probabilità che vinca k volte se eseguo n prove?

$X \sim \text{Bin}(n,p)$ è una distribuzione **discreta**. Ogni prova può terminare o con successo o senza successo. $P(X=k) = \binom{n}{k} p^k (n-p)^{n-k}$; $\text{E}(X) = np$; $\text{Var}(X) = np(1-p)$. La binomiale serve a calcolare il numero di successi in un numero dato di prove.

$$\text{Approssimazioni} \rightarrow \left\{ \begin{array}{ll} \text{P} & \left(n > 20 \wedge p < \frac{1}{20}\right) \vee (n > 100 \wedge np < 10) \\ \text{N} & n \text{ tende all'infinito lasciando } p \text{ fisso} \end{array} \right\}$$
$$\rightarrow \left\{ \begin{array}{c} \text{Poiss}(np) \\ \text{Norm}(np, npq) \end{array} \right\}$$

6.3 Poisson

Qual'è la probabilità che vinca k volte se mediamente ci sono λ vincite?

$X \sim \text{Poiss}(\lambda)$; il numero di arrivi in un certo numero di tempo è la variabile di poisson. $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$; $\text{E}(X) = \lambda$; $\text{Var}(X) = \lambda$. Se X sono le telefonate urbane ed Y quelle interurbane la somma di queste due variabili è la somma delle telefonate.

La distribuzione di Poisson è una distribuzione di probabilità **discreta** che esprime le probabilità per il numero di eventi che si verificano successivamente ed indipendentemente in un dato intervallo di tempo, sapendo che mediamente se ne verifica un numero λ. Questa distribuzione è anche nota come *legge degli eventi rari*.

λ rappresenta il numero medio di eventi per intervallo di tempo. Quindi determinare $P(X = k)$ significa determinare la probabilità che in una unità di tempo si verifichino k successi sapendo che mediamente ce ne sono λ.

La poissoniana può essere anche interpretata come il tempo che intercorre mediamente tra due eventi vincenti.

La poissoniana è un'approssimazione della binomiale. There is a rule of thumb stating that the Poisson distribution is a good approximation of the binomial distribution if n is at least 20 and p is smaller than or equal to 0.05, and an excellent approximation if n $\geq$ 100 and np $\leq$ 10

Es: vogliamo contare il numero di guasti che accadono in una determinata centrale elettrica, conoscendo λ ovvero il numero medio di guasti nell'unità di tempo, useremo una variabile di Poisson così definita:

$P(X_t = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$ Probabilità che nell'intervallo di tempo $[0, t]$ ci siano stati k guasti

6.4 Esponenziale

Qual'è la probabilità che non si guasti prima del tempo t_1 ?

$X \sim \epsilon(\lambda)$; $f_x(k) = \left\{ \begin{array}{cc} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{altirmenti} \end{array} \right\}$. Assenza di memoria, ovvero $P(X > t + s | X > s) = P(X > t)$ Nonostante il suo storico, il suo tempo di funzionamento futuro non dipende dal passato.

La distribuzione esponenziale è una distribuzione di probabilità **continua** che descrive la durata di vita di un fenomeno che non invecchia (ovvero è priva di memoria)

Formalmente è espressa nel seguente modo: $f(x, \lambda) = \lambda e^{-\lambda x}$ ma nel corso di statistica è preferibile modificarla nella seguente forma $f(x, \beta) = \frac{1}{\beta} e^{-\frac{1}{\beta}}$

L'esponenziale è un'approssimazione della distribuzione Γ.

La distribuzione esponenziale rovescia il discorso fatto nella poissoniana, perchè serve a determinare qual'è il tempo di attesa per il primo guasto. Quindi passiamo da una VA discreta ad una continua T: il tempo di attesa per il primo guasto.

Quindi interpretiamo la funzione di ripartizione di T $\rightarrow F(t) = P(T \leq t)$ come la sopravvivenza del sistema che stiamo studiando. Infatti la probabilità che l'oggetto sia ancora attivo dopo pochissimo tempo è maggiore rispetto alla probabilità che l'oggetto sia ancora attivo dopo un lungo periodo. Il grafico che rappresenta meglio questo andamento è l'esponenziale che una curva tanto ripida quanto maggiore è il valore λ, che indica il numero di 'disattivazione' nell'arco temporale.

6.5 Normale

È una distribuzione di probabilità continua, considerata il caso base delle distribuzioni di probabilità a causa del suo ruolo nel teorema del limite centrale[2].

$X \sim N(\mu, \sigma^2)$... ??? $P(Z \leq x) = \Phi(x)$

La Normale è un'approssiamazione della binomiale

6.6 Variabili aleatorie discrete e continue

In una VA continua la probabilità che X assuma un determinato valore è pari a 0. Si deve pertando calcolare la probabilità (l'area) su range di valori.

7 Esercizi

Ripasso delle distribuzioni di densità note

[2]La somma (normalizzata) di un grande numero di VA è distribuita approssimativamente come una VA normale standard

7.1 Esercizio

Abbiamo un batteria la cui durata è una va la cui densità è

$$f_x(x,\theta) = \frac{1}{2\theta\sqrt{x}} e^{-\frac{\sqrt{x}}{\theta}} 1_{[0,+\infty]}(x) = \left\{ \begin{array}{cc} \frac{1}{2\theta\sqrt{x}} e^{-\frac{\sqrt{x}}{\theta}} & 0 < x < +\infty \\ 0 & \text{altrove} \end{array} \right\}$$

La funzione indicatrice è un modo compatto per dire che l'espressione in in funzione di x è valida nell'intervallo [0,+∞], 0 altrimenti.

Determinare la probabilità che la batteria funzioni ancora dopo 11h: $P(X > 11)$. O la calcoliamo direttamente con la funzione densità, oppure scriviamo la F grande (funzione di ripartizione) e calcolare $1 - \int_0^{11}$.

$$\begin{aligned} P(X > 11) &= 1 - \int_0^{11} \frac{1}{2\theta\sqrt{x}} e^{-\frac{\sqrt{x}}{\theta}} dx \\ &= \text{procedo con la sostituzione } \frac{\sqrt{x}}{\theta} = t \\ &= 1 - \int_{\frac{0}{\theta}}^{\frac{\sqrt{11}}{\theta}} \frac{1}{2\theta t\theta} e^{-t} 2\theta^2 t \; dt \\ &= 1 - \int_0^{\frac{\sqrt{11}}{\theta}} e^{-t} dt \\ &= 1 - \left[-e^{-t}\right]_0^{\frac{\sqrt{11}}{\theta}} \\ &= e^{-\frac{\sqrt{11}}{\theta}} \end{aligned}$$

θ è la VA che indica in quanto tempo la batteria di scarica.

7.2 Funzione densità di probabilità di Y a partira da quella di X

Data una funzione di una X dobbiamo determinare la funzione di una $Y = \sqrt{X}$. La prima via è una formula precotta che funziona solo se la trasformazione g è invertibile; l'altra strada è una strategia che si può usare sempre, sia che la trasformazione sia invertibile, sia che non la sia.

7.2.1 Prima strada

$Y = g(X) \rightarrow X = h(Y)$, con g invertibile

$f_y(y) = f_x(h(y))\,|h'(y)|$.

Nel nostro esercizio $h(y) = y^2$

$f_y(y) = \frac{1}{2\theta y} e^{-\frac{y}{\theta}}\,|2y|\,1_{[0,+\infty]}\,(y^2) = \frac{1}{\theta} e^{\frac{-y}{\theta}} 1_{[0,+\infty]}(y)$. E' diventata la funzione esponenziale nota.

y e non modulo di y nel denominatore della prima frazione perchè secondo come definita è sempre positiva; la funzione indicatrice è sempre verificata perchè $y = \sqrt{x}$.

7.2.2 Seconda strada

Dobbiamo ancora vederla

7.3 Esercizio

Un esercizio possibile è quello di riuscire a determinare una funzione densità di probabilità di una Y a partire da una X di cui si conosce la funzione densità di probabilità e la sua relazione con Y

$f_x(x) = \left\{ \begin{array}{cc} \frac{1}{2}x & 0 \leq x \leq 2 \\ 0 & \text{altrimenti} \end{array} \right\} = \frac{x}{2} 1_{[0,2]}(x)$ con $y = x^2$ la trasformazione è invertibile grazie al dominio ristretto $[0, 2]$. Sarebbe infatti $x = \pm\sqrt{y}$ ma $x = -y$ non è possibile perchè x assume solo valori positivi (y sempre positiva).

$h(y) = \sqrt{y} \rightarrow h = \frac{1}{2\sqrt{y}}$

$$\begin{aligned} f_y(y) &= f_x(h(y))\,|h'(y)| \\ &= \frac{\sqrt{y}}{2} 1_{[0,+2]}\,(\sqrt{y})\,\frac{1}{2\sqrt{y}} \\ &= \frac{1}{4} 1_{[0,+2]}\,(\sqrt{y}) \\ &= \frac{1}{4} 1_{[0,+4]}\,(y) \text{ più bella da scrivere ma equivalente} \end{aligned}$$

E' la VA nota uniforme, perchè è un quarto solo tra 0 e 4.

7.4 Esercizio

$f_x(x) = \frac{3}{2}x^2 1_{[-1,1]}(x)$ con $Y = X^2$. Non possiamo usare la trasformazione perchè la trasformazione non invertibile nell'intervallo imposto.

Lo scopo del gioco è vedere come è fatta Y sapendo X, passando dalla funzione di ripartizione e non quella di densità.

$$F_x(x) = \left\{ \begin{array}{cc} 0 & x < -1 \\ \int_{-1}^{x} \frac{3}{2}t^2 dt = \left[\frac{t^3}{2}\right]_{-1}^{x} = \frac{x^3+1}{2} & -1 \leq x \leq 1 \\ 1 & x > 1 \end{array} \right\}$$

Il dominio di Y è $[0, 1]$ perchè i valori negativi di X diventano tutti positivi elevandoli al quadrato. Quindi subito possiamo dire che al di fuori dei suoi estremi la F_y assume 0 quando $y < 0$ e 1 quando $y > 1$.

$F_y(y) = \left\{ \begin{array}{cc} 0 & y < 0 \\ stellina & 0 \leq y \leq 1 \\ 1 & y > 1 \end{array} \right\}$; $stellina = F_y(y) = P(Y \leq y) = P(x^2 \leq y)$ riusciamo a porre tutto in funzione di x? $P\left(-\sqrt{y} \leq x \leq \sqrt{y}\right) = F_x\left(\sqrt{y}\right) - F_x\left(-\sqrt{y}\right) = \frac{\left(\sqrt{y}\right)^3+1}{2} - \frac{\left(-\sqrt{y}\right)^3+1}{2} = y^{\frac{3}{2}}$.

7.5 Esercizio

Z è la normale standard, vogliamo calcolare $Y = Z^2$. Le operazioni sono quelle di prime con la differenza che tratto con variabili normali. Partiamo da considerazioni di base

$F_y(y) = \left\{ \begin{array}{cc} 0 & y < 0 \\ stellina & y \geq 0 \end{array} \right\}$; stellina sarà in funzione di Φ. $F_y(y) = P(Y \leq y) = P\left(-\sqrt{y} \leq z \leq \sqrt{y}\right) = \Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right) = 2\Phi\left(\sqrt{y}\right) - 1$ [3]

$f_y(y) =$ è la derivata della funzione di ripartizione $= 2\frac{d}{dy}\left(\Phi\left(\sqrt{y}\right)\right) = 2\Phi'\left(\sqrt{y}\right)\frac{1}{2\sqrt{y}}$ dove Φ' è la funzione di densità della normale che sappiamo scrivere con $\mu = 0$ e $\sigma^2 = 1$.

Quindi $\Phi'(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \rightarrow \frac{d}{dy}\left(\Phi\left(\sqrt{y}\right)\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y}$

In conclusione:

[3] $\Phi(-z) = 1 - \Phi(z)$, grazie alla proprietà di simmetria della distribuzione normale

$$
\begin{aligned}
F_y(y) &= 2\Phi(\sqrt{y}) - 1 \\
f_y(y) &= F'_y(y) \\
&= 2\Phi'(\sqrt{y})\frac{1}{2\sqrt{y}} \\
&= 2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y}\frac{1}{2\sqrt{y}} \\
&= \frac{1}{\sqrt{2\pi}}y^{\frac{-1}{2}}e^{-\frac{1}{2}y}1_{[0,+\infty)}(y) \\
&= \chi^2(1) \text{ Chi-square distribution}
\end{aligned}
$$

In teoria delle probabilità una distribuzione χ^2 (chi quadrato o chi quadro) è una distribuzione di probabilità che descrive la somma dei quadrati di alcune variabili aleatorie indipendenti aventi distribuzione normale standard.

L'esercizio ha *dimostrato* come la VA normale standard elevata al quadrato risulti in un'altra VA che ha grado di libertà 1

7.5.1 Anticipazioni sulla χ^2 e inferenza statistica

In statistica χ^2 viene particolarmente utilizzata per l'omonimo *test di verifica d'ipotesi* (test χ^2), la quale si utilizza per verificare la bontà di un ipotesi e gioca un ruolo importante nell'*inferenza statistica.*

$$Z \Longrightarrow \chi^2 \Longrightarrow \text{test di verifica d'ipotesi} \Longrightarrow \text{inferenza statistica}$$

L'inferenza statistica è il procedimento per cui si inducono le caratteristiche di una popolazione dall'osservazione di una parte di essa, detta campione, selezionata solitamente mediante un esperimento casuale (aleatorio). Da un punto di vista *filosofico*, si tratta di tecniche matematiche per quantificare il processo di apprendimento tramite l'esperienza.

Esempio: dtata un'urna con composizione nota: 6 palloni bianchi, 4 palloni rossi, utilizzando le regole del calcolo delle probabilità possiamo dedurre che se estraiamo un pallone a caso dall'urna, la probabilità che esso sia rosso è 0,4. Si ha invece un problema di inferenza statistica se abbiamo un'urna di cui non conosciamo la composizione, estraiamo n palloni e ne osserviamo il colore e, a partire da questo, vogliamo indurre la composizione dell'urna.

7.5.2 Distribuzioni gamma

Le distribuzioni gamma sono frequentemente utilizzati come modelli di probabilità per i tempi di attesi. Infatti sono una generalizzazione del modello esponenziale.

$f(x) = \frac{1}{\Gamma(\alpha)} \frac{1}{\theta^\alpha} x^{\alpha-1} e^{-\frac{1}{\theta}\alpha} 1 \rightarrow X \sim \Gamma(\alpha, \theta)$. Con $\alpha = 1$ ritorno all'esponenziale.

7.6 Esercizio

$$f(x, \theta) = \left\{ \begin{array}{cc} \frac{1}{30^\theta}(x-30)^{\theta-1} & 30 < x < 60 \\ 0 & \text{altrove} \end{array} \right\}$$

$P(X < 35) = \int_{30}^{35} f_X$

Se registro i tempi di viaggio per 20 giorni consecutivi... Ho 10 giorni in cui alcune volte di metterò 35min, altri più o meno. Utilizzerò la binomiale per calcolare la aleatorietà.

$X \sim \text{Bin}\left(n = 100, p = \frac{1}{6^\theta}\right)$; $P(x \geq 40)$ =come approssimazione della binomiale conosciamo la poisson o la normale. La prima la possiamo utilizzare con n grande e p piccolo (un sesto alla theta è troppo grande).

La normale è un'approssimazione che deriva con il teorema centrale del limite (valido con p intorno a un mezzo); l'approssimazione è valida con n grande; $\mu = np$ e $\sigma^2 = np(1-p)$. Quindi

$$X \sim \underbrace{N\left(\mu = \frac{100}{6^\theta}, \sigma^2 = \frac{100}{6^\theta}\left(1 - \frac{1}{6^\theta}\right)\right)}_{Y}$$

e risolvo il calcolo della probabilità standardizzando ed utilizzando la tabella.

7.6.1 Correzione di continuità

Approssimiamo la X discreta, con quindi valori distribuiti lungo tutto l'asse. La probabilità del 38 si spalmerà pertanto tra 37.5 a 38.5. Quindi $P(X > 40) = P(X > 39.5) = P\left(Z > \frac{39.5 - \frac{100}{6^\theta}}{\sqrt{\frac{100}{6^\theta}\frac{6^\theta - 1}{6^\theta}}}\right) = P(Z > \ ) = 1 - \Phi\left(\frac{3.95\ 6^{\theta-10}}{\sqrt{6^\theta - 1}}\right)$

7.7 Esercizio

$X_1, X_2, ..., X_n \sim \text{Poisson}(\lambda = 2)$; Voglio calcolare $P\left(\bar{X} \leq 1.9\right)$ con n che va da 1 a 100

$n = 1 \rightarrow P(X_1 \leq 1.9) = e^{-2} + 2e^{-2} = 3e^{-2}$

$n = 2 \rightarrow P\left(\frac{X_1+X_2}{2} \leq 1.9\right) = P(X_1 + X_2 \leq 3.8)$, ma $X_1 + X_2$ è la poissoniana di parametro 4

$n = 100 \rightarrow P(X_1 + ... + X_n \leq 190) \sim N(200, 200) = Y$

$P(Y \leq 190.5) \rightarrow P\left(Z - \frac{190.5-200}{\sqrt{200}}\right)$

7.8 Riepilogo

Scopo degli esercizi è stato osservare quali sono gli strumenti a disposizione per osservare una VA Y (non nota) partendo dalla sua relazione con una VA X (nota), di cui si conosce la funzione densità di probabilità. Conoscere una VA significa conoscere o la sua **funzione densità di probabilità** o la sua **funzione di ripartizione**.

Nel procedere con i calcoli occorre fare attenzione al dominio di entrambe le variabili aleatorie.

Utile è sfruttare la correlazione che c'è tra le due funzioni che descrivono una VA: la funzione densità di probabilità è la derivata della funzione di partizione. Quest'ultima inoltre ha la proprietà di partire da 0 ed arrivare a 1.

8 Funzioni Γ

Continuiamo con la famiglia di distribuzione gamma per fare inferenza sulla varianza di un modello gaussiano, sia che la media sia nota che non.

Consideriamo $X_1, ..., X_n$ iid $\sim \xi(\beta)$, $\beta > 0$

$M_x(t) = E\left(e^{tX}\right)$; $M_{\sum X_j}(t) = [M_x(t)]^n = \left(\frac{1}{1-\beta t}\right)^n$ se $t < \frac{1}{\beta}$. Essendo $\frac{1}{1-\beta t}$ la funzione generatrice di momenti nel caso dell'esponenziale (più in generale $\left(\frac{1}{1-\beta t}\right)^n$ questa è la funzione generatrice dei momenti di $\Gamma(n, \beta)$).

Due VA hanno stessa funzione di ripartizione e densità sse $M_x(t) = M_y(t)$, infatti media e varianza (momenti di primo e secondo ordine) coincidono.

Distribuzione di Erlang n,β. Modello per l'ennesimo arrivo. Quando mi arriva il guasto ed aggiorno l'orologio. I tempi sono indipendenti.

8.1 Analisi della funzione Γ

Data la $\Gamma(\alpha, \beta)$ con $\beta > 0$ e $\alpha > 0$, la sua funzione di densità è $f(x, \alpha, \beta) = \frac{\left(\frac{1}{\beta}\right)^{\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)} 1_{(0,+\infty)}(x)$

Se $\alpha = 1 \Rightarrow \Gamma(1, \beta) = \xi(\beta)$

$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$ è la definizione di una funzione Γ ad una variabile reale, ed è una generalizzazione del fattoriale.

Nel caso di un numero naturale n $\Gamma(n) = (n-1)!$ con $n \geq 1$, con proprietà $\Gamma(0) = 1$.

Condivide proprietà del fattoriale $n! = n(n-1)!$ allora $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.

Casi particolari della gamma:

- $\Gamma\left(\frac{1}{2}, 2\right) =$ è la densità di una normale al quadrato $= \chi^2_1$ dove il pedice indica i gradi di libertà e l'abbiamo ottenuta con $X = Z^2$, dove $Z \sim N(0, 1)$
- $\Gamma\left(\frac{n}{2}, 2\right) = \chi^2_n$ n gradi di libertà.

8.2 Analizziamo χ^2_n

$X_1, ..., X_n$ iid $\sim N(\mu, \sigma^2)$ standardizzo ed ottengo $\frac{X_1-\mu}{\sigma}, ..., \frac{X_n-\mu}{\sigma}$ iid $\sim N(0, 1)$. Di ciascuna adesso ne prendo il quadrato $\left(\frac{X_1-\mu}{\sigma}\right)^2, ..., \left(\frac{X_n-\mu}{\sigma}\right)^2$ iid $\sim \chi^2_1\left(\Gamma\left(\frac{1}{2}, 2\right)\right)$. Così come una distribuzione normale N accetta come parametri μ e σ^2 e la distribuzione esponenziale accetta come parametri α e β, la funzione χ^2 accetta come parametro una funzione di densità $\Gamma(\alpha, \beta)$.

Adesso

$\sum_{j=1}^n \left(\frac{X_j-\mu}{\sigma}\right)^2 = \sum_{j=1}^n (\chi^2_1)_j = W$ somma di n $\chi^2_1\left(\Gamma\left(\frac{1}{2}, 2\right)\right)$ indipendenti. Calcolo la funzione generatrice dei momenti di W.

$$M_W(t) = \left[\underbrace{\left(\frac{1}{1-2t}\right)^{\frac{1}{2}}}_{M_\Gamma\left(\frac{1}{2},2\right)}\right]^n = \left[\frac{1}{1-2t}\right]^{\frac{n}{2}} = \Gamma\left(\frac{n}{2}, 2\right)$$

Il significato dei gradi di liberta di un χ^2 indice il numero di χ^2 indipendenti sommate per ottenere la W.

8.3 Trasformazioni su Γ

Torniamo alle funzioni gamma e cambiamo la scala o le unità di misura $X \sim \Gamma(\alpha, \beta) \to cX$, con $c > 0$.

$M_{cX}(t) = E\left(e^{tcX}\right) = E\left(e^{(tc)X}\right) = \left(\frac{1}{1-\beta ct}\right)^{\alpha}$ se $ct < \frac{1}{\beta}$ è fgm di$\Gamma(\alpha, c\beta)$. Questo modello va bene tipicamente per le misurazioni.

A partire da X, Y indipendenti $X \sim \Gamma(\alpha_1, \beta)$ e $Y \sim \Gamma(\alpha_2, \beta)$. Qual'è la funzione di distribuzione congiunta $X + Y$?

$X+Y$; $M_{X+Y}(t) = E\left(e^{t(X+Y)}\right) = E\left(e^{tX}\right) E\left(e^{tY}\right) = \left(\frac{1}{1-\beta t}\right)^{\alpha_1}\left(\frac{1}{1-\beta t}\right)^{\alpha_2} = \left(\frac{1}{1-\beta t}\right)^{\alpha_1+\alpha_2}$ è fgm di $\Gamma(\alpha_1 + \alpha_2, \beta)$. β è parametro di scala e pertanto non può cambiare, altrimenti sommo mele con pere. La funzione generatrice della somma di due variabili Γ è la funzione generatrice della funzione Γ con $\alpha = \alpha_1 + \alpha_2$ e $\beta = \beta$.

$--$

La mia situazione è questa: con X ed Y indipendenti, $X \sim \Gamma(\alpha_1, \beta)$ mentre $X + Y \sim \Gamma(\delta, \beta)$ con $\delta > \alpha_1$ (perchè c'è una α_2). Riesco a ricavare Y? (algebra di distribuzioni). In questo caso è possibile risolvere il problema con la funzione generatrice di momenti.

$\left(\frac{1}{1-\beta t}\right)^{\delta} = \left(\frac{1}{1-\beta t}\right)^{\alpha_1} \cdot M_y(t) \Rightarrow M_y(t) = \left(\frac{1}{1-\beta t}\right)^{\delta-\alpha_1}$

$--$

$W = Y_1 + Y_2$ se so che $Y_1 \sim \chi^2_1$ e $W \sim \chi^2_n$ e Y_1, Y_2 indipendenti. In questa situazione deduco che $Y_2 \sim \Gamma\left(\frac{n-1}{2}, 2\right) = \chi^2_{n-1}$. La Y_2 deve essere di grado $n-1$ poichè la somma dei gradi di Y_1 e Y_2 deve essere n.

Qual'è la media e la varianza di una gamma? O calcoliamo gli integrali oppure continuamo ad usare la generatrice di momenti.

$M_X(t) \to \frac{d^k M_x(t)}{d^k t}_{t=0} = E\left(X^k\right)$. Applicata alla gamma $\Rightarrow E(\Gamma(\alpha,\beta)) = \alpha\beta$; $Var(\Gamma(\alpha,\beta)) = \alpha\beta^2$

$E(X) = \frac{dM_x(t)}{dt}_{t=0} = \frac{d}{dt}\left(\frac{1}{1-\beta t}\right)^{\alpha} = \alpha\left(\frac{1}{1-\beta t}\right)^{\alpha-1}\left(-\frac{1-\beta t}{(1-\beta t)^2}\right)(-\beta) = \alpha\beta$

La funzione χ^2_n è definita dalla funzione $\Gamma\left(\frac{n}{2},2\right)$, pertanto calcolare $E(\chi^2) = E\left(\Gamma\left(\frac{n}{2},2\right)\right)$.

$E(\chi^2_n) = \frac{n}{2}2 = n$; $Var(\chi^2_n) = \frac{n}{2}2^2 = 2n$.

Se n piccolo ..., se n grande posso usare un'appossimazione. Se sarà necessario calcolare i quantili di un χ^2_n posso utilizzare il teorema centrale del limite $F_{\chi^2_n} \sim F_{N(n,2n)} \to q_{\chi^2_n}(\lambda) \sim z_\lambda\sqrt{2n} - n$.

9 Inferenza sulle normali.

$X_1, ..., X_n$ iid $\sigma^2 = Var(X_1)$ (non c'è nessuna ipotesi di distribuzione). Stimiamo σ^2 con $S^2 = \frac{\sum_{j=1}^n\left(X_j-\bar{X}\right)^2}{n-1}$ è uno stimatore non distorto.

9.1 Dimostrazione $E\left(S^2\right) = \sigma^2$

$$\begin{aligned}
\sum_j\left(X_j-\bar{X}\right)^2 &= \sum_j\left[(X_j-\mu)-\left(\bar{X}-\mu\right)\right]^2 \\
&= \sum_j\left[(X_j-\mu)^2+\left(\bar{X}-\mu\right)^2-2(X_j-\mu)\left(\bar{X}-\mu\right)\right] \\
&= \sum_j(Xj-\mu)^2 \text{ ...sostituisco con sotto} \\
&= \sum(X_j-\mu)^2 - n\left(\bar{X}-\mu\right)^2
\end{aligned}$$

con $\sum(X_j-\mu) = \sum X_j - n\mu = n\bar{X} - n\mu = n\left(\bar{X}-\mu\right)$

$$\begin{aligned}
E\left[\sum_j\left(X_j-\bar{X}\right)^2\right] &= E\left(\sum(X_j-\mu)^2\right) - nE\left(\bar{X}-\mu\right)^2 \\
&= \sum_j E(X_j-\mu)^2 - nVar\left(\bar{X}\right) \\
&= \sum Var(X_j) - n\frac{\sigma^2}{n} \\
&= \sum\sigma^2 - \sigma^2 \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

$E(S^2) = \frac{1}{n-1}E\left(\sum\left(X_j-\bar{X}\right)^2\right) = \sigma^2$ è indistorto

Questo stimatore è consistente in media quadratica? Sì

$Var\left(S^2\right) \sim \frac{E\left(X^4\right)c}{n-1}$ solo accenno (con c costante)

9.2 $X_1, ..., X_n$ iid $N\left(\mu, \sigma^2\right)$ con media e varianza incognite $\Rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Vediamo la distribuzione di S^2 e mi serve per dare una stima intervallare della varianza.

Ripartiamo con un set di VA **normale**. La prima proprietà in questo caso dello stimatore è per niente intuitiva. Mi interessano sapere media e varianza. La legge congiunta di $\bar{X}, S^2$, se il campione è gaussiano media e varianza campionaria sono indipendenti. La dimostrazione prevede la generatrice dei momenti o i vettori gaussiani (?).

Parentesi culturale: l'indipendenza non è un proprietà deterministica.

La proprietà rilevante è $\bar{X}, S^2$ sono indipendenti nel senso della probabilità. Abbiamo tutti gli elementi per la distribuzione di S^2.

Parto da una trasformazione di $S^2 \to \frac{S^2(n-1)}{\sigma^2}$ questa non è una statistica perchè dipende da un parametro incognito.

$\frac{S^2(n-1)}{\sigma^2} = \sum_{j=1}^{n} \left(\frac{X_j - \bar{X}}{\sigma}\right)^2$. Vediamo la distribuzione di questa identità.

Riprendo

$$\sum_j \left(X_j - \bar{X}\right)^2 = \overbrace{\sum_j \left(X_j - \mu\right)^2 - n\left(\bar{X} - \mu\right)^2}^{\text{vedere dimostrazione sopra}}$$

$$\begin{array}{ccc} \sum_j \left(X_j - \bar{X}\right)^2 & = & \sum \left(X_j - \mu\right)^2 - n\left(\bar{X} - \mu\right)^2 \\ \frac{\sum_j \left(X_j - \bar{X}\right)^2}{\sigma^2} & = & \frac{\sum (X_j - \mu)^2 - n\left(\bar{X} - \mu\right)^2}{\sigma^2} \\ Y_2 & = & W - Y_1 \end{array}$$

$Y_1 = n\left(\frac{\bar{X} - \mu}{\sigma}\right)^2$; $W = \sum_j \left(\frac{X_j - \mu}{\sigma}\right)^2$; $Y_2 = \frac{(n-1)S^2}{\sigma^2}$

$W \sim \chi^2_n$; Y_1 è il quadrato della media campionaria standardizzata $Y_1 \sim \chi^2_n$.

Y_1, Y_2 sono dipendenti o indipendenti? Indipendenti per quella proprietà peculiare del modello gaussiamo che abbiamo riportato (quale?).

Tutto questo per dire $\Rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Se ho i valori di 19 misurazioni e la media campionaria, posso ricavare anche la misurazione della ventesima. Per questo motivo c'è $n-1$ e non n.

Domanda, se sappiamo $\Rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ sappiamo ricavare la distribuzione di S^2? La differenza è una costante che è strettamente positiva. Se prendo questa particolare chi^2 e moltiplico per $\frac{\sigma^2}{n-1}$ ottengo la S^2.

$S^2 = \left(\frac{\sigma^2}{n-1}\right) \chi^2_{n-1}$ allora $S^2 \sim \Gamma(\cdot, \cdot)$

9.3 Intervallo di confidenza $P(T_1 < \sigma^2 < T_2) = \gamma$

Voglio determinare una forbice di valori in cui sono sicuro al TOT% che la varianza sia all'interno di questa forbice.

Obiettivo formalizzato: $T_1(x_1, ..., x_n) < T_2(x_1, ..., x_n)$ tale che

$$P\left(T_1 < \sigma^2 < T_2\right) = \gamma$$

con $\gamma \in (0, 1)$, γ prossimo ad 1.

Vedere FIG 0.

Primo passo determina q_1, q_2 tale che $P\left(q_1 < \frac{(n-1)S^2}{\sigma^2} < q_2\right) = \gamma$

Il gamma visivamente è una probabilità, quindi l'area sottesa tra q_1 e q_2 nel grafico della funzione di distribuzione. Il problema è che la distribuzione è asimmetrica. Voglio che l'area a sinistra di q_0 sia uguale all'area a destra di q_1 che è uguale a $= \frac{1-\gamma}{2}$

$$1 - P(X < q_1) - P(X > q_2) = \gamma$$

Con questa richiesta so quanto vale q_1.

$$P\left(\frac{(n-1)S^2}{\sigma^2} \leq q_1\right) = \frac{1-\gamma}{2}$$

$$q_1 = \underbrace{\chi^2_{n-1;\left(\frac{1-\gamma}{2}\right)}}_{\text{quantile di } \chi^2_{n-1} \text{ tale che l'area è } \frac{1-\gamma}{2}}$$

$$q_2 = \chi^2_{n-1;\left(\frac{1-\gamma}{2}+\gamma\right)} = \chi^2_{n-1;\left(\frac{1+\gamma}{2}\right)}$$

Riscrivo l'evento $P\left(q_1 < \frac{(n-1)S^2}{\sigma^2} < q_2\right)$ in funzione di σ^2

$$P\left(q_1 < \frac{(n-1)\,S^2}{\sigma^2} < q_2\right) = \gamma = P\left(\frac{(n-1)\,S^2}{\chi^2_{n-1;\left(\frac{1+\gamma}{2}\right)}} < \sigma^2 < \frac{(n-1)\,S^2}{\chi^2_{n-1;\left(\frac{1-\gamma}{2}\right)}}\right)$$

Perchè considero il rapporto piuttosto che la differenza? Perchè la precisione la leggo in termini di rapporto.

Cos'è aleatorio in quest'evento? σ^2 è un parametro, una costante che non conosco. È aleatorio l'intervallo di $P\left(\frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1+\gamma}{2}\right)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1-\gamma}{2}\right)}\right)$. Se arrivano i dati, calcoliamo il valore di S^2 gamma lo scegliamo noi, allora la realizzazione si chiama intervallo di confidenza γ.

Se invece sono interessato soltanto alla frontiera dal basso? Oppure voglio tutelarmi da un massimo? Non ti posso dare una probabilità certa. Siano interessati ad un intervallo di confidenza unilatero.

Sono sempre statistiche questo lower|upper-bound.

Cerco una statistica T_{Lower} tale che $P\left(T_l < \sigma^2\right)$ confine inferiore di σ^2. $T_l = \frac{(n-1)S^2}{\chi^2_{n-1}(\gamma)} = \gamma$.

$P\left(\frac{S^2(n-1)}{\chi^2_{n-1}(\gamma)} < \sigma^2\right) = P\left(\frac{S^2(n-1)}{\sigma^2} < \chi^2_{n-1}\left(\gamma\right)\right)$

T_U tale che $\gamma = P\left(T_u > \sigma^2\right) \Rightarrow T_u = \frac{(n-1)S^2}{\chi^2_{n-1}(1-\gamma)}$

9.4 $X_1, ..., X_n$ iid $N\left(\mu, \sigma^2\right)$ con $\mu = \mu_0$ nota e σ^2 incognita

Che stimatore utilizzo? $\frac{\sum(X_j-\mu_0)^2}{n} = S_0^2$

$n\frac{S_0^2}{\sigma^2} \sim \chi_n^2$ Se questa è la situazione procedo con lo stesso procedimento di oggi, ma cambia

μ nota	μ incognita
S_0^2	S^2
χ_n^2	χ_{n-1}^2
n	$n-1$

Obiettivo: trovare un intervallo di validità quando la varianza è incognita.

Punto di partenza $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}}$. Il problema è che questa quantità non è più una normale 0,1 per piccoli campioni. Rimane approssimativamente normale per grandi campioni. Per piccoli campioni, qual'è la distribuzione?

$\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$ (**t di Student** con n-1). Vedi FIG 1 (Fabio)

Se $n = 1$ la densità di probabilità $\frac{1}{(1+x^2)}\frac{1}{\pi}$. Una distribuzione a campana di cui abbiamo le tavole. Se n grande $n \geq 60$. Converge alla densità normale 0,1. $t_{n-1} \sim N(0,1)$

L'intervallo di confidenza che viene fuori ha la stessa forma di quella con varianza nota (utilizzo quantili della normale), la differenza utilizzo la t di student per i quantili. Dove prima avevo σ^2 adesso ho S^2.

σ^2 nota	σ^2 incognita
σ^2	S^2
z_λ	$t_{n-1}(\lambda)$

10 Esercizi

10.1 Esercizi sulla Normale

$X \sim N(8,9)$

$P(X \leq 13) = P\left(Z \leq \frac{13-8}{\sqrt{9}}\right) = P\left(Z \leq \frac{5}{3}\right) \sim \Phi(1,67) = 0.9525$

$P(X < 7) = P\left(Z < \frac{7-8}{\sqrt{9}}\right) = \Phi\left(-\frac{1}{3}\right) = 1 - \Phi\left(\frac{1}{3}\right) = 1 - 0.6293$

Determinare il valore da dare a γ affinchè $P(X < \gamma) = 0.95$

Indichiamo il quantile con la lettera z (minuscola) con pedice l'area alla sua sinistra

$0.95 = P\left(Z < \frac{\gamma-8}{\sqrt{9}}\right)$; $\frac{\gamma-8}{3} = z_{0.95} \rightarrow \gamma = 8 - 3z_{0.95}$

Prendo allora la tabella della normale e leggo nelle celle cetrali 0.95 (il valore più approssimato), le coordinate di questo valore mi danno il quantile che sto cercando.

$1.64 \Rightarrow 0.9495$ e $1.65 \Rightarrow 0.95053$

La distribuzione di t di Student è come una normale ma con le code più grandi. Con n che tende all'infinito la t di Student tende alla normale.

Determinare

$$\begin{aligned} P(X > \gamma) &= 0.9 \\ &= P\left(Z > \frac{\gamma - 8}{3}\right) \\ 1 - P\left(Z > \frac{\gamma - 8}{3}\right) &= 1 - 0.9 \\ P\left(Z < \underbrace{\frac{\gamma - 8}{3}}_{z_{0.1}}\right) &= 0.1 \end{aligned}$$

$z_{0.1} = -z_{0.9} = -1.282 \rightarrow \gamma = 8 - 3 \cdot 1.282$

10.2 Esercizio dell'eserciziario 1.3.4

Variabile aleatoria X con $f_x = (x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} 1_{(0,+\infty)}(x)$

Se X VA qualsiasi con $E(X) = m$ e $Var(X) = v$, costruiamo $\bar{X}$ e quali sono i legami tra le due medie e le due varianze? $E\left(\bar{X}\right) = m$ e $Var\left(\bar{X}\right) = \frac{v}{n}$.

Qual'è la distribuzione della media campionaria di quest'esponenziale?

Ripasso: $\xi(\theta) \sim \Gamma(1,\theta)$. Se $Y \sim (\alpha,\beta) \to f(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-\frac{1}{\beta}y} 1_{(0,+\infty)}(y)$

$\Gamma(\alpha_1,\beta) + \Gamma(\alpha_2,\beta) = \Gamma(\alpha_1+\alpha_2,\beta)$

$\bar{X} = \frac{1}{n} \underbrace{\sum_{i}^{n} X_i}_{\Gamma(n,\theta)} = \frac{1}{n}\Gamma(n,\theta)$

$\Gamma(n+1) = n!$; $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$Y = \sum X_i \to f_Y(y) = \frac{y^{n-1}}{\Gamma(n)\beta^{\alpha}} e^{-\frac{1}{\beta}y} 1_{(0,+\infty)}(y)$

$\bar{X} = \frac{1}{n}Y \to Y = n\bar{X} \to f_{\bar{X}}(x) = f_x(nx) \cdot n$ (ricordando $f_y(y) = f_x(h(y)) \cdot h'(y)$)

quindi: $f_{\bar{X}}(x) = f_x(nx) \cdot n = \frac{(nx)^{n-1}}{\Gamma(n)\theta^n} e^{-\frac{nx}{\theta}} 1_{(0,+\infty)}(x) \underbrace{=}_{\frac{n}{\theta}=\frac{1}{\frac{\theta}{n}}} \frac{x^{n-1}}{\Gamma(n)\left(\frac{\theta}{n}\right)^n} e^{-\frac{x}{\frac{\theta}{n}}} 1_{(0,+\infty)}(x)$

$--$

Definire la densità $Y = \frac{2n\bar{X}}{\theta} = \frac{2}{\theta}\sum X_i \sim \Gamma(n,2) \sim \chi^2_{2n}$

Fissiamo: $\alpha = 0.05$, $n = 3$, $\theta = 2$, determiniamo k tale che $P\left(\bar{X} < k\right) = \alpha$

Per l'esponenziale il gioco consiste nell'arrivare alla χ^2, perchè dobbiamo arrivare a distribuzioni note di cui abbiamo le tavole. Quindi l'argomento della probabilità lo devo trasformare affichè si arrivi alla $\chi^2 \to \frac{2n}{\theta}\bar{X} \sim \chi^2_{2n}$.

$P\left(\frac{2\cdot3}{2}\bar{X} < \frac{2\cdot3}{2}k\right) = 0.05 \to P(\chi^2_6 < 3k) = 0.05$

$3k = \chi^2_{6;0.05} = 1.64$

10.3 Intervalli di confidenza

10.3.1 Esercizio 2.1.1

Abbiamo due statistici che lavorano sullo stesso campione, ma il primo costruisce un intervallo al 99%, il secondo al 90%. Qual'è quello più largo? Quello del 99%. La confidenza è tipo una probabilità che il nostro intervallo sia corretto. Se voglio aumentare la precisione diminuisco la confidenza e viceversa.

Una ditta produce ?, si trovano n punte dello stesso diametro producendo n fori: $X_1, ..., X_n$ sono i diametri ed ogni X_i sono distribuiti nel modo con μ incognito e σ^2nota. Vogliamo un intervallo di confidenza per la μ di livello γ. $X_i \sim N(\mu, \sigma^2)$

Abbiamo visto intervalli di confidenza per la media (varianza nota e non nota), stesso ragionamento ma inverso per la varianza.

La media campionaria è distribuita come una normale $\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n}\right)$.

$$P\left(\underbrace{-z_{\frac{1+\gamma}{2}}}_{\text{quantile sx}} < \underbrace{\frac{\bar{Y}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}}}_{Z} < \underbrace{z_{\frac{1+\gamma}{2}}}_{\text{quantile dx}} \right) = \gamma$$

Parentesi sui quantili: considerando una funzione di densità normale standard γ rappresenta un'area suddivisa simmetricamente dall'asse delle ordinate, a sinistra e a destra di tale area ci saranno pertanto due aree di uguali dimensione pari a $\frac{1-\gamma}{2}$; leggendo da sinitra verso destra le aree incontriamo $\frac{1-\gamma}{2}; \gamma; \frac{1-\gamma}{2}$. I due quantili sono $z_{\frac{1-\gamma}{2}}$ e $z_{\frac{1-\gamma}{2}+\gamma} = z_{\frac{1+\gamma}{2}}$. Notare che $-z_{\frac{1+\gamma}{2}} = z_{\frac{1-\gamma}{2}}$

Scopo del gioco è trovare l'intervallo di confidenza. I due quantili sono simmetrici e sono tali da avere area γ, quindi sono valori tale per cui l'area è $\frac{1-\gamma}{2}$.

$\bar{Y} - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma_0^2}{n}} < \mu < \bar{Y} + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma_0^2}{n}}$. Perchè abbiamo scelti i quantili simmetrici? Avremmo potuto prendere qualsiasi coppia di valori, ma c'è una bilancia tra confidenza e precizione. Se prendo i due quantili simmetrici l'intervallo che ottengo è il più stretto di tutti.

10.3.2 Esercizio

$n = 100$; $\bar{Y} = 5$ mm; $\sigma_0^2 = 10^{-2}$mm; $\gamma = 95\%$

Sapendo $\sigma_0^2 = 10^{-2}$mm, quanto deve essere grande n affinchè stima di μ sia precisa entro 10^{-2}mm? L'intervallo $\bar{Y} - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma_0^2}{n}} < \mu < \bar{Y} + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma_0^2}{n}}$ deve essere largo al massimo 10^{-2}mm. Scopo del gioco è determinare n tale che $2z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma_0^2}{n}} < 10^{-2} (= l) \rightarrow n > \sigma_0^2 \left(\frac{2z_{\frac{1+\gamma}{2}}}{l}\right)^2$ Sostituendo i numeri abbiamo che $n = 385$ (Nota a margine: a me risulta 1536)

10.3.3 Esercizio

$\mu \in (4.50, 4.54)$

1. Quanto vale la media campionaria $\bar{Y}$? immediato, perchè è $4.52 = \frac{4.50+4.54}{2}$
2. Quante sono il numero delle osservazioni? $z_{\frac{1+\gamma}{2}} \sqrt{\frac{\sigma_0^2}{n}} = 0.02 \rightarrow n = \sigma_0^2 \left(\frac{z_{\frac{1+\gamma}{2}}}{0.02} \right)^2 = 97$

10.3.4 Esercizio 2.1.1

$X \sim N(\mu, \sigma^2)$, con 12 osservazioni che danno come risultato $\bar{X} = 21.18$ e $S^2 = 61.15$. Voglio calcolare un intervallo di confidenza (IC) del 95% per μ.

Varianza incognita, quindi

$$\mu = \left[\bar{X} \pm t_{n-1} \left(\tfrac{1+\gamma}{2} \right) \sqrt{\tfrac{S^2}{n}} \right] \rightarrow \mu \in 21.18 \pm t \underbrace{0.975}_{1-\frac{(1-0.95)}{2}=0.95+0.025} \sqrt{\tfrac{61.51}{12}}$$

IC al 90% per $\sigma^2 \rightarrow \sigma^2 \in \left[\frac{11 \cdot 61.51}{19.675}, \frac{11 \cdot 61.51}{4.575} \right]$, ottenuto come $\sigma^2 \in \left[\frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1+\gamma}{2}\right)}, \frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1-\gamma}{2}\right)} \right]$. Il grafico della χ^2 è definita solo per x positiva. [disegno] I due quantili ora sono asimmetrici $\chi^2_{11}(0.95)$ e $\chi^2_{11}(0.05)$

$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \rightarrow q_1 < \frac{(n-1)S^2}{\sigma^2} < q_2 \rightarrow \left[\frac{(n-1)S^2}{q_2}, \frac{(n-1)S^2}{q_1} \right]$, attenzione al corretto posizionamento dei quantili al denominatore. Perchè l'intervallo ha come estremo sinistro un valore minore rispetto all'estermo destro.

10.3.5 Fare a casa esercizio 2.1.2

11 Riepilogo

X va $\rightarrow M_x(t) = E\left[e^{tX}\right]$

$X_1, ..., X_n \rightarrow M_{\sum_j^n X_j}(t) = [M_x(t)]^n$

$X \sim \xi(\beta) \rightarrow M_X(t) = \frac{1}{1-\beta t}$

$X \sim \Gamma(n, \beta) \rightarrow M_x(t) = \left(\frac{1}{1-\beta t}\right)^n$

$X \sim \Gamma\left(\frac{1}{2}, 2\right) \rightarrow X \sim \chi^2_1 \rightarrow M_x(t) = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}$

$X \sim \Gamma\left(\frac{n}{2}, 2\right) \rightarrow X \sim \chi^2_n \rightarrow M_x(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$

$\frac{d^k M_x(t)}{d^k t} = E\left(X^k\right) \rightarrow \begin{cases} k = 1 & \text{media} \\ k = 2 & \text{varianza} \end{cases}$

$$\Gamma(\alpha, \beta) = \begin{cases} \alpha = 1 & \xi(\beta) \\ \alpha = \frac{1}{2}, \beta = 2 & \chi^2_1 \\ \alpha = \frac{n}{2}, \beta = 2 & \chi^2_n \\ \Gamma(\alpha + 1) = \alpha\Gamma(\alpha) & \text{fattoriale} \end{cases}$$

11.1 Inferenza sulle normali: μ, σ^2, S^2, χ^2_{n-1} e t_{stud}

Il campione di riferimento con cui abbiamo a che fare è il seguente:

$$X_1, ..., X_n \text{ iid } \sim N(\cdot, \cdot)$$

di cui posso conoscere o meno la media e varianza. Di seguito saranno riportate le procedure da seguire quando uno o due di questi valori non sono conosciuti (ovviamente se li sappiamo entrambi abbiamo tutte le informazioni necessarie per procedere con i calcoli statistici).

11.1.1 S^2, σ^2 e χ^2_{n-1} con μ non nota e $n \rightarrow \infty$

Per definizione

$$\sigma^2 = Var[X]$$

Per determinare la varianza ho bisogno della media, che se non è nota è possibile ricavare con lo stimatore non distorto $\bar{X}$. Lo stimatore tende alla non distorsione se il numero di campioni è elevato ($n \rightarrow +\infty$). Su questa condizione posso utilizzare la varianza campionaria per stimare la varianza

$$S^2 = \frac{\sum_{j=1}^n \left(X_j - \bar{X}\right)^2}{n-1}$$

Abbiamo dimostrato inoltre che $E(S^2) = \sigma^2$. Lo stimatore della varianza è indistorto ed è consistente in media quadratica.

A partire dalla definizione di varianza campionaria abbiamo eseguito la seguente trasformazione:

$$\frac{S^2(n-1)}{\sigma^2} = \sum_{j=1}^{n} \left(\frac{X_j - \bar{X}}{\sigma} \right)^2 \sim \chi^2_{n-1}$$

Dato che χ^2 è una distribuzione notevole, posso considerarla come punto di partenza per ricavare S^2

$$S^2 = \left(\frac{\sigma^2}{n-1} \right) \chi^2_{n-1} \rightarrow \underbrace{S^2 \sim \Gamma(\cdot, \cdot)}_{\chi^2 \subseteq \Gamma}$$

(Non detto in classe) Dato che la varianza campionaria la posso stimare a partire dai valori campionati e dalla media, da quest'ultima equazione posso osservare che la varianza reale $\sigma^2 = f\left(S^2, \chi^2_{n-1}\right)$ che sarà tanto più corretta quanto più n è grande.

11.1.2 Cosa succede se n piccolo o X_j non noti? t_{stud}

La t di Student è uno strumento che ci viene in aiuto quando il numero di campioni a disposizione è piccolo (quindi la media campionaria non è più una normale), ma comunque conosciamo i singoli risultati per poter estrarre la media campionaria.

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

11.1.3 S^2, σ^2 e χ^2_n con $\mu = \mu_0$ nota

Il ragionamento ed i procedimenti sono gli stessi con media non nota, con la differenza che qualche definizione è leggermente diversa perchè utilizza l'informazione della media nota.

Varianza campionaria

$$S_0^2 = \frac{\sum_{j=1}^{n} (X_j - \mu_0)^2}{n}$$

$$n\frac{S_0^2}{\sigma^2} = \sum_{j=1}^{n}\left(\frac{X_j - \mu_0}{\sigma}\right)^2 \sim \chi_n^2$$

$$S_0^2 = \frac{\sigma^2}{n}\chi_n^2$$

12 Stima puntuale

12.1 Introduzione

Mentre quando si fa della probabilità è normale supporre che le distribuzioni in gioco siano completamente note, in statistica è vero il contrario, e il problema centrale è quello di dire qualcosa (ovvero *fare dell'inferenza*) sui parametri sconosciuti, usando i dati osservati. Il problema è noto come **stima parametrica**. Per **stima puntuale** si intende l'utilizzo di uno stimatore che porti a fornire un singolo valore come stima del parametro.

$$Stimatore \rightarrow \begin{cases} \text{puntuale} & MLE \\ \text{non puntuale} & IC \end{cases}$$

Con gli stimatori non puntuali siamo in grado di ottenere non un singolo punto, come stima del parametro θ, ma un intervallo di valori plausibili per θ. A ciascuno di questi intervalli è associato un *livello di confidenza* nei confronti dell'ipotesi che θ vi appartenga.

Abbiamo parlato finora di intervalli di confidenza per popolazioni guassiane.

Con un campione gaussiano $\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$ non è una statistica perchè μ non è noto. Al denominatore c'è una variabile aleatoria che a che fare con la distribuzione di χ^2. Quando devo calcolare l'intervallo di confidenza della media utilizzo la σ^2 se a disposizione, altrimenti la S^2.

12.2 Definizione t di student

Prendiamo due VA indipendenti

$$Z, W \quad Z \sim N(0,1) \quad W \sim \chi_a^2 \quad T = \frac{Z}{\sqrt{W}}\sqrt{a}$$

Facendo il rapporto tra due continue ho ancora una continua.

$$f_T(t) = \frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{\pi a}\Gamma\left(\frac{a}{2}\right)}\left(1+\frac{t^2}{a}\right)^{-a\frac{(a+1)}{2}} \qquad t \in \mathbb{R} \quad a = 1, 2, ...$$

questa densità è detta densità t di student con a gradi di libertà.

Qual'è l'applicazione statistica? La distribuzione di Student viene utilizzata per definire degli intervalli di confidenza per la media di una popolazione, sulla base degli stimatori puntuali $\bar{X}$ e S_n^2 della sua media e della sua varianza.

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \quad W = \frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow \frac{Z}{\sqrt{W}}\sqrt{a} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$$

12.3 Stima puntuale

Abbiamo già introdotto lo stimatore. Nel caso del modello gaussiano abbiamo guardato lo stimatore per modelli guassiani.

Ora, il modello di riferimento della popolazione non è più strettamente guassiano.

$$X_1, ..., X_n \text{ iid} \sim f(x, \theta_1, \theta_2, ..., \theta_m) \quad m \geq 1$$

con i parametri θ_i incogniti che rappresentano le caratteristiche della popolazione.

K caratteristica della popolazione di stimare $K = K(\theta_1, \theta_2, ..., \theta_m)$. Un esempio è $f(x, \alpha, \beta) = \Gamma(\alpha, \beta)$ con α e β incogniti.

$$P(X \geq 5) = \int_5^{+\infty} f_{\Gamma(\alpha,\beta)}(x)\, dx$$

12.3.1 Come costruire stimatori $\hat{\theta}$, $\hat{k}$?

Metodo dei momenti Idea: $\mu_1(\theta_1, \theta_2, ..., \theta_m)$ (momento di ordine 1) $= E(X^1)$, quindi $\mu_r(\theta_1, \theta_2, ..., \theta_m) = E(X^r)$. Quindi penso che le caratteristiche di interesse siano i momenti; se ho 5 parametri incogniti dovrò calcolare 5 momenti. Con la media campionaria posso stimare la media, quindi μ_1.

$$\mu_1 \quad \bar{X} = \frac{\sum X_j^1}{n}$$

$$\mu_2 \quad M_2 = \frac{\sum X_j^2}{n}$$

M_i sono detti momenti campionari di ordine i-esimo. In questo modo ho ottenuto una stima non distorta dei momenti, ma il mio obiettivo sono le θ. La cosa semplice da fare è mettere a sistema i momenti teorici con quelli campionari.

$$\begin{cases} \mu_1(\theta_1, \theta_2, ..., \theta_m) = M_1 \\ \mu_m(\theta_1, \theta_2, ..., \theta_m) = M_m \end{cases}$$

È un sistema di m incognite (θ). I momenti campionari sono statistiche.

Se esiste una soluzione $\hat{\theta}_1, ...$ essa è costituita da statistiche che chiamo stimatori di $\theta_1, \theta_2, ...$ ottenuto con il metodo dei momenti.

Quindi posso avere **campioni completamente diversi** ma che daranno come stimatori dei parametri gli **stessi stimatori**, perchè utilizzo solo le informazioni contenute nelle medie (può essere che una distribuzione continua e discreta possano avere gli stessi momenti).

NB: se le μ dipendono linearmente dai θ, allora gli stimatori sono non distorti (asintoticamente).

Esercizio $X_1, ..., X_n$ iid $\sim \Gamma(\alpha, \beta)$. Quali sono $\hat{\alpha}_{mom}$ e $\hat{\beta}_{mom}$?

Sono due parametri quindi avrò bisogno di due equazioni. Il ragionamento alla base di questo esercizio è quello di individuare $\alpha = f\left(\bar{X}, S^2\right)$ e $\beta = f\left(\bar{X}, S^2\right)$ perchè la media e varianza campionaria sono gli unici dati che ho a disposizione.

$$\begin{cases} E(X) = \bar{X} \\ E(X^2) = \frac{\sum X_j^2}{n} \end{cases} = \begin{cases} \alpha\beta = \bar{X} \\ \sigma^2 + (E(x))^2 = M_2 \end{cases} = \begin{cases} \alpha\beta = \bar{X} \\ \sigma^2 = M_2 - \bar{X}^2 \end{cases}$$

$$M_2 - \bar{X}^2 = \frac{\sum X_j^2}{n} - \bar{X}^2 \overset{*}{=} \frac{\sum_{j=1}^n (x_i - \bar{x})^2}{n} = \frac{S^2(n-1)}{n}$$

$$\begin{cases} \alpha\beta = \bar{X} \\ \alpha\beta^2 = S^2\frac{(n-1)}{n} \end{cases} = \begin{cases} \alpha\beta = \bar{X} \\ \bar{X}\beta = \frac{S^{2(n-1)}}{n} \end{cases} = \begin{cases} \hat{\alpha} = \left(\frac{\bar{X}}{S}\right)^2 \frac{n}{n-1} \\ \hat{\beta} = \frac{S^2}{\bar{X}}\frac{n-1}{n} \end{cases}$$

* = Il fatto è che devo fare questo conto in modo semplice

$$\sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2 \overset{\text{sviluppo}}{=} \sum_j X_j^2 + n\bar{X}^2 - 2\bar{X} \cdot n\bar{X} = \sum_j X_j^2 - n\bar{X}^2$$

Esercizio $X_1, ..., X_n$ iid $\sim$ $unif\,(0, \theta)$ con $\theta > 0$ la funzione di distribuzione è uniforme di altezza $\frac{1}{\theta}$ nell'intervallo $[0, \theta]$.

$E\,(X) = \bar{X} \to \frac{\theta}{2} = \bar{X} \Rightarrow \hat{\theta}_{mom} = 2\bar{X}$

Esercizio $unif\,(-\theta, \theta) \quad \theta > 0$

$$E\,(X) = \bar{X} = 0 \quad E\left(X^3\right) = 0 \quad E\left(X^{2k-1}\right) = 0$$

quindi i momenti dispari non posso utilizzarli, posso utilizzare solo quelli pari.

$$E\left(X^2\right) = M_2 = \frac{(2\theta)^2}{12} \Rightarrow \hat{\theta} = \sqrt{M_2 \cdot 3}$$

Limita la complessità del problema utilizzando i momenti di ordine minore.

Metodi di massima verosomiglianza (MLE)[4] È una classe di stimatori largamente utilizzata in statistica. Uno stimatore di questo tipo si ottiene con il ragionamento seguente. Denotiamo con $f\,(x_1, ..., x_n|\theta)$ la funzione di massa congiunta di $X_1, ..., X_n$ oppure la loro densità congiunta, a seconda che siano variabili aleatorie discrete o continue. Poichè stiamo supponendo che θ sia un'incognita, mostriamo esplicitamente che f dipende da θ. Se interpretiamo $f\,(x_1, ..., x_n|\theta)$ come la verosomiglianza che si realizzi la n-pla di dati $x_1, ..., x_n$ quando θ è il vero valore assunto dal parametro, sembra ragionevole adottare come stima di θ quel valore che rende massima la verosomiglianza per i dati osservati. In altri termini, la stima di massima verosomiglianza $\hat{\theta}$ è definita come il valore di θ che rende massima $f\,(x_1, ..., x_n|\theta)$, quando i valori osservati sono $x_1, ..., x_n$.

Sfrutta più informazioni teoriche: forma e tipo di densità del campione.

$$X_1, ..., X_n \text{ iid} \sim f\,(x, \underline{\theta}) \quad \underline{\theta} = (\theta_1, ..., \theta_m)$$

[4]Su Wikipedia (it) questo argomento è spiegato molto bene

La densità congiunta delle VA è $\Pi_{i=1}^{n} f(x_i, \underline{\theta})$ (produttoria perchè le VA sono indipendenti).

$$L_{\underline{\theta}}(x_1, ..., x_n) = \Pi_{i=1}^{n} f(x_i, \underline{\theta})$$

Letta come funzione di $\underline{\theta}$ per assegnata il vettore di x è detta funzione di verosomiglianza del campione, mentre se è letta come funzione $\underline{x}$ per assegnato il vettore $\underline{\theta}$ è la funzione di densità congiunta.

Supponiamo che le x siano discrete

$X_1, ..., X_n \sim Poisson(\theta) \Rightarrow L_{\underline{\theta}} = P_{\underline{\theta}}(X_1 = x_1, ..., X_n = x_n)$. Riesco ad ottenere qualche informazione su θ. Tra tutti i θ possibili scegli quello che ti permette di massimizzazione l'osservazione del dato che hai bisogno.

Stimo $\underline{\theta}$ con il valore che individua la densità $f(x, \theta)$ per cui il campione $x_1, ..., x_n$ effettivamente osservato è più verosimile a

$$max\ L(x_1, ..., x_n) \quad \theta \in \Theta$$

dove Θ è lo spazio parametrico.

Formalizzazione del MLE $x_1, ..., x_n$ realizzazione congiunta di $X_1, ..., X_n$

$L_\theta(x)$ è funzione di verosomiglianza (non è derivabile in θ) che devo studiare per trovare il massimo.

Se esiste una funzione $g(x_1, ..., x_n) \in \Theta$ tale che

$$L_g(x) = \max_{\theta \in \Theta} L_\theta(x) \rightarrow T = g(X_1, ..., X_n) = \theta_{\text{MLE}}$$

con $T = g(X_1, ..., X_n)$ punto di massima della funzione di verosomiglianza.

Uno stimatore MLE di $k = k(\underline{\theta}) \rightarrow \hat{k}_{MLE} = k\left(\hat{\theta}_{MLE}\right)$

Proprietà di questo stimatore: **simmetria** (cambiando l'ordine delle osservazioni lo stimatore non cambia);

La verosomiglianza è un prodotto delle osservazioni. Passo da prodotti a somme con la funzione di logaritmo, per semplificare il problema del derivare tanti moltiplicandi.

Esercizio $X_1, ..., X_n$ iid $unif(0, \theta)$ $\left(\hat{\theta}_{mom} = 2\bar{X}\right)$ $\hat{\theta}_{MLE}$?

$$\begin{aligned} L_\theta(x) &= \Pi_{i=1}^n f(x, \theta) \\ &= \frac{1}{\theta} 1(0 < x_1 < \theta) \cdot ... \cdot \frac{1}{\theta} 1(0 < x_m < \theta) \\ &= \frac{1}{\theta^2} (\theta > x_1 \quad ... \quad \theta > x_m) \end{aligned}$$

$L_\theta(x) = \frac{1}{\theta^n} 1(\theta > max\{x_1, ..., x_n\})$

FIG 0

$\hat{\theta}_{ML} = max\{X_1, ..., X_n\}$

Confronto tra stimatori diversi I due metodi potrebbero risultare stimatori diversi, come faccio a confrontarli?

$P\left(\left|2\bar{X} - \theta\right| < \delta\right) \leq P(|max\{X_1, ..., X_n\}| < \delta)$ allora $\hat{\theta}_{mom}$ è meno accurato di $\hat{\theta}_{ML}$ e scelgo $\hat{\theta}_{ML}$. Scelgo di confrontare stimatori con l'errore quadratico medio, perchè con la strada precedente dovrei conoscere troppe informazioni sulla distribuzione.

$$\begin{aligned} MSE\left(\hat{\theta}_{mom}\right) &= E\left(\left[2\bar{X} - \theta\right]^2\right) \\ &\geq \\ MSE\left(\hat{\theta}_{ML}\right) &= E\left(\left[max\{X_1, ..., X_n\} - \theta\right]^2\right) \end{aligned}$$

Con il metodo della verosomiglianza uso informazioni sulla distribuzione della VA, mentre con il metodo dei momenti uso solo informazioni sulle somme.

$MSE\left(\hat{\theta}_{mom}\right) = Var\left(2\bar{X}\right) + \left[E\left(2\bar{X}\right) - \theta\right]^2 = 4 \cdot Var\left(\bar{X}\right) = 4\frac{\frac{\theta^2}{12}}{n} = \frac{\theta^2}{3n}$

Distorsione dello stimatore MLE $E\left(\hat{\theta}_{ML}\right) = \frac{n\theta}{n+1}$ stimatore distorto, ma con $n \to \infty$ non è distorto. Gli stimatori di massima verosomiglianza possono essere non distorti.

$Var\left(\hat{\theta}_{ML}\right) = \frac{n\theta^2}{(n+1)^2(n+2)} \Rightarrow MSE\left(\hat{\theta}_{ML}\right) = \frac{2\theta^2}{(n+1)(n+2)}$

Gli stimatori di massima verosomiglianza sono quasi sempre consistenti ed asintoticamente non distorti.

Il metodo della verosomiglianza è più accurato su tanti parametri rispetto a quello dei momenti.

13 Esercitazioni

13.1 Esercizio 2.1.2

13.1.1 Punto 1

$X \sim N(2\theta, \sigma^2)$. Conosciamo $\bar{X}$ e S^2 con $n = 25$ e dobbiamo ricavare $P\left(\bar{X} - 0.342 \cdot S - 2\theta \leq 0\right)$

$$X \sim N\left(\mu, \sigma^2\right) \rightarrow \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

$$\begin{aligned} P\left(\bar{X} - 0.342 \cdot S - 2\theta \leq 0\right) &= P\left(\bar{X} - 2\theta < 0.342 \cdot S\right) \\ &= P\left(\frac{\bar{X} - 2\theta}{\sqrt{\frac{S^2}{25}}} \leq 1.71\right) \\ &= P\left(t_{24} \leq 1.71\right) \\ &= 0.95 \end{aligned}$$

13.1.2 Punto 2

Abbiamo calcolato il momento primo $\bar{X} = 120.0$ e secondo $\frac{1}{25}\sum X_i^2 = 146160$. Qual'è la confidenza di questo intervallo $\theta \geq 57435$?

$$\bar{X} - 0.342 \cdot S - 2\theta \leq 0 \quad \Rightarrow \quad \overbrace{\theta \geq \frac{\bar{X} - 0.342 \cdot S}{2}}^{\text{intervallo al } 95\% \text{ per} \theta}$$

Come calcoliamo la varianza campionaria? $\bar{X} = \frac{1}{n}\sum X_i$ e $S^2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{n-1} = \frac{(\sum X_i)^2 - \left(\bar{X}^2\right)}{n-1}$

$$\begin{aligned} S^2 &= \frac{1}{24}\left(25 \cdot 14616 - 25 \cdot (120)^2\right) \\ &= \frac{25}{24} \cdot 216 \\ &= 25 \cdot 9 \\ S &= 15 \end{aligned}$$

$\theta \geq \frac{\bar{X} - 0.342 \cdot S}{2} = \frac{120.0 - 0.342 \cdot 15}{2} = 57.435$

Se devo trovare un intervallo di confidenza per la media non conoscenza nè la media nè la varianza dovrò nel procedimento sicuramente utilizzare la t_{stud}

13.1.3 Punto 3

Sono stati raccolti altri 39 dati (Y) per i quali $\frac{1}{39}\sum Y_i = 110$ e $\frac{1}{39}\sum (Y_i)^2 = 12715$. Quali sono la nuova media e varianza?

Sulla base di tutte le 64 osservazioni

$$\bar{X}_{tot} = \frac{1}{64}\left(25 \cdot 120 + 39 \cdot 110\right)$$

$$\sum X_i^2 = 25 \cdot 14616 + 39 \cdot 12715$$

13.2 Stimatori

13.2.1 Esercizio 3.1.4

$X \quad f_x(x, a, b) = \frac{1}{2b} 1_{(a-b, a+b)}(x)$ è una distribuzione uniforme centrata in a e di intervallo $2b$.

Dobbiamo stimare due parametri. Stimare il momento campionario non sarà sufficiente, vediamo però cosa succede

$$X_1, ..., X_n \quad X \sim (a-b, a+b)$$

$$E(X) = a \quad E\left(X^2\right) = Var(X) + (E(X))^2 = \frac{4b^2}{12} + a^2 = \frac{b^2}{3} + a^2$$

$\hat{a} = \bar{X}$

$b = \sqrt{3 \cdot (E(X) - a^2)} \quad \rightarrow \quad \hat{b} = \sqrt{3\left(\frac{\sum X_i^2}{n} - \bar{X}^2\right)}$

Cosa succede se $a = 0$? $f_x(x) = \frac{1}{2b} 1_{(-b,b)}(x)$ determinare allora lo stimatore di massima verosomiglianza di b, perchè con il metodo dei momenti non recupero informazioni utili.

Lo stimatore di massima verosomiglianza consiste nell'inferire la distribuzione della popolazione sulla base delle osservazioni fatte (es: della scatola con n palline).

$L(x_1, ..., x_n, b) = \Pi_{i=1}^n f(x_i)$

$f_x(x) = \frac{1}{2b} 1_{(-b,b)}(x) = \begin{cases} \frac{1}{2b} & |x| \leq b \\ 0 & \text{altrove} \end{cases}$

$\Pi_{i=1}^n f(x_i) = \begin{cases} \frac{1}{(2b)^n} & |x_i| \leq b \quad \forall i \in [0, n] \\ 0 & \text{altrove} \end{cases} = \begin{cases} \frac{1}{(2b)^n} & b \geq max\{x_i\} \\ 0 & b < max\{x_i\} \end{cases}$

Il massimo di quest'ultima funzione non è possibile perchè non è continua. Quindi lo stimatore di massima verosomiglianza è il massimo dei moduli (sulla base dell'osservazione del grafico).

Se $a = 0 \rightarrow \hat{b}_L = max\{|x_i|\}$

$X \sim unif(-b, b) \rightarrow |X| \sim unif(0, b)$. Dalla prima definizione di X vogliamo trovare lo stimatore di massima verosomiglianza per b. Posso dire $X \sim unif(0, b) \rightarrow \hat{b} = max\{|x_i|\}$.

13.2.2 Esercizio 3.1.9 (modificato)

$X \quad f(x, \theta)_{\theta > 0} = \begin{cases} \frac{1}{\theta} x^{\frac{1}{\theta} - 1} & 0 < x < 1 \\ 0 & \text{altrove} \end{cases}$

Punto 1: Stimatore di θ col metodo dei momenti

$$E(X) = \int_0^1 x \frac{1}{\theta} x^{\frac{1}{\theta}-1} dx = \frac{1}{\theta}\left[\frac{1}{\frac{1}{\theta}+1} x^{\frac{1}{\theta}-1}\right] = \frac{1}{\theta+1}$$

$$\theta = 1 - \frac{1}{E(X)} \rightarrow \hat{\theta}_M = 1 - \frac{1}{\bar{X}}$$

Punto 2: Stimatore di θ con il metodo della massima verosomiglianza

$$L(x_1, ..., x_n, \theta) = \frac{1}{\theta^n} (\Pi x_i)^{\frac{1}{\theta}-1} \quad 0 < x_1 < 1$$

Passo alla rappresentazione logaritmica, perchè mi interessa il valore delle ascisse e non dell'ordinata in sè. Pertanto data la proprietà monotona di $f(x) = e^{g(x)}$posso limitarmi a cercare il massimo di $g(x)$

$$\ln(L(\theta)) = -n \ln \theta + \left(\frac{1}{\theta} - 1\right) \ln(\Pi x_i)$$

$$\begin{aligned} \frac{d}{d\theta} \ln(L(\theta)) &= -\frac{n}{\theta} - \frac{1}{\theta^2} \sum \ln x_i = 0 \\ \theta &= -\frac{1}{n} \sum \ln x_i \end{aligned}$$

Come faccio a sapere se è un punto di massimo? La funzione è definita per $\theta > 0$. Quanto θ tende a zero, la funzione $L(\theta)$ tende a meno infinito, quando θ tende ad infinito, la funzione tende a meno infinito pure. Pertanto θ è il punto di massimo assoluto.

$$\hat{\theta}_L = -\frac{1}{n} \sum \ln x_i$$

È uno stimatore sensato? Sì perchè valore positivo (i logaritmi sono tutti negativi, poichè l'argomento è minore di 1).

Punto 3: $Y = -\ln X \quad g(x) = -\ln X \iff h(y) = e^{-Y}$

$f_Y(y) = \frac{1}{\theta}\left(e^{-y}\right)^{\frac{1}{\theta}-1} \cdot |e^{-y}| \quad y > 0$

$$f_Y(y) = \begin{cases} \frac{1}{\theta}e^{-\frac{y}{\theta}} & y > 0 \\ 0 & \text{altrove} \end{cases}$$

Punto 4: Discutere le proprietà dello stimatore

- **Indistorto**: il valore atteso coincide con il parametro ricercato

$$E\left(\hat{\theta}_L\right) = E\left(-\frac{1}{n}\sum \ln x_i\right) = \frac{1}{n}E\left(\sum\left(-\ln x_i\right)\right) = \frac{1}{n}\sum\left(\underbrace{E\left[-\ln x_i\right]}_{\theta}\right) = \frac{1}{n}n\theta = \theta$$

- **Consistente**: all'aumentare delle osservazioni, l'errore quadratico medio tende a zero. Dato che è indistorto è sufficiente osservare se la varianza tende a zero.

$$Var\left(\hat{\theta}_L\right) = Var\left(\frac{1}{n}\sum\left(-\ln x_i\right)\right) = \frac{1}{n^2}\overbrace{\sum\left(\underbrace{Var\left(-\ln x_i\right)}_{\theta^2}\right)}^{\text{va indipendenti}} = \frac{\theta^2}{n} \to 0$$

13.2.3 Esercizio 3.1.10

$$f_x(x) = \begin{cases} \overbrace{\frac{2}{\theta}\left(1-\frac{x}{\theta}\right)}^{\frac{2}{\theta^2}(\theta-x)} & 0 < x < \theta \\ 0 & \text{altrove} \end{cases}$$

Punto 1 Vogliamo calcolare il valore atteso e varianza della media campionaria.

$$E(X) = \frac{2}{\theta^2} = \int_0^\theta x(\theta - x)\,dx = \frac{2}{\theta^2}\left[\frac{\theta}{2}x^2 - \frac{x^3}{3}\right]_0^\theta = \frac{\theta}{3}$$

$$E\left(X^2\right) = \frac{2}{\theta^2}\int_0^\theta x^2\left(\theta - x\right)dx = \frac{2}{\theta^2}\left[\frac{\theta}{3}x^3 - \frac{x^4}{4}\right]_0^\theta = \frac{\theta^2}{6}$$

$$Var\left(X\right) = \frac{\theta^2}{6} - \frac{\theta^2}{9} = \frac{\theta^2}{18}$$

Quindi $E\left(\bar{X}\right) = \frac{\theta}{3} \quad Var\left(\bar{X}\right) = \frac{\theta^2}{18n}$

Punto 2 $\hat{\theta} = 3\bar{X}$

$E\left(\hat{\theta}\right) = \theta$ non distorto

$MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) = Var\left(3\bar{X}\right) = \theta \cdot Var\left(\bar{X}\right) = \frac{\theta^2}{2n}$

Punto 3 La situazione è che $n = 1$. Determinare lo stimatore di massima verosomiglianza di θ.

Abbiamo un unica osservazione, quindi la funzione di verosomiglianza è f_x

$$L\left(\theta\right) = \frac{2}{\theta^2}\left(\theta - x\right) \quad \theta \geq x$$

$$L'\left(\theta\right) = -\frac{2}{\theta^2} + \frac{2x}{\theta^3} = \frac{2}{\theta^3}\left(x - \theta\right) < 0$$

Morale $\hat{\theta}_L = x$ (ho solo un caso, quindi x è unico)

Punto 4: Calcolare l'errore quadratico medio $MSE = E\left[\left(\hat{\theta}_L - \theta\right)^2\right] = Var\left(\hat{\theta}\right) + \underbrace{\left(E\left(\hat{\theta} - \theta\right)\right)^2}_{\text{bias o distorsione}} =$

$\underbrace{Var\left(X\right)}_{\frac{\theta^2}{18}} + \left(\underbrace{E\left(X\right)}_{\frac{\theta}{3}} - \theta\right)^2 = \frac{\theta^2}{18} + \frac{4}{9}\theta^2 = \frac{\theta^2}{2}$

Conclusione $\hat{\theta}_M = 3X \quad MSE = \frac{\theta^2}{2}$ ottenuto col metodo dei momenti.

Lo stimatore ottenuto con la verosomiglianza è distorto, mentre quello ottenuto dai momenti non lo è. Dato che l'errore quadratico medio è uguale per entrambi, allora preferisco utilizzare lo stimatore non distorto.

Se non avessero avuto lo stesso MSE utilizzo lo stimatore in base a quello che mi interessa (non distorsione o minore MSE).

14 Ottimalità degli stimatori

Possiamo usare l'MSE per stabilire qual'è lo stimatore che ha MSE più piccolo rispetto a tutti gli stimatori? Così posto il problema non ha soluzione

$X_1, .., X_n$ iid $\sim f(x, \theta)$

k caratteristica

Classe di tutti i possibili stimatori -> {T stimatore di k che hanno MSE} = δ

$\exists$ un $T^* \in \delta : MSE(T^*) \leq MSE(T) \, \forall T \in \delta, \forall \theta$? No

perchè è una classe troppo grossa, per cui esistono stimatori inutili.

Supponiamo che k (caratteristica da stimare) sia la media. $T = \bar{X}$, allora il candidato è $T* = \bar{X}$

$T = 10.23$ è un possibile stimatore di μ

$MSE(T) = Var(10.23) + (E[10.23] - \mu)^2 = 0 + (10.23 - \mu)^2$

Se $\mu = 10.23 \Rightarrow MSE(T) = 0$

$MSE\left(\bar{X}\right) = \frac{\sigma^2}{n} > 0 \Rightarrow \nexists T$ che ha $MSE(T) \leq MSE(10.23)$

Come identifico una classe più piccola? La proprietà da richiedere proviene dalla MSE.

$$MSE(T) = Var(T) + \underbrace{(E[T] - k)^2}_{\text{T non distorto} = 0} \quad \Rightarrow MSE(T) = Var(T)$$

Invece di minimizzare la somma, pongo il secondo membro = 0 e minimizzo la varianza

Cercare lo stimatore migliore T^* significa quello per cui $Var(T^*) \leq Var(T)$ nella classe degli stimatori $\delta_U = \{TdK : E(T) = k, \forall \theta\}$

1. Esiste lo stimatore non distorto $E(T) = k$, quindi restringo il campo ai stimatori non distorti
2. ,Parlo di una classe di stimatori, perchè lo stimatore non distorto di quella caratteristica (una volta trovato), non è detto che sia unico.

14.1 Stimatori non distorti

$$E(T) = k \quad \forall\theta$$

1. **Possono non esistere**. Es: $X_1 \sim Bin(5,\theta) \quad 0 < \theta < 1 \quad k = \frac{1}{\theta} \Rightarrow \nexists T$ non distorto di $\frac{1}{\theta}$. Questo non significa che non posso stimare
2. **Unicità**. $X_1, ..., X_n \sim Poisson(\theta) \quad \theta > 0 \Rightarrow \theta = \mu = E(X_1) \Rightarrow \hat{\theta}_1 = \bar{X}$ e $\theta = Var(X_1) \Rightarrow \hat{\theta}_2 = S^2$. $E\left(\hat{\theta}_1\right) = \theta$ e $E\left(\hat{\theta}_2\right) = Var(X_1) = \theta$

   Se esistono due stimatori, allora ne esistono infiniti, perchè li ottengo come combinazione lineare con pesi a somma 1.

$$\hat{\theta}_a = a\bar{X} + (1-a)S^2 \quad a \in R$$

14.1.1 Esempio

$X_1 \overset{\text{iid}}{\sim} Poiss(\theta) \quad k = P(\text{nessuna telefonata 2° e 3° giorno}) = P(X_2 = 0, X_3 = 0) \overset{\text{iid}}{=} P(X_2 = 0) P(X_3 = 0) = f_\theta(0) f_\theta(0) = e^{-\theta}e^{-\theta} = e^{-2\theta}$

$X_1 \to T = g(X_1) \quad k = \underbrace{e^{-2\theta}}_{\text{caratteristica}}$

$T : E(T) = e^{-2\theta}$

$E(T) = \sum_{k=0}^{\infty} g(k) f_\theta(k) = e^{-2\theta} \Rightarrow \sum_{k=0}^{\infty} g(k) \frac{e^{-\theta}\theta^k}{k!} = \sum_{k=0}^{\infty} \frac{(-2\theta)^k}{k!} = e^{-2\theta} \quad \forall\theta$

$\sum_{k=0}^{\infty} g(k) \frac{\theta^k}{k!} = \sum_{k=0}^{\infty} (-1)^k \frac{\theta^k}{k!} \iff g(k) = (-1)^k \Rightarrow T = (-1)^{X_1}$

Questo stimatore non ha senso perchè stima una quantità che è nell'intervallo [0,1] con una quantita discreta -1,+1. Quindi ho trovato uno stimatore non distorto, ma è inutile. Se mi pongo il problema dello stimatore ottimale la soluzione è stata trovata.

14.2 Problema della ricerca dello stimatore ottimale

Quello con varianza uniformemente minima: varianza più piccola rispetto alla varianza degli altri con qualsiasi valore di θ. Richiesta necessaria da cui non posso prescindere. Allora questa condizione è necessaria.

Ci basta sapere che esiste una metodologia che permette di dare le **condizioni perchè stimatori non distorti esistono**, ma che noi non tratteremo per la complessità degli argomenti sull'ottimalità da introdurre.

Ogni **stimatore non distorto ha una certa varianza**, posso dire qual'è il valore più piccolo che la varianza dello stimatore può avere? Andando così a calcolare la varianza reale e lo confronto con quello dello stimatore. Se sono vicino allora la strategia applicata può essere corretta, altrimenti cambio strategia.

È buona cosa che lo stimatore abbia **varianza minima** perchè così **i valori si concentrano nella media**. Ma per lo stimatore la media corrisponde proprio alla caratteristica da cercare.

14.3 Diseguaglianza di Frechet-Cramer-Kau

Dati

- $X_1, ..., X_n \sim f(x, \theta) \quad \theta \in \Theta \leq R$
- $k = k(\theta)$ da stimare
- T è stimatore di K tale che
  - $E(T) = k \quad \forall \theta$
  - $Var(T) < \infty$

Ipotesi

1. Θ intervallo aperto di R (perchè lavoro con molte derivate e non voglio pormi problemi sui punti limite)
2. $S = \underbrace{\{x : f(x, \theta) > 0\}}_{\text{supporto della densità}}$ è indipendente da θ
3. $\theta \vdash f(x, \theta)$ è derivabile $\forall x \in S$

4. $E\left[\frac{d}{d\theta}\ln f\left(X_1,\theta\right)\right]=0 \quad \forall\theta$
5. $0<E\left[\left(\frac{d}{d\theta}\ln f\left(X_1,\theta\right)\right)^2\right]<\infty$
6. k derivabile e $k'\left(\theta\right)=E\left[T\cdot\frac{d}{d\theta}\ln L_\theta\left(x_1,...,x_n\right)\right]$

Enunciato

$$Var\left(T\right)\geq\frac{\left(k'\left(\theta\right)\right)^2}{n\cdot I\left(\theta\right)}\geq 0 \quad \forall\theta\in\Theta$$

con $I\left(\theta\right)=\overbrace{E\left[\left(\frac{d}{d\theta}\ln f\left(X_1,\theta\right)\right)^2\right]}^{\text{informazione di Fischer}}$

Inoltre $Var\left(T\right)=\frac{\left(k'\left(\theta\right)\right)^2}{n\cdot I\left(\theta\right)}$ sse $\frac{d}{d\theta}\ln L_\theta\left(x_1,...,x_n\right)$ scopro di poterla rappresentare come $a\left(n,\theta\right)\left[T-k\left(\theta\right)\right]$ che succede con probabilità uno (ovvero succede per ogni parametro). Linearizzo la derivata del logaritmo di L.

Più osservazioni hai, una migliore stima puoi ottenere. Per modelli regolari (soddisfacenti quelle ipotesi) $\frac{1}{n}$ è il tasso ottimale cui cui la varianza va a zero.

Più grande è l'informazione di Fischer e maggiore sarà la concentrazione delle informazioni.

Il meglio che posso fare dipende da tre membri: n, informazione di Fischer e la derivata della caratteristica.

Lo stimatore migliore ce l'ho quando il modello.. Lo stimatore è un buon stimatore quando la dipendenza delle osservazioni è presente solo in T

$$\frac{d}{d\theta}\ln L_\theta\left(x_1,...,x_n\right)=\underbrace{a\left(n,\theta\right)\left[\overbrace{T}^{x_i}-k\left(\theta\right)\right]}_{\text{separo }\theta\text{ e }x}$$

Integrando il secondo membro ottengo la funzione di verosomiglianza.

Ipotesi 2. $f\left(x,\theta\right)=\varepsilon\left(\theta\right)\rightarrow S=\{\}=\left(0,+\infty\right)$ ipotesi indipendente da θ. Stesso ragionamento con la poissoniana. Se prendo il modello $unif\left(0,\theta\right)$ e cambio θ allora cambiano le osservazioni, per cui in questo caso l'ipotesi non è indipendente da θ.

Ipotesi 3. Richiesta necessaria perchè dopo utilizzo le derivate

Ipotesi 4. La derivata dell'integrale è l'integrale della derivata (della densità).

$1 = \int_R f(x,\theta)\,dx = \int_S f(x,\theta)\,dx \Rightarrow 0 = \frac{d}{d\theta}\int_S f(x,\theta)\,dx \overset{\text{ipotesi}}{=} \int_S \frac{d}{d\theta} f(x,\theta)\,dx = \int_S \frac{\frac{d}{d\theta}f(x,\theta)}{f(x,\theta)} f(x,\theta)\,dx = \int_S \frac{d}{d\theta}\ln f(x,\theta)\, f(x,\theta)\,dx = E\left[\frac{d}{d\theta}\ln f(x_1,\theta)\right]$

14.3.1 Esempio

$X_1,..,X_n \sim Poiss(\theta)$

1. Determinare uno stimatore $\hat{\theta}_{ML}$ di θ
2. Determinare $\hat{k}_{ML}$ di $P(X_1 > 0)$
3. Determinare se esiste T non distorto con varianza $\frac{(k'(\theta))^2}{n \cdot I(\theta)}$

$L_\theta(x_1,..,x_n)$?

$f(x,\theta) = \frac{e^{-\theta}\theta^x}{x!} \quad \theta > 0$

Ipotesi FCR?

- $\Theta = (0,+\infty)$ ok
- $\{x : f(x,\theta) > 0\} = naturali$
- $f(x,\theta)$ è ∞ volte derivabili in $\theta \quad \forall x = 0,1,2,...$
- $\ln f(x_1,\theta) = \ln\left(\frac{e^{-\theta}\theta^{X_1}}{X_1!}\right) = -\theta + X_1 \ln(\theta) - \ln(X_1!) \Rightarrow \frac{d}{d\theta}\ln f(x_1,\theta) = -1 + \frac{X_1}{\theta}$
  $E\left[\frac{d}{d\theta}\ln f(x_1,\theta)\right] = E\left[-1+\frac{X_1}{\theta}\right] = E\left[\frac{X_1-\theta}{\theta}\right] = \frac{1}{\theta}[\theta-\theta] = 0$ OK
- $E\left[\left(\frac{X_1-\theta}{\theta}\right)^2\right] = \frac{1}{\theta^2}E[X_1-\theta]^2 = \frac{Var(X_1)}{\sigma^2}$ OK?

$L_\theta(x)$?

$k_1(\theta) = \theta \to k_1'(\theta) = 1 \Rightarrow$ confine $FCR = \frac{1}{n \cdot I(\theta)} = \frac{\theta}{n}$ è il meglio che posso fare per stimare k1

$k_2(\theta) = 1 - P(X_1 = 0) = 1 - e^{-\theta} \Rightarrow k_2'(\theta) = e^{-\theta} \Rightarrow$ confine FCR $= \frac{\left(e^{-\theta}\right)^2}{n \cdot I(\theta)} = \frac{\theta e^{-2\theta}}{n} =$ meglio che posso fare per stimare k2

... GUARDARE ESERCIZIARIO

$\frac{d}{d\theta} \lg L_\theta = -n + \frac{\sum X_i}{\theta} \geq 0 = \frac{n}{\theta}\left[-\theta + \frac{\sum X_j}{n}\right] = \frac{n}{\theta}\left[\bar{X} - \theta\right] \geq 0 \leftrightarrow \bar{X} \geq \theta$

Lo stimatore di massima verosimiglianza è $\bar{X} \to \hat{\theta}_{ML} = \bar{X}$.

Domanda 3. $\frac{d}{d\theta} \lg L_\theta = \frac{n}{\theta}\left[\bar{X} - \theta\right]$

$a(n, \theta) = \frac{n}{\theta} \quad T = \bar{X} \quad k = \theta \Rightarrow \frac{d}{d\theta} \lg L_\theta = a(n, \theta)(T - k)$. T stimatore con varianza più piccola possibile. Allora $\bar{X}$ è lo stimatore ottimo. È non distorto? $E\left(\bar{X}\right) = \mu = \theta$

Non è un caso che lo stimatore ottimo l'abbia trovato con il metodo della max verosomiglianza.

Stimatore efficiente = quelli di FCR.

Se uno stimatore efficiente esiste allora è perforza di massima verosomiglianza.

15 Esercitazioni

Il limite CR è interessante perchè è il limite inferiore della varianza dello stimatore. Per cui potenzialmente lo stimatore con il più basso limite è quello potenzialmente migliore.

15.1 Esercizio 3.1.5

$$f_x(x, a, b) = \begin{cases} \frac{2(b-x)}{(b-a)^2} & a \leq x \leq b \\ 0 & \text{altrove} \end{cases}$$

Dobbiamo stimare due parametri. Quindi se con il metodo dei momenti li ottengo entrambi sono a posto.

$$\begin{aligned} E[X] &= \int_a^b x \frac{2(b-a)}{(b-a)^2} dx \\ &= \frac{2}{(b-a)^2} \int_a^b \left(bx - x^2\right) dx \\ &= \frac{2}{(b-a)^2} \left[\frac{b}{2}x^2 - \frac{1}{3}x^3\right]_a^b \\ &= \frac{2}{(b-a)^2} \left(\frac{1}{6}b^3 - \frac{ba^2}{2} + \frac{a^3}{3}\right) \\ &= \frac{2}{(b-a)^2} \left(\frac{b^3 - 3a^2b + 2a^3}{6}\right) \\ &= \bar{X}_n \end{aligned}$$

Come suggerimento l'esercizio dava il valore atteso di

$$\begin{aligned} E\left[X^2\right] &= \frac{(b-a)^2}{18} + \frac{(2a+b)^2}{9} \\ &= \frac{1}{n}\sum X_i^2 \end{aligned}$$

$$\begin{aligned} b^3 - 3a^2b + 2a^3 &= b^3 - a^3b - 2a^2b + 2a^3 \\ &= b\left(b^2 - a^2\right) - 2a^2(b-a) \\ &= (b-a)^2[b+2a] \end{aligned}$$

$$E[X] = \frac{b+2a}{3} = \bar{X}_n$$

$$\begin{cases} 2a + b = 3\bar{X}_n \\ (b-a)^2 = 18\left(\frac{1}{n}\sum X_i^2 - \bar{X}_n^2\right) \end{cases} = \begin{cases} 2a + b = 3\bar{X}_n \\ b - a = \sqrt{18\left(\frac{1}{n}\sum X_i^2 - \bar{X}_n^2\right)} \end{cases}$$

15.2 Esercizio 3.1.6

$X \sim N(\mu, \sigma^2)$. Trovare $\hat{\mu}_L$ e σ^2 nel caso la media sia nota e non. Utilizzo la funzione di verosomiglianza

15.2.1 Stima della media

$L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2}$

$l = \ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(X_i-\mu)^2$

$\frac{dl}{d\mu} = \frac{1}{\sigma^2}\sum(X_i-\mu) = 0 \quad \Longrightarrow \quad \sum\mu_i - n\mu = 0 \Rightarrow \hat{\mu}_L = \frac{1}{n}\sum x_i = \bar{X}_n$

Lo stimatore è efficiente (ovvero raggiunge il limite di CR)? Finora sappiamo che è non distorto e consistente in media quadratica.

15.2.2 Stima della varianza

$$-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(X_i-\mu)^2 \overbrace{=}^{\text{prop. log}} -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(X_i-\mu)^2$$

$\frac{dl}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(X_i-\mu)^2 = 0$

$\frac{-n\sigma^2+\sum(X_i-\mu)^2}{2(\sigma^2)^2} = 0 \quad \Rightarrow \quad \sigma^2 = \frac{1}{n}\sum(X_i-\mu)^2$

Media nota Lo stimatore è $\frac{1}{n}\sum(X_i-\mu)^2$, corretto è non distorto

Media non nota Lo stimatore è $\frac{1}{n}\sum\left(X_i-\bar{X}_n\right)^2$, distorto perchè $S^2 = \frac{1}{n-1}\sum\left(X_i-\bar{X}_n\right)^2$ ed allora lo stimatore ha al denominatore $\frac{1}{n}$ invece di $\frac{1}{n-1}$ come volevamo aspettarci. Comunque lo stimatore è asintoticamente non distorto.

Sui stimatori ottenuti col metodo dei momenti non possiamo dire nulla, ma quelli ottenuti con il metodo della verosomiglianza sono consistenti ed asintoticamente non distorti.

15.3 Esercizio 3.1.14

$$\underbrace{p\left(x,\theta\right)_{\theta\in R-\{0\}}}_{\text{densità discreta}} = \begin{cases} \frac{e^{-\theta^2}\theta^{2x}}{x!} & x\in N \\ 0 & \text{altrove} \end{cases}$$

Funzione di verosomiglianza: $L = \frac{e^{-n\theta^2}\theta^{2\sum X_i}}{\Pi_i(x_i!)}$

15.3.1 Trovare $\hat{\theta^2_M}$

Questa VA assomiglia a quella poissoniana $\rightarrow\sim Poisson\left(\lambda=\theta^2\right)$. Sappiamo che $E\left[X\right]=\theta^2$, per cui $\hat{\theta^2_M}=\bar{X}$. Lo stimatore è non distorto, perchè il valore atteso individua il parametro cercato.

Questo stimatore è efficiente?

Come trovo il limite FCR? $l_{CR} = \frac{\left[k'(\theta)\right]^2}{n\cdot E\left[\left(\frac{d}{d\theta}\ln f(x)\right)^2\right]} = \frac{\left[k'(\theta)\right]^2}{-n\cdot E\left[\frac{d^2}{d\theta^2}\ln f(x)\right]}$

$k\left(\theta\right)=\theta^2 \rightarrow \ln f\left(x\right) = -\theta^2 + 2x\ln\theta - \ln x!$

$\frac{d}{d\theta}\ln f\left(x\right) = -2\theta + \frac{2x}{\theta}$

$$l_{CR} = \frac{4\theta^2}{n\cdot E\left[\underbrace{\left(-2\theta+\frac{2x}{\theta}\right)^2}_{\text{non uso la der}}\right]}$$

$$E\left[\frac{\left(2x-2\theta^2\right)^2}{\theta^2}\right] = \frac{4}{\theta^2}\underbrace{E\left[\left(x-\theta^2\right)^2\right]}_{Var(X)\underbrace{=\theta^2}_{Poiss}}$$

$l_{CR} = \frac{4\theta^2}{n4} = \frac{\theta^2}{n}$

Domanda $Var\left(\bar{X}_n\right) = \frac{\theta^2}{n}$ (UMVUE = stimatore non distorto a varianza uniformemente minima)

15.4 Esercizio 3.1.16

$X, Y \sim N\left(0, \underbrace{\sigma^2}_{\text{inc}}\right)$ vogliamo stimare la precisione del tiratore

$Q_i = x_i^2 + y_i^2$ distanza tra punto colpito e centro del bersaglio (0,0)

precisione $\tau = \frac{1}{\sigma^2}$

15.4.1 Risoluzione

Se X e Y fossero due normali standard allora $Q_i \sim \chi_2^2$. Ma le due variabili aleatorie non sono standard.

$V = X^2$

$$\begin{aligned} F_V(v) &= P\left(X^2 \leq \theta\right) \\ &= P\left(-\sqrt{v} < X < \sqrt{v}\right) \\ &= P\left(-\sqrt{\frac{v}{\sigma^2}} < Z < \sqrt{\frac{v}{\sigma^2}}\right) \\ &= 2\Phi\left(\sqrt{\frac{v}{\sigma^2}}\right) - 1 \end{aligned}$$

$$\begin{aligned} f_V(v) &= 2\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{v}{\sigma^2}} \cdot \frac{1}{2\sqrt{v\sigma^2}} \\ &= \frac{1}{\sqrt{\pi}} \left(\frac{1}{2\sigma^2}\right)^{\frac{1}{2}} v^{-\frac{1}{2}} e^{-\frac{v}{2\sigma^2}} \end{aligned}$$

$X^2 \sim \Gamma\left(\frac{1}{2}, 2\sigma^2\right) \quad Y^2 \sim \Gamma\left(\frac{1}{2}, 2\sigma^2\right)$

$X^2 + Y^2 \sim \Gamma\left(1, 2\sigma^2\right) \rightarrow Q_i \sim \varepsilon\left(2\sigma^2\right)$

$f(q) = \frac{1}{2\sigma^2} e^{-\frac{1}{2\sigma^2}q} \quad q > 0$

$L = \left(\frac{1}{2\sigma^2}\right)^n e^{-\frac{1}{2\sigma^2}\sum q_i} \rightarrow l = \ln L = -n\ln\left(2\sigma^2\right) - \frac{1}{2\sigma^2}\sum q_i$

$\frac{dl}{d\sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum x_i = 0 \quad \Rightarrow \quad \sigma^2 = \frac{1}{2n}\sum q_i$

$\hat{\sigma_L^2} = \frac{1}{2}\bar{Q}_n$

$\hat{\tau}_L \overbrace{=}^{\text{invarianza}} \frac{2}{\bar{Q}_n} \qquad \widehat{g\left(\theta\right)_L} = g\left(\hat{\theta}_L\right)$

15.4.2 Lo stimatore è efficiente?

Calcolo il limite di CR e valuto la varianza

$Var\left(\sigma_L^2\right) = \frac{1}{4}Var\left(\bar{Q}_n\right) = \frac{1}{4n}Var\left(Q\right) = \frac{1}{4n}\cdot 4\sigma^4 = \frac{\sigma^4}{n}$

$\ln f = -\ln\left(2\sigma^2\right) - \frac{q}{2\sigma^2}$

$\frac{d\ln f}{d\sigma^2} = -\frac{1}{\sigma^2} + \frac{q}{2(\sigma^2)^2} = \frac{q-2\sigma^2}{2(\sigma^2)^2}$

$l_{CR} = \frac{1}{n\cdot E\left[\frac{(q-2\sigma^2)^2}{(2\sigma^4)^2}\right]} = \frac{4\sigma^8}{\underbrace{n\cdot E\left[\left(Q-2\sigma^8\right)^2\right]}_{Var(Q)=4\sigma^4}} = \frac{\sigma^4}{n}$

15.5 Esercizio 3.1.18

X è una VA log-normale, ovvero $\ln X \sim N\left(m, v\right)$

$E\left(X\right) = e^{m+\frac{v}{2}}$

$Var\left(X\right) = e^{2m+v}\left(e^v - 1\right)$

$m = \ln\theta \quad v = 4$

Stimare θ con i momenti e calcolarne l'MSE

...

Stimare θ con il metodo della verosomiglianza, si vedrà che è distorto con la differenza di un coefficiente. Quindi come correggerlo?

Suggerimento $Y \sim N\left(\mu, \sigma^2\right)$, $E\left(e^{tY}\right) = e^{\mu t + \frac{\sigma^2}{2}t^2}$

16 Dimostrazione diseguaglianza FCR

$X_1, ..., X_n$ iid $\sim f(x, \theta)$, k_j T stimatore non distorto di K con varianza

16.1 Ipotesi

1. $\Theta \subseteq R$ intervallo aperto
2. $S = \{x : f(x, \theta) > 0\}$ indipendenti da θ
3. $\theta \longmapsto f(x, \theta)$ derivabile $\forall x \in \delta$
4. $E\left[\frac{d}{d\theta} \ln f(x_1, \theta)\right] = 0$
5. $0 < I(\theta) < \infty \quad I(\theta) = E\left[\left(\left(\frac{d}{d\theta}\right) \ln f(X_1, \theta)\right)^2\right]$
6. k è derivabile e $k'(\theta) = E\left(T \cdot \frac{d}{d\theta} \ln L_\theta(X_1, ..., X_n)\right)$

16.2 Tesi

1. $Var(T) \geq \frac{(k'(\theta))^2}{n \cdot I(\theta)} \quad \forall \theta$
2. $Var(T) = \frac{(k'(\theta))^2}{n \cdot I(\theta)}$

sse $\frac{d}{d\theta} \ln L_\theta(x_1, ..., x_n) = a(n, \theta)(T - k)$ con probabilità 1.

16.3 Pre-Dimostrazione

$Cov(x, y) = E[(x - \mu_x)(y - \mu_y)] = E(xy) - E(x) E(y)$

$\rho(x, y) = \frac{Cov(x,y)}{\sqrt{Var(x) Var(y)}}$

- $|\rho(x, y)| \leq 1$

- $\rho = 1$ sse $P(y = ax + b) = 1$
  - $a = \frac{Cov(x,y)}{Var(x)}$
  - $b = \mu_y - a\mu_x$

16.4 Dimostrazione

$X_1, ..., X_n$ iid. Trasformiamo ciascuna variabile nel seguente modo $Y_1, ..., Y_n$ tale che $Y_1 = \frac{d}{d\theta} \ln f(X_1, \theta)$. Se le X sono indipendenti lo saranno anche le Y, inoltre le Y avranno le stesse leggi di X.

Calcoliamo ora media e varianza: $E(Y_1) = E\left(\frac{d}{d\theta} \ln f(X_1, \theta)\right) \overbrace{=}^{(4)} 0$. $Var(Y_1) = E(Y_1^2) - E(Y_1)^2 = E(Y_1^2) =$
$\underbrace{E\left(\left(\frac{d}{d\theta} \ln f(X_1, \theta)\right)^2\right)}_{\text{inf di Fischer}} = I(\theta) < \infty$

$\sum_{i=1}^n Y_i = \sum \frac{d}{d\theta} \ln f(X_1, \theta) = \frac{d}{d\theta} \sum \ln f(X_1, \theta)$

$$\begin{aligned} & = \frac{d}{d\theta} \ln \Pi_{i=1}^n f(X_1, \theta) \\ \sum_{i=1}^n Y_i & = \frac{d}{d\theta} \ln L_\theta(X_1, ..., X_n) \end{aligned}$$

$--$

$$\begin{aligned} E\left(\frac{d}{d\theta} \ln L_\theta\right) & = E\left(\sum Y_i\right) \\ & = \sum E(Y_i) \\ & = 0 \end{aligned}$$

$$\begin{aligned} Var\left(\frac{d}{d\theta}\ln L_\theta\right) &= Var\left(\sum_i Y_i\right) \\ &= \sum_i Var\left(Y_i\right) \\ &= \sum_i I\left(\theta\right) \\ &= n \cdot I\left(\theta\right) \\ &= E\left[\left(\frac{d}{d\theta}\ln\left(\theta\right)\right)^2\right] \end{aligned}$$

$n \cdot I\left(\theta\right)$ informazione dell'insieme di campioni.

– –

$$\begin{aligned} Cov\left(T, \frac{d}{d\theta}\ln L_\theta\left(X_1, ..., X_n\right)\right) &= E\left(T \cdot \frac{d}{d\theta}\ln L_\theta\left(X_1, ..., X_n\right)\right) - E\left(T\right)\cdot 0 \\ &\overbrace{=}^{(6)} k'\left(\theta\right) \end{aligned}$$

$$\begin{aligned} \rho\left(T, \frac{d}{d\theta}, ...\right) &= \frac{k'\left(\theta\right)}{\sqrt{Var\left(T\right)\cdot Var\left(\frac{d}{d\theta}\ln L\right)}} \\ &= \frac{k'\left(\theta\right)}{\sqrt{Var\left(T\right)\cdot n\cdot I\left(\theta\right)}} \end{aligned}$$

Valido quando le varianze sono strettamente maggiori di zero. Quindi il denominatore è maggiore di 0? $I\left(\theta\right)$ è maggiore di zero per ipotesi, n pure, mentre $Var\left(T\right)$ è maggiore di zero perchè lo stimatore è non distorto (non chiaro).

$\left|\frac{k'(\theta)}{\sqrt{Var(T)\cdot Var\left(\frac{d}{d\theta}\ln L\right)}}\right| = \frac{|k'(\theta)|}{\sqrt{Var(T)\cdot n\cdot I(\theta)}} \leq 1$

$[...]^2 \leq 1$

$\frac{k'(\theta)}{Var(T)nI(\theta)} \leq 1 \Longleftrightarrow Var(T) \geq \frac{(k'(\theta))^2}{nI(\theta)}$ come volevasi dimostrare (1)

$--$

$\left|\rho\left(T, \frac{d}{d\theta}\ln(\theta)\right)\right| \leq 1 \rightarrow \left|\rho\left(T, \frac{d}{d\theta}\ln(\theta)\right)\right| = 1 \Longleftrightarrow \frac{d}{d\theta}\ln L_\theta(X_1, ..., X_n) = a(n,\theta)\cdot T + b(n,\theta)$

$$\begin{aligned} 0 &= E\left[\frac{d}{d\theta}\ln L_\theta(X_1, ..., X_n)\right] \\ &= E\left[a(n,\theta)\cdot T + b(n,\theta)\right] \\ &= aE(T) + b \\ -aE(T) &= b \\ -ak_\theta &= b \end{aligned}$$

$$\frac{d}{d\theta}\ln L_\theta = a(n,\theta)(T-k) \text{ come volevasi dimostrare}$$

$--$

$X_1, ..., X_n$ iid $\sim f(x,\theta)$, con k caratteristica da stimare. Sono inoltre soddisfatte tutte le ipotesi di FCR.

Che relazione c'è tra stimatori efficienti e stimatori di verosomiglianza?

$\Longrightarrow \hat{k}_{ML}$ soddisfa P' equazioni, $\frac{d}{d\theta}\ln L_\theta(X_1, ..., X_n) = 0 \quad (*)$.

Supponete che esiste uno stimatore T **efficiente** per k $\Longleftrightarrow \frac{d}{d\theta}\ln L_\theta = a(n,\theta)(T-k) \quad (**)$

$(*)(**) \rightarrow a(n,\theta) = 0$ oppure $T - k = 0 \Longrightarrow \frac{d}{d\theta}\ln L_\theta = 0 \quad \forall\theta \in \Theta \Longrightarrow \ln L_\theta$ è indipendente da $\theta \Longrightarrow L_\theta$ è indipendente da θ. Ma questo è assurdo $\Longrightarrow a(n,\theta) \neq 0$ per qualche θ.

Supponete che esiste uno stimatore T **efficiente** per k. Allora $\hat{k}_{ML} = T$

Per ricercare lo stimatore efficiente determino prima quello di max verosomiglianza, se questo è efficiente sei a posto, se non lo è allora non esiste alcun stimatore efficiente.

$T_{effic} \Longrightarrow \hat{k}_{ML} = T \qquad \hat{k}_{ML} = T \not\Longrightarrow T_{effic}$

16.4.1 Controesempio

$X_1, ..., X_n \sim Poiss(\theta) \rightarrow \hat{\theta}_{ML} = \bar{X}, \quad E(\bar{X}) = \theta, \quad \bar{Var}(\bar{X}) = \frac{\theta}{n} = \frac{1}{nI(\theta)}$ tale che $\bar{X}$ è efficiente per θ

$\hat{k}_{ML} = 1 - e^{-\bar{X}}$

$$\begin{aligned} E\left(\hat{k}_{ML}\right) &= 1 - E\left(e^{-\bar{X}}\right) \\ &= 1 - E\left(e^{-\frac{1}{n}(\sum X_j)}\right) \end{aligned}$$

$[\sum X_j \sim Poiss(n\theta)] = M_{Poiss(n\theta)}\left(-\frac{1}{n}\right) \rightarrow [M_x(t) = E(e^{tx})] = e^{n\theta\left(e^{-\frac{1}{n}}-1\right)} \neq \theta \quad \forall\theta$

Lo stimatore di massima verosomiglianza non è efficiente. Quindi per quanto detto finora non esiste alcun stimatore efficiente.

$\lim_{n\rightarrow+\infty} E\left(\hat{k}_{ML}\right) = 1 - e^{-\theta}$ che è la caratteristica. Gli stimatori di verosomiglianza possono essere distorti, ma se sono soddisfatte le condizioni di FCE vale in generale che lo stimatore è asintoticamente non distorto.

16.5 Proprietà stimatori verosomiglianza

1. Se uno stimatore efficiente esiste allora è di verosomiglianza

16.5.1 Proprietà asintotiche

[Prop asintotiche e nel caso di più parametri da stimare] Siano soddisfatte ipotesti di regolarità del modello (FCR, derivabile di ordine tre

1. $\hat{k}_{ML}$ esiste ed è unico con probabilità 1
   (a) Deriva da questo $\frac{d}{d\theta}\ln L_\theta = a(n,\theta)(T-k)$ (perchè?)
   (b) Uno stimatore efficiente può esistere ma su dei punti che hanno probabilità zero (?)
2. $\lim_{n\rightarrow\infty} E\left(\hat{k}_{ML}\right) = k \quad \forall\theta$ quindi asintoticamente non distorto

3. $\lim_{n\to\infty} Var\left(\hat{k}_{ML}\right) = 0$ (2+3 : consistenza)

4. $\lim_{n\to\infty} P\left(\frac{\hat{k}_{ML}-k}{\sqrt{\frac{(k'(\theta))^2}{nI(\theta)}}}\right) = \Phi(z) \quad \forall z \in R$

   (a) Applicazione: Per n grandi $\hat{k}_{ML} \sim N\left(k, \frac{(k'(\theta))^2}{nI(\theta)}\right)$

      i. Cosa mi serve la normalità asintotica? Perchè mi permette di risolvere problemi di stima intervallare e di verifica di ipotesi quando non sono sotto ipotesi di dati gaussiani, ma ho tante osservazioni.

Lo stimatore efficiente è consistente (punti 2 e 3 già verificati)

16.5.2 Intervallo di confidenza di k per grandi campioni

$\gamma \sim P\left(-q < \frac{\hat{k}-k}{\sqrt{\frac{(k'(\theta))^2}{nI(\theta)}}} < q\right)$ sse $q = z_{\frac{1+\gamma}{2}}$, bilatero: $T_1 < k < T_2$.

$X_1, ..., X_n$ iid $f(x, \theta)$

Esempio Poisson $k = \theta \quad \hat{k}_{ML} = \bar{X}$

$-z_{\frac{1+\gamma}{2}} < \frac{\bar{X}-\theta}{\sqrt{\frac{\theta}{n}}} < z_{\frac{1+\gamma}{2}} \iff \theta \in \bar{X} \mp z_{\frac{1+\gamma}{2}}\sqrt{\frac{\bar{X}}{n}}$

$-q < \frac{\hat{k}-k}{\sqrt{\frac{(k'(\theta))^2}{nI(\theta)}}} < q \quad \rightarrow \quad k - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{(k'(\theta))^2}{nI(\theta)}} < \hat{k} < \underbrace{k + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{(k'(\theta))^2}{nI(\theta)}}}_{\text{IC per grandi campioni}}$

17 Tabella sugli intervalli di confidenza

$N(\mu, \sigma^2)$	$\varepsilon(\theta)$	grandi campioni
μ	θ	k
σ^2nota: $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$	$\frac{\bar{X} \cdot 2n}{\theta} \sim \chi^2_n$	$\frac{\hat{k}-k}{\sqrt{\frac{(k'(\theta))^2}{nI(\theta)}}}$
$\frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$		

Quantità pivotali che non dipendono da parametri ma che non sono statistiche (?). Metodo della quantità pivotale? parte da uno stimatore, costruisco una variabile la cui distribuzione non dipende dalla variabile, risolvo gli estremi dell'intervallo; ora inverto per trovare la caratteristica.

18 Metodo della quantità pivotale

$X_1, ..., X_n$ iid $\sim f(x, \theta)$, $k = k(\theta)$ e $x_1, ..., x_n$ come dati

Come stima intervallare intendo $\overbrace{t_1(x_1, ..., x_n) < k < t_2(x_1, ..., x_n)}^{\text{intervallo di confidenza bilatero}}$ sono intevalli numerici, tale che $P_\theta(T_1 < k < T_2) \geq \gamma$. Con γ che è la confidenza. Ho la diseguaglianza invece dell'uguaglianza quando sono i casi in cui le statistiche sono osservazioni discrete (binomiale, poisson, geometrica).

18.1 Metodologie per creare intervalli di confidenza

18.1.1 Metodo della quantità pivotale

Per costruire IC posso usare metodo della quantità pivotale (Q). $Q(\theta, x_1, ..., x_n)$ ma il modello probabilistico che spiega Q non dipende da parametri incogniti

Passi

1. determinare q_1 e q_2 tale che $P(q_1 < Q(\theta, X_1, ..., X_n) < q_2) = \gamma$
2. invento $q_1 < Q(\theta, x_1, ..., x_n) < q_2$ per ottenere $t_1 < k < t_2$

Esempio $X_1, ..., X_n$ iid $\sim \varepsilon(\theta)$ con $\theta > 0$ incognito. Allora $\hat{\theta}_{ML} = \bar{X}$. $IC(\theta)$?

Le quantità pivotali si costruiscono a partire da uno stimatore di cui conosco la distribuzione.

$\bar{X} = \frac{\sum X_j}{n} \quad \varepsilon(\theta) = \Gamma(1, \theta) \rightarrow \frac{\sum X_j}{n} \sim \Gamma\left(n, \frac{n}{\theta}\right)$

$\frac{\bar{X}}{\theta} \sim \Gamma\left(n, \frac{1}{n}\right)$

Cerco di ricondurmi alla chi^2 con due gradi di libertà.

$n \cdot 2 \cdot \frac{\bar{X}}{\theta} \sim \Gamma(n, 2)$ è candidata ideale per il calcolo l'intervallo di confidenza.

Quindi la Q per questo problema è $Q = \frac{2n\bar{X}}{\theta} \sim \chi^2_{2n}$

$\gamma = P\left(q_1 < \frac{2n\bar{X}}{\theta} < q_2\right) = P(q_1 < \chi^2_{2n} < q_2)$

FIG 0

$$\chi^2_{2n}\left(\frac{1-\gamma}{2}\right) < \frac{2n\bar{X}}{\theta} < \chi^2_{2n}\left(\frac{1+\gamma}{2}\right)$$

$$\frac{2n\bar{X}}{\chi^2_{\underbrace{2n}_{n-1}}\left(\frac{1+\gamma}{2}\right)} < \underbrace{\theta}_{\sigma^2} < \frac{2n\bar{X}}{\chi^2_{2n}\left(\frac{1-\gamma}{2}\right)}$$

con $\theta \rightarrow \sigma^2$ e $2n \rightarrow n-1$ ho un caso particolare che abbiamo affrontato in precedenza.

19 Test verifica d'ipotesi

Ho come strumenti gli intervalli di confidenza e gli stimatori per verificare che una congettura sia plausibile o meno.

19.1 Formalmente

Verificare una congettura su una popolazione si traduce in verificare una congettura su un parametro oppure su una caratteristica sull'intera distribuzione della popolazione.

19.2 Esempio

Consideramo un'azienda che produce un certo bene e brevetta un procedimento di costruzione. Se la vita media della cinghia era di 50000 km con il nuovo procedimento si arriva fino a 56000. Ma prima di produrre su larga scala il nuovo prodotto l'azienda si preoccupa di verificare che effettivamente un campione del nuovo prodotto soddisfi le aspettative.

Se la durata media campionaria del nuovo campione supera i 57000 km allora posso lanciare la produzione, se non li supera allora gli ingegneri che hanno prototipato il prodotto non sono così affidabili. La caratteristica su cui faccio congettura è la durata media m_x.

19.2.1 Congettura

Preoccupazione dell'azienda è fondata $\iff$ Durata media non superiore a 56000 Km.

$$\mu \leq 56000$$

$X_1, ..., X_{35}$[5] iid con media incognita $\mu \to \bar{X}$. L'azienda ritiene non fondata la sua preoccupazione se $\bar{X} \geq 57000 \iff$Secondo l'azienda ritiene che effettivamente il congegno funziona, ovvero $\mu > 56000$

L'**ipotesi statistica** è un'affermazione su un parametro che sono finora $\overbrace{\mu \leq 56000}^{H_0}$ e $\overbrace{\mu > 56000}^{H_1}$[6] che sono in contraddizione e va bene perchè devono essere incompatibili perchè i dati mi devono portare a discriminare tra più situazioni incompatibili. Incompatibili però non significa che l'unione delle ipotesi sia l'insieme nel totale.

Le ipotesi statistiche sono congetture su parametri.

Raccolgo dati per falsificare l'ipotesi nulla (H_0). Data l'incertezza dei dati, la decisione finale dipende proprio dal campione che utilizzo. Per questo motivo è importante il terzo elemento per la verifica d'ipotesi, ovvero la **regola per prendere decisione** $\bar{X} \geq 57000$ [7], indica che la media campionaria è significativamente più grande di 56000 (quindi scelgo 57000).

[5] Primo elemento per verifica d'ipotesi
[6] Secondo elemento per verifica d'ipotesi
[7] Terzo elemento per verifica d'ipotesi

19.3 Riassumendo

Una procedura di verifica d'ipotesi si avvia con un test d'ipotesi

- H_0 e H_1
- $X_1, ..., X_n$ iid $\sim f$
- Regola decisionale: regola di decisione di rifiutare H_0.
- $G = \{(x_1, ..., x_n) : \bar{x} \geq 57000\}$ regione di rifiuto di H_0 o **regione critica**.
  - La regione critica (G) è un sottoinsieme dei risultati sperimentali tale che se $x_1, ..., x_n \in G$ rifiuto H_0.

Abbiamo preso la media campionaria perchè la nostra congettura era sulla durata media e sappiamo che la media campionaria è valida sia su molti che su pochi dati a disposizione.

Se invertiamo il problema non è detto che arriviamo alla stessa conclusione. Siccome le ipotesi non giocano un ruolo simmetrico intendo H_0 l'ipotesi che voglio rifiutare.

19.3.1 Come possono essere queste ipotesi?

H può essere su un parametro

1. $H : \mu \leq 56000$ rimane comunque incognito μ
2. $H : X \sim \varepsilon(\theta)$ il parametro rimane incognito
3. $H : X \sim \Gamma(5, 3)$
4. $H : X \sim N(0, 1)$

1 e 2 sono *ipotesi composte*. 3 e 4 sono *ipotesi semplici* perchè caratterizzano completamente la distribuzione della popolazione. 2, 3 e 4 utilizzo statistiche parametriche o distribution free. Nel primo caso sto invece sfruttando la distribuzione sottostante la popolazione.

19.3.2 Esempio

$X_1, ..., X_n$ iid $\sim f(x, \theta_1, \theta_2)$

1. Problema: $H_0 : \theta_1 = 0 \quad H_1 : \theta_1 = 1$
2. Problema: $H_0 : \theta_1 \leq \bar{\theta} \quad H_1 : \theta_1 > \bar{\theta}$
3. Problema: $H_0 : \theta_1 = \bar{\theta} \quad H_1 : \theta_1 \neq \bar{\theta}$
4. Problema: $H_0 : \theta_1 \geq \bar{\theta} \quad H_1 : \theta_1 < \bar{\theta}$

1a, 3a e 3b sono semplici o composte? Dipende da θ_2, perchè se è noto allora sono semplici, ma se è incognito allora sono composte.

19.4 Qual'è la probabilità di prendere la decisione corretta?

	Accettare H_0	Rifiutare H_0
H_0 Vera		Errore (I tipo)
H_0 Falsa	Errore (II tipo)	$\Pi(\theta) = P_\theta(G)$ con $\theta \in \Theta_1$

È importante notare che in qualunque test per verificare un'ipotesi nulla, il risultato può essere sbagliato in due modi differenti. Si ha infatti un *errore di prima specie* quando i dati ci portano a rifiutare una ipotesi H_0 che in realtà è corretta, e un *errore di seconda specie* quando finiamo con l'accettare H_0 ed essa è falsa.

Non vi è simmetria tra i due errori. Ricordiamo infatti che l'obiettivo di una verifica di H_0 non è quello di dire se questa ipotesi è vera o falsa, ma piuttosto di dire se l'ipotesi fatta sia anche solo compatibile con i dati raccolti. In effetti vi è un ampio livello di tolleranza nell'accettare H_0, mentre per rifiutarla occorre che i dati campionari siano molto improbabili quando H_0 è soddisfatta.

Commettiamo un errore di I tipo perchè accetto H_0 e quindi i dati che ho raccolto ricadono in in G. Come posso determinare la probabilità dell'errore di primo tipo?

$$P_{H_0}((x_1, ..., x_n) \in G)$$

- $X_1, ..., X_n \sim f(x, \theta)$

- $H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$
- G

$\alpha(\theta) = P_\theta(G), \theta \in \Theta_0$ è la probabilità di errore di I tipo. Un test con livello di significatività α deve avere una probabilità di errore di prima specie minore o uguale ad α .

$\beta(\theta) = P_\theta = (G^c), \theta \in \Theta_1$ è la probabilità di errore di II tipo.

$$\beta(\theta) = 1 - \alpha(\theta)$$

Quest'ultima equazione è sbagliata perchè possono fare riferimento a due distribuzioni diverse.

$$\beta(\theta_2) = 1 - \alpha(\theta_1)$$

19.4.1 Esempio

$\alpha(\mu) = P_\mu\left(\bar{X} \geq 57000\right) \quad \mu \leq 56000$

$\beta(\mu) = P_\mu\left(\bar{X} < 57000\right) \quad \mu > 56000$

––

$\Pi(\theta) = P_\theta(G), \theta \in \Theta_1 = 1 - \beta(\theta), \theta \in \Theta_1$ è definita funzione di potenza del test.

Il test migliore è quello che ti porta ad avere errori del I e del II tipo più piccolo possibile. Succede però se provo a controllare α allora si alza β e viceversa. Darò una priorità allora a questi due errori minimizzando uno piuttosto che l'altro.

Non riesco a minimizzare contemporaneamente entrambi gli errori allora preferisco quei test che hanno errori del primo tipo inferiore ad una certa soglia che chiamo **significatività del test**.

$$\alpha = sup_{\theta \in \Theta_0} \alpha(\theta)$$

Livello ampiezza del test. Più è basso α più voglio essere sicuro di non commettere errore di primo tipo. I dati devono essere significativamente in contrasto con l'ipotesi H_0.

19.5 Esempio

$X_1, ..., X_n$ iid $\sim N(\mu, 1)$

- $H_0 : \mu \leq 0 \quad H_1 : \mu > 0$
- $G = \{(x_1, ..., x_{25}) : \bar{x} \geq 0.4\}$

$$\alpha(\mu) = P_\mu\left(\bar{X} \geq 0.4\right), \mu \leq 0 \quad = \quad 1 - P\left(N\left(\mu, \tfrac{1}{25}\right) \leq 0.4\right) = 1 - \Phi\left(\frac{0.4-\mu}{\sqrt{\frac{1}{25}}}\right), \mu \leq 0 = 1 - \Phi(2 - 5\mu), \mu \leq 0$$

FIG 1

$$\begin{aligned} \Pi(\mu) &= P_\mu\left(\bar{X} \geq 0.4\right), \mu > 0 \\ &= 1 - \Phi(2 - 5\mu), \mu > 0 \end{aligned}$$

20 Esercitazioni

20.1 Esercizi sugli intervalli di confidenza

$f(x) = \frac{3x^2}{\theta} e^{-\frac{x^2}{\theta}} \quad x > 0$

1. Calcolare $\hat{\theta}_L$
2. Sapere se è efficiente
3. $Y = X^3$
4. IC per θ al 90%

20.1.1 Risoluzione punto 1

$L = \frac{3^n \prod_i x_i^2}{\theta^n} e^{-\frac{1}{\theta}\sum_i x_i^2}$

$\ln L = n \ln 3 + \ln \prod_i x_i^2 - n \ln \theta - \frac{1}{\theta} \sum x_i^3$

$\frac{d}{d\theta} (\ln L) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_i x_i^3 = 0 \Longleftrightarrow \theta = \frac{1}{n} \sum x_i^3$

$$\frac{d^2}{d\theta^2} (\ln \theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum x_i^3 = \underbrace{\frac{1}{\theta^3}}_{>0} \left(\underbrace{n\theta - 2\sum x_i^3}_{<0} \right) < 0$$

$$\hat{\theta}_L = \frac{1}{n} \sum x_i^3$$

20.1.2 Risoluzione punto 2

Solitamente si calcola se uno stimatore è efficiente se la varianza coincide con il limite di FCR. Ma qui c'è una scorciatoia che possiamo prendere.

$\frac{d}{d\theta} (\ln L) = a(\theta) [\tau(\theta) - T]$ se è possibile scrivere questa fattorizzazione sappiamo che T è lo stimatore efficiente.

$\frac{d}{d\theta} (\ln L) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_i x_i^3 = -\frac{n}{\theta^2} \left[\theta - \frac{1}{n} \sum x_i^3\right]$. Qui c'è scritto di più: per ogni densità che è possibile scrivere la fattorizzazione è univoca. Significa che per ciascuna densità c'è un solo oggetto che può essere stimato in maniera efficiente.

θ^2 non può essere stimato in maniera efficiente perchè non riesco a mettere dentro la parentesi quadra un θ^2, quindi non esiste uno stimatore efficiente per θ^2. D'altronde questo lo sapevo già perchè lo stimatore efficiente è unico per l'intera famiglia (?).

20.1.3 Intervalli di confidenza

Utilizzo uno stimatore per stimare il parametro incognito. Tale stimatore lo manipolo al fine di poterlo standardizzare e poter utilizzare tabelle di distribuzioni note. Per esempio $\frac{\bar{X}-n}{\sqrt{\frac{\sigma^2}{n}}}$ non dipende più strettamente dal parametro incognito, ma è una quantità che posso limitare attraverso due quantili (Intervallo di confidenza).

20.1.4 Risoluzione punto 3

$y = x^3 \leftrightarrow x = h(y) = \sqrt[3]{y} \qquad h'(y) = \frac{1}{3\sqrt[3]{y^2}}$

$$\begin{aligned} f_y(y) &= f_x(\sqrt[3]{y}) \cdot \frac{1}{3\sqrt[3]{y^2}} \\ &= \frac{3\left(\sqrt[3]{y^2}\right)}{\theta} e^{-\frac{y}{\theta}} \frac{1}{3\sqrt[3]{y^2}} \\ &= \frac{1}{\theta} e^{-\frac{y}{\theta}} \end{aligned}$$

$Y = X^3 \sim \varepsilon(\theta)$

Con quest'informazioni mi dice che $\hat{\theta}_L \sim \Gamma(n, \theta)$.

$T \quad f_T(t) = \frac{\alpha t^{\alpha-1}}{\theta} e^{-\frac{\alpha t}{\theta}} \quad \alpha > 0$ densità di Weibull. Usate in pratica perchè l'esponenziale è una VA che gode dell'assenza di memoria .

Ripasso Se $T \sim \varepsilon(\theta)$, la VA $V = \frac{2}{\theta} \underbrace{T}_{E(T)=\theta} \sim \varepsilon(2)$

$Q = \frac{2}{n\theta} \sum X_i^3 \sim \Gamma(n, 2) \sim \chi^2_{2n}$

20.1.5 Risoluzione punto 4

$P\left(q_1 < \frac{2}{n\theta} \sum X_i^3 < q_2\right) = 0.9$

$\theta \in \left[\frac{2}{nq_2} \sum X_i^3; \frac{2}{nq_1} \sum X_i^3\right]$ con $q_1 = \chi^2_{2n;0.05}$ e $q_2 = \chi^2_{2n;0.95}$

20.2 Esercizio

$$X \sim Geom(\theta) \rightarrow E[X] = \frac{1}{\theta}, Var[X] = \frac{1-\theta}{\theta^2}$$

$X \sim geom(\theta)$, stimare $\hat{\theta}_L$ e trovare un intervallo di confidenza asintotico al 90% per θ . Questa è una situazioni in cui non ho alcun suggerimento per arrivare ad una distribuzione nota.

Proprietà garantite dello stimatore di massima verosomiglianza.

- Consistenti
- Asintoticamente normali (server per IC)
- ...

Quindi se standardizzo lo stimatore di massima verosomiglianza troverò alla fine una normale.

$f_x(x) = (1-\theta)^{x-1}\theta \quad x \in \{1,2,3,...\}$

$L(\theta) = (1-\theta)^{\sum X_i - n}\theta^n \quad \ln L = (\sum X_i - n)\ln(1-\theta) + n\ln\theta$

$$\begin{aligned} \frac{d}{d\theta}\ln L &= -\frac{\sum X_i - n}{1-\theta} + \frac{n}{\theta} \\ &= \frac{-\theta(\sum X_i - n) + n - n\theta}{\theta(1-\theta)} \\ &= \frac{n - \theta\sum X_i}{\theta(1-\theta)} \end{aligned}$$

$\frac{n-\theta\sum X_i}{\theta(1-\theta)} = 0 \iff \theta = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}$

$\frac{d}{d\theta}\ln L = \frac{n-\theta\sum X_i}{\theta(1-\theta)}$

$$\hat{\theta}_L = \frac{1}{\bar{X}}$$

Visto che utilizziamo l'asintotica normalità, significa che visto che lui è uno stimatore MLE $\hat{\theta}_L = \frac{1}{\bar{X}} \approx N\left(\theta, ? = \frac{1}{nI}\right)$

$$E\left[\left(\frac{n-\theta\sum X_i}{\theta(1-\theta)}\right)^2\right] = \frac{n^2}{(1-\theta)^2} E\left[\left(\underbrace{\frac{1}{\theta}}_{E[X]} - \underbrace{\frac{\sum X_i}{n}}_{\bar{X}}\right)^2\right] = \frac{n^2}{(1-\theta)^2} \overbrace{Var\left(\bar{X}\right)}^{\frac{1-\theta}{n\theta^2}} = \frac{n}{\theta^2(1-\theta)}$$

$$nI = \frac{n}{\theta^2\left(1-\theta\right)}$$

$-z_{0.95} < \frac{\hat{\theta}_L - \theta}{\sqrt{\frac{\theta^2(1-\theta)}{n}}} < z_{0.95}$ c'è il problema che θ è anche al denominatore, quindi eseguo la sostituzione di θ con $\hat{\theta}$. Quindi

$$\theta \in \hat{\theta} \pm z_{0.95}\sqrt{\frac{\hat{\theta}^2\left(1-\theta\right)}{n}}$$

20.3 Esercizio

$X \sim N\left(\mu, \sigma_x^2\right) \quad Y \sim N\left(\mu, \sigma_y^2\right)$ con σ_x^2 e σ_y^2 note. L'incognita è la media. Abbiamo m campioni di X ed n campioni di Y.

$T = \alpha\bar{X} + (1-\alpha)\bar{Y}$ trovare α al fine di minimizzare l'errore quadratico medio. Abbiamo scelto α ed $1-\alpha$ perchè così siamo sicuri che non sia distorto. $E\left(T\right) = \alpha E\left(\bar{X}\right) + (1-\alpha) E\left(\bar{Y}\right) = \mu$

$$\begin{aligned} MSE\left(T\right) &= Var\left(T\right) \\ &= \alpha^2 Var(\bar{X}) + (1-\alpha)^2 Var\left(\bar{Y}\right) \\ &= \alpha^2\frac{\sigma_x^2}{m} + ... \\ &= \alpha^2\left(\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right) - 2\alpha\frac{\sigma_y^2}{n} + \frac{\sigma_y^2}{n} \end{aligned}$$

$\alpha = \frac{\frac{\sigma_y^2}{n}}{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}$

20.4 Esercizio

$f_T(t) = \frac{1}{6^{\frac{1}{\theta}}} \cdot \frac{1}{2\theta} t^{\frac{1}{2\theta}-1} 1_{[0,36]}(t)$

$\sum_{i=1}^{169} \ln \sqrt{T_i} = -35.49$

Calcolare $\hat{\theta}_L$ e $\hat{k}_L$ e $k = P(T > 6)$

20.4.1 Risoluzione

$$\begin{aligned} P(T>6) &= 1 - \int_0^6 \frac{1}{6^{\frac{1}{\theta}}} \cdot \frac{1}{2\theta} t^{\frac{1}{2\theta}-1} dt \\ &= 1 - \frac{1}{6^{\frac{1}{\theta}}} \left[t^{\frac{1}{2\theta}} \right]_0^6 \\ &= 1 - \frac{6^{\frac{1}{2\theta}}}{6^{\frac{1}{\theta}}} \\ &= 1 - 6^{-\frac{1}{2\theta}} \end{aligned}$$

Ottengo quindi che $k = 1 - 6^{-\frac{1}{2\theta}}$

$L = 6^{-\frac{n}{\theta}} \cdot \frac{1}{2^n \theta^n} \left(\prod_i t_i\right)^{\frac{1}{2\theta}-1}$

$\ln L = -\frac{n}{\theta} \ln 6 - n \ln 2 - n \ln \theta + \frac{1}{2\theta} \ln \prod t_i - \ln \prod t_i$

$$\begin{aligned} \frac{d}{d\theta} \ln L &= \frac{n}{\theta^2} \ln 6 - \frac{n}{\theta} - \frac{1}{2\theta^2} \ln \prod t_i \\ &= \frac{1}{\theta^2} \left(n \ln 6 - n\theta - \frac{1}{2} \ln \prod t_i \right) \end{aligned}$$

$\frac{1}{\theta^2} \left(n \ln 6 - n\theta - \frac{1}{2} \ln \prod t_i \right) = 0 \iff \theta = \ln 6 - \frac{1}{n} \sum \ln \sqrt{t_i}$

$\hat{\theta}_L = \ln 6 - \frac{1}{n} \sum \ln \sqrt{T_i}$

$\hat{k} = 1 - 6^{-\frac{1}{2\hat{\theta}_L}}$

20.5 Esercizio

$X \sim Unif\,(0, \theta)$.

21 Lemma di Neyman-Person e generalizzazione

- $X_1, ..., X_n$ iid $f\,(x, \theta)$
- $H_0 : \theta \in \Theta_0$
- G regione critica
- $\alpha\,(\theta) = P_\theta\,(G)\,, \theta \in \Theta_0$
- $\beta\,(\theta) = P_\theta\,(G^c)\,, \theta \in \Theta_1$
- $T\,(\theta) = 1 - \alpha\,(\theta)$
- $\alpha = \sup_{\theta \in \Theta_0} P_\theta\,(G)$

21.1 Obiettivo in generale

L'obiettivo è quello di determinare fra i test di ampiezza $\leq \alpha$ (α prefissato) un test con $p\,(\theta)$ minimo $\forall \theta \in \Theta_1$. Minimizzare β significa massimare la funzione di potenza. Cioè un test che abbia Π massima $\forall \theta \in \Theta_1 =:$ test uniformemente più potente fra i test di ampiezza α.

21.2 Lemma di Neyman-Pearson

21.2.1 Ipotesi

$X_1, ..., X_n$ campione causale con verosomiglianza $L_\theta = (x_1, ..., x_n)$

- $H_0 : \theta = \theta_0$ (ipotesi semplice)
- $H_1 : \theta = \Theta_1$ (ipotesi semplice)

Il test più potente è fornito dal lemma di Neyman-Pearson

21.2.2 Definizione

$$G = \left\{ (x_1, ..., x_n) : \frac{L_\theta (x_1, ..., x_n)}{L_{\theta_1} (x_1, ..., x_n)} \leq \delta \right\}$$

allora il test con regione critica G ha potenza massima rispetto a ogni regione F di ampiezza $\leq$ l'ampiezza di G. La verosomiglianza sotto θ è significativamente più piccola (dipende da δ) della verosomiglianza sotto θ_1. Rifiuto H_0, quando la probabilità di ciò che ho effettivamente osservato sotto θ sia minore rispetto alla probabilità di ciò che ho osservato sotto θ_1.

21.2.3 Dimostrazione

Osservazione:

$A \subseteq G : L_{\theta_0} \leq \delta L_{\theta_1}$

$$\begin{aligned} P_{\theta_0} (A) &= \int_A f (x_1, ..., x_n, \theta_0) \, dx_1 \cdot dx_2 \cdot ... \cdot dx_n \\ &= \int_A L_{\theta_0} dx_1 \cdot dx_2 \cdot ... \cdot dx_n \leq \delta \int_A L_{\theta_1} = \delta P_{\theta_1} (A) \end{aligned}$$

$P_{\theta_0} \leq \delta P_{\theta_1} (A) \quad B \subseteq G^c : P_{\theta_0} (B) \geq \delta P_{\theta_1} (B)$

Dimostrazione: Ipotesi:

[FIG 1]

- α di $F = P_{\theta_0} (F)$ ho eliminato il sup perchè tanto l'ipotesi è semplice. $P_{\theta_0} (F) \leq P_{\theta_0} (G) \iff P_{\theta_0} (F \cap G^c) \leq P_{\theta_0} (F^c \cap G)$

Tesi:

- $\Pi_F \leq \Pi_G$

- $P_{\theta_1}(F) \leq P_{\theta_1}(G) \Longleftrightarrow \mathbf{P_{\theta_1}(F \cap G^c) \leq P_{\theta_1}(F^c \cap G)}$

$$\begin{aligned} P_{\theta_1}\left(\underbrace{F \cap G^c}_{B}\right) &\leq \frac{1}{\delta} P_{\theta_0}(F \cap G^c) \\ &\leq \frac{1}{\delta} P_{\theta_0}\left(\underbrace{F^c \cap G}_{A}\right) \\ &\leq \frac{1}{\delta} \delta P_{\theta_1}(F^c \cap G) \end{aligned}$$

Applichiamo adesso il lemma per costruire il test d'ipotesi migliore per una popolazione gaussiana con varianza nota e media incognita.

21.3 Verifica d'ipotesi popolazione gaussiana

$X_1, ..., X_n$ iid $\sim N(\mu, \sigma^2)$

$$\begin{aligned} L_{\mu,\sigma^2}(x_1, ..., x_n) &= \prod_{i=1}^{n} f\left(x, \mu, \sigma^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \\ &= (n-1)S^2 + n(\bar{x} - \mu)^2 \end{aligned}$$

Verifica d'ipotesi su μ con σ^2 nota.

$H_0 : \mu = \mu_0 \quad H_1 : \mu = \mu_1$

Test più potente: quello di NP

$$\begin{aligned}
\frac{L_{\mu_0,\sigma^2}}{L_{\mu_0,\sigma^2}} &= e^{\left\{-\frac{n(\bar{x}-\mu_0)^2+n(\bar{x}-\mu)}{2\sigma^2}\right\}} \\
&= \underbrace{e^{\left\{\overbrace{\frac{1}{2\sigma^2}\left[-n\mu_0^2+n\mu_1^2\right]}^{\text{pezzi che non dipende da x}}\right\}}}_{c>0} e^{\left\{\frac{2n\bar{x}\mu_0-2n\bar{x}\mu_1}{2\sigma^2}\right\}} \\
&= c \cdot e^{\left\{\frac{n\bar{x}}{\sigma^2}(\mu_0-\mu_1)\right\}}
\end{aligned}$$

Ipotizziamo che $\mu_1 < \mu_0$

$\frac{L_{\mu_0}}{L_{\mu_1}}$ cresce al crescere di $\bar{x}$. Se vale questo, quand'è allora che $\frac{L_{\mu_0}}{L_{\mu_1}} \leq \delta \Longleftrightarrow \bar{x} \leq k$ per quale k e quale δ ?

[FIG 2]

Tutto dipende dalla significatività del test. A questo punto entra in gioco α.

$\alpha = \alpha(\mu_0) = P_{\mu_0}(\text{rif } H_0) = P_{\mu_0}(\bar{x} \leq k)$

Sotto H_0 ho che $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ e quindi $\alpha = \Phi\left(\frac{k-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \Longrightarrow \frac{k-\mu_0}{\frac{\sigma}{\sqrt{n}}} = z_\alpha = -z_{1-\alpha} \Longrightarrow k = \mu_0 - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$

Il test più potente di livello α ha regione critica $G = \left\{(x_1, ..., x_n) : \bar{x} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}\right\}$ rifiuto ... quando la media campionaria è più piccola di μ_0 che dipende dal numero di osservazioni.

Se $\mu_1 > \mu_0$ rif H_0 se $\bar{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$.

$\Longrightarrow$poichè G non dipende dal valore μ_1 ma solo dal fatto che $\mu_1 < \mu_0$, allora G è la migliore regione (nel senso che mi massimizza la potenza) $\forall \mu_1 < \mu_0$, quindi uniformemente in $\mu_1 < \mu_0$. Quindi posso usare questa G per verificare $H_0 : \mu = \mu_0 \quad H_1 : \mu < \mu_0$ Gratuitamente ho ottenuto anche il miglior test per un'ipotesi semplice contro una composta.

$--$

Se avessi il caso $H_0 = \mu \geq \mu_0 \quad H_1 : \mu < \mu_0$ si può dimostrare che se utilizziamo sempre quella regione G, allora ho costruito quella più potente tra tutti i possibili test? No, non esiste, ma è la più potente tra quelle che hanno funzioni di potenza maggiori a quella di α e sono non distorti.

Se cerco il tst migliore, mi devo limitare con quelli con una certa ampiezza ed in questo modo risolvo il problema solo per alcuni problemi (sto parlando di modelli gaussiani addirittura). Anche se questo speciale modello, modello unilatero contro unilatero, il test migliore lo devo cercare in una sott'area (di α?)

La regione ricavata in questa dimostrazione è quella con cui avremo più a che fare. $G = \left\{(x_1, ..., x_n) : \bar{x} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} \cdot z$ è la regione più potente $\forall \mu < \mu_0$ fra le regioni di test **non distorti**.

La media campionaria svolge il ruolo di statistica test. Questa statistica complicata (il rapporto tra verosomiglianza).

21.4 Test di ... verosomiglianza generalizzato (rispetto NP)

$H_0 : \theta \in \Theta_0 \quad H_1 : \theta_1 \in \Theta_1$ caso particolare $H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$ e $H_0 : \theta \leq \theta_0 \quad H_1 : \theta > \theta_0$

$\Lambda = \frac{\sup_{\theta \in \Theta_0} L_\theta(x_1,...,x_n)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L_\theta(x_1,...,x_n)}$

Rif H_0 se $\Lambda \leq \delta$ con $\delta : \alpha = \sup_{\theta \in \Theta_0} P_\theta (\Lambda \leq \delta)$

$\Lambda = \frac{L_{\theta_0}}{\sup_{\theta \in \Theta} L_\theta} = \frac{L_{\theta_0}}{L_{\hat{\theta}_{ML}}}$

Nello stimatore vado a cercare il punto in cui la funzione ha valore massimo, qui mi serve sapere il valore massimo. Quello che è stato fatto per la teoria della stima è utilizzato anche per la verifica d'ipotesi.

Ultimo passaggio (per arrivare alla normale)

$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$

$\frac{L_{\mu_0}}{L_{\hat{\theta}_{ML}}} = \frac{L_{\theta_0}}{L_{\bar{x}}} = e^{\left\{-\frac{n(\bar{x}-\mu_0)^2}{2\sigma^2}\right\}} \leq \delta$

$\frac{n}{\sigma^2}(\bar{x} - \mu_0)^2 \geq k \Longrightarrow \left(\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}}\right)^2 \geq k \Longrightarrow \frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}} \geq q$

Rifiuto (rif) a livello α $\mu = \mu_0$ a favore di $\mu \neq \mu_0$ se $\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}} \geq q$ con $q : \alpha = \sup_{\mu=\mu_0} P_\mu \left(\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}} \geq q \right) =$

$$P_{\mu_0} \left(\underbrace{\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}}}_{\sim N(0,1) \text{ sotto } H_0} \geq q \right)$$

quest'ultima equazione è rappresentata graficamente in [FIG 3]

$q = z_{a-\frac{\alpha}{2}}$

Se la varianza è incognita, dovremo usare il test del rapporto di verosomiglianza... si arriva a regioni della stessa forma con la differenza che dove con σ^2 prendo S^2 e lavorerò con quantili della t di student al posto di z.

22 Esercitazioni

$X \sim$lognormale (μ, σ^2) se $\ln X \sim N(\mu, \sigma^2)$

$f_x(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{[\ln(x-\mu)]^2}{2\sigma^2}}$

$X_1, ..., X_n \quad \mu = 0 \quad \sigma^2$ incognito

22.1 Funzione di verosomiglianza per trovare σ^2

$$\begin{aligned} L(\sigma^2) &= \frac{1}{\pi x_i (2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum[\ln(x_i)]^2}{2\sigma^2}} \\ &= -\ln(\pi x_i) - \frac{n}{2}\ln 2\pi - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum[\ln x_i]^2 \end{aligned}$$

$$\begin{aligned}
\frac{d}{d\sigma^2} &= -\frac{n}{2}\cdot\frac{1}{\sigma^2}+\frac{1}{2\left(\sigma^2\right)^2}\sum\left(\ln x_i\right)^2 \\
&= \frac{-n\sigma^2+\sum\left(\ln x_i\right)^2}{2\left(\sigma^2\right)^2} \\
0 &= \frac{-n\sigma^2+\sum\left(\ln x_i\right)^2}{2\left(\sigma^2\right)^2} \\
\sigma^2 &= \frac{1}{n}\sum\left(\ln x_i\right)^2
\end{aligned}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum\left(\ln x_i\right)^2$$

stimatori massima verosomiglianza, proprietà:

- Asintoticamente non distorto: il suo valore atteso per $n \to \infty$ tende al parametro incognito. L'unico modo è attraverso il calcolo della media.
- consistente
- asintoticamnte normale

$E\left(\hat{\sigma^2}\right) = \frac{1}{n}E\left[\sum\left(\ln x_i\right)^2\right] = E\left[\left(\ln x_i\right)^2\right] = \sigma^2$

È efficiente? Sì, da calcolare

– –

$\sum_{n=1}^{55}\left(\ln x_i\right)^2 = 36440$, dobbiamo trovare un intervallo di confidenza per σ^2 al 95%. Ci sono due strade diverse che portano a risultati leggermente differenti, ma comunque validi.

1. Lo stimatore di massima verosomiglianza è asintoticamente normale.
2. Il dato fornito è una somma di quadrati di normali, quindi lavoro con delle χ^2.

23 Introduzione al test d'ipotesi

23.1 Esempi reali

23.1.1 Giudice, l'innocente e le prove

- Perchè faccio un test? per prendere una decisione
- Esempio di test d'ipotesi: l'imputato è innocente fino a prova contraria. l'ipotesi nulla è vera fino a prova contraria, i dati servono a decidere al di là di ogni ragionevole dubbio l'ipotessi nulla è da rifiutare.
- Chi è α? Quanto il giudice è predisposto a condannare l'innocente. Se il giudice è estremamente cauto avrà bisogno di prove molto evidenti, ma in questo modo ci sarà più probabilità di commettere errore di secondo tipo (assolvere un colpevole).
- L'errore di primo grado è quello di condannare l'imputato quando è innocente.
- Cos'è il p-value? è il minimo α per cui inizio ad accettare l'ipotesi alternativa.

Il giudice deve valutare se condannare o rilasciare l'imputato. Secondo la legge un cittadino è innocente fino a prova contraria, significa che l'ipotesi nulla (H_0) è innocente. L'accusa deve raccogliere delle prove $(x_1, ..., x_n)$ per dimostrare che l'imputato è colpevole.

Il giudice in base alle prove raccolte decide se accettare o rifiutare l'ipotesi nulla, la sua decisione è influenzata dalla sua personalità (α): se estremamente cauto avrà bisogno di prove consistente perchè non vuole condannare un innocente e sarà quindi molto tollerante (α piccolo); se tende invece a condannare l'imputato (α grande) allora incorrerà più probabilmente nell'errore di primo grado.

In base alla personalità del giudice, infatti, le prove possono essere o non sufficienti a condannare l'imputato

23.1.2 Lo studente e la febbre: giornata universitaria vs giornata in vacanza

- [giornata da studente] mi sveglio la mattina è penso di star bene, ci sono dati statistici (dolore alla schiena, mal di testa, ecc..) che mi fanno pensare che magari ho la febbre.
- la statistica test è il valore del termometro
- Sopra al 37 rifiuto l'ipotesi di star bene e mi ritengo ammalato

- [giornata di vacanza] stessa situazione ma il livello di significatività cambia. Il livello di significatività lo prendo più piccolo (regione critica ristretta) e devo essere sicuro di essere veramente ammalato prima di perdermi un giorno di vacanza.

23.2 Conclusione degli esempi

Abbiamo un ipotesi di partenza, dei dati che la contraddicono. Dobbiamo valutare se questi dati sono sufficienti e forti sufficienti per rifiutare l'ipotesi nulla. Se scelgo l'ipotesi alternativa è perchè sono molto convinto di questa scelta; scelgo l'ipotesi nulla perchè non ho prove sufficienti a dimostrare il contrario, è una scelta debole.

23.3 Esercizio

- $X \quad f_x(x) = 2\theta x (1-x^2)^{\theta-1} \quad 0 < x < 1$
- $H_0 : \theta = 1$
- $H_1 : \theta = 10$
- Costruire il test più potente di livello $\alpha = 4\%$ sulla base di un'unica osservazione x_1

Le ipotesi a disposizione sono semplici, pertanto per il lemma di NP il test più potente è il rapporto delle funzioni di verosomiglianza e rifiuto l'ipotesi nulla se quel rapporto è più piccolo di un δ che è legato in un qualche modo al livello di significatività.

Nel nostro caso $\theta_0 = 1$ e $\theta_1 = 10$. Quanto vale $\frac{L_{\theta_0}}{L_{\theta_1}} = \frac{2\theta_0 x\left(1-x^2\right)^{\theta_0-1}}{2\theta_1 x(1-x^2)^{\theta_1-1}} = \frac{1}{10\cdot(1-x^2)^9}$. Rifiuta H_0 se $\frac{1}{10(1-x^2)^9} < \delta$. Quanto vale δ ? In qualche modo è legato al 4%. Il livello di significativià è la massima probabilità che ci promettiamo di commettere un errore del primo tipo, che consiste nel rifiutare H_0 con H_0 vero. Se H_0 sappiamo come è distribuita la X. Riusciamo ad esprimere il rapporto in termini di X ?

$(1-x)^9 > \frac{1}{10\delta} \rightarrow (1-x^2) > \sqrt[9]{\frac{1}{10\delta}} \rightarrow x^2 < 1 - \sqrt[9]{\frac{1}{10\delta}}$

$\frac{1}{10(1-x^2)^9} < \delta \Longleftrightarrow x < \sqrt{1-\sqrt[9]{\frac{1}{10\delta}}}$ abbiamo costruito il test e sappiamo che è il più potente perchè abbiamo seguito le indicazioni di NP. Rifiuto H_0 se la mia singola osservazione $x < \sqrt{1-\sqrt[9]{\frac{1}{10\delta}}}$. Questa è la regola, che dobbiamo metterla insieme ad α ed al fatto che l'ipotesi sia vera.

$$\alpha = P\left(\text{rif } H_0 | H_0 \text{ vera}\right) = P_{\theta=1}\left(\text{rif } H_0\right) = P_{\theta=1}\left(X < \sqrt{1 - \sqrt[9]{\frac{1}{10\delta}}}\right)$$

$\theta = 1 \rightarrow f_x\left(x\right) = 2x$

$$F_x\left(x\right) = \begin{cases} 0 & x = 0 \\ x^2 & 0 < x < 1 \\ 1 & x > 1 \end{cases}$$

$$P_{\theta=1}\left(X < \sqrt{1 - \sqrt[9]{\frac{1}{10\delta}}}\right) \overbrace{=}^{\text{def F(x)}} \left(\sqrt{1 - \sqrt[9]{\frac{1}{10\delta}}}\right)^2$$

$$= \quad 1 - \frac{1}{\sqrt[9]{10\delta}}$$

$$\alpha = 1 - \frac{1}{\sqrt[9]{10\delta}}$$

Rifiuto H_0 se $x < \sqrt{\alpha} \rightarrow x < 0.2$

– –

Calcolara probabilità di errore di secondo tipo?

$$\begin{aligned} P\left(\text{non rif } H_0 | H_0 \text{ falsa}\right) &= P_{\theta=10}\left(X > 0.2\right) \\ &= \int_{0.2}^{1} 20x\left(1 - x^2\right)^9 dx \\ &= \left[-\left(1 - x^2\right)^{10}\right]_{0.2}^{1} \\ &= 0.96^{10} \\ &= 0.66 \end{aligned}$$

- Se $\theta = 10$ qual'è la probabilità di prendere la decisione corretta $1 - 66\% = 34\%$.
- Se la nostra osservazione vale $x = 0.5$ quanto vale di p-v? Intanto con questi valori non riufito l'ipotesi nulla perchè x non è sufficientemente piccolo. Il p-value è più ... del 4%
  - Il p-value è α tale per cui cambio decisione con i dati in possesso
  - Il livello di signif è il massimo rischio che voglio correre nel incorrere nell'errore del primo tipo. Con $x = 0.5$ ho una probabilità di commettere errore di primo tipo più alta (7%?)
  - Se non rifiuto H0 significa che il rischio è troppo elevato.

Come facciamo a calcolare il p-v? [FIG 4] È quel particolare valore di α per cui la statistica test diventa il confine della regione critica.

$G = \{x < \sqrt{\alpha}\}$

$\alpha = 4\% \quad G = \{x < 0.2\}$

In questo caso il p-v $= 0.5^2 = 0.25$

23.4 Esercizio

- $\theta(\theta+1)x^{\theta-1}(1-x) \quad 0 < x < 1$ L'ing informatico non si ricorda se il valore di θ è 1 o 10.
- $H_0 : \theta = 1$
- $H_1 : \theta = 10$
- Costruire il test più potente di grado α

$\frac{L_{\theta_1}}{L_{\theta_{10}}} = \frac{2(1-x)}{110x^9(1-x)} < \delta \rightarrow \frac{2}{110x^9} < \delta \rightarrow x > \sqrt[9]{\frac{1}{55\delta}}$

Rifiuto H_0 se $x > \clubsuit$

$\alpha = P_{\theta=1}(X > \clubsuit) = \int_{\clubsuit}^{1} 2(1-x)\,dx = \left[-(1-x)^2\right]_{\clubsuit}^{1} = (1-\clubsuit)^2$

$\alpha = (1-\clubsuit)^2 \rightarrow$Rifiuto H_0 se $x > \clubsuit = 1 - \sqrt{\alpha}$ Questa è la regola ora

con $\alpha = 2.5\%$ e $x = 0.88 \rightarrow 0.88 \overset{?}{>} 1 - \sqrt{0.025}$ Sì, quindi rifiutiamo H_0.

– –

Calcolare la probabilità di errore di secondo tipo.

$P_{\theta=10}\left(X < 1 - \sqrt{0.025}\right) = \int_0^{1-\sqrt{0.025}} 110x^9 (1-x)\, dx = \left[110^{10} - 10x^{11}\right]_0^{1-\sqrt{0.025}}$

23.5 Esercizio 4.1.3

23.5.1 Punto 1

$F_a(x) = \int_0^x 2xe^{-x^2} dx = ... = 1 - e^{-2x^2} \quad x \geq 0$

$F_b(x) = \int_0^x 20xe^{-10x^2} dx = -\left[e^{-10x^2}\right]_0^x = -\left(e^{-10x^2} - 1\right) = 1 - e^{-10x^2} \quad x \geq 0$

23.5.2 Punto 2

- $H_0 : f = f_a \quad H_1 : f = f_b$
- $\alpha \in [0, 1]$

$\alpha = P_{H_0}(X \in G)$

Lemma di Neyman-Pearson: Rifiuto H_0 se $\frac{L_{f_a}(x)}{L_{f_b}(x)} < \delta$

$$\begin{aligned}
\alpha &= P_{H_0}\left(\frac{L_{f_a}(x)}{L_{f_b}(x)} < \delta\right) \\
&= P_{H_0}\left(\frac{2xe^{-x^2}}{20xe^{-10x^2}} < \delta\right) \\
&= P_{H_0}\left(\frac{1}{10e^{-9x}} < \delta\right) \\
&\overset{*}{=} P_{H_0}\left(X < \sqrt{\frac{\ln(10\delta)}{9}}\right) \\
&= F_{f_a}\left(\sqrt{\frac{\ln(10\delta)}{9}}\right) \\
&\overbrace{=}^{1-e^{-2x^2}} 1 - e^{-2\left(\sqrt{\frac{\ln(10\delta)}{9}}\right)^2} \\
\alpha &= 1 - e^{-2\frac{\ln(10\delta)}{9}} \\
&= 1 - e^{\ln(10\delta)^{-\frac{2}{9}}} \\
&= 1 - (10\delta)^{-\frac{2}{9}}
\end{aligned}$$

* Ricavo $x = f(\delta) \to x < \pm\sqrt{\frac{\ln(10\delta)}{9}} \overbrace{\to}^{1_{(0,\infty)}(x)} \sqrt{\frac{\ln(10\delta)}{9}}$ è il test più potente grazie alle indicazioni NP. α è la probabilità che $f = f_a$ quando x è molto piccolo, ovvero $< \sqrt{\frac{\ln(10\delta)}{9}}$

Dopo aver trovato $x = f(\delta)$ devo ricavare $\delta = f(\alpha)$ per ottenere $x = f(\alpha)$

$$\delta = \frac{10}{(1-\alpha)^4}$$

$$x < \sqrt{\frac{\ln(10\delta)}{8}} = \sqrt{\frac{\ln\left(10\frac{10}{(1-\alpha)^4}\right)}{8}}$$

$\alpha = P_{H_0}(x < \gamma) = F_x(\gamma) = 1 - e^{-\gamma^2} \iff \gamma = \sqrt{-\ln(1-\alpha)}$

Rifiuto H_0 al livello α se $x < \sqrt{-\ln(1-\alpha)}$

Riepilogo L'obiettivo è trovare $x = f(\alpha)$, per raggiungere questo obiettivo devo utilizzare una variabile intermedia δ che posso sfruttare grazie al lemma di Neyman-Pearson. Sapendo che $\alpha = P_{H_0}(\text{lemma}) = F(x)$ ricavo prima la correlazione $x = f(\delta)$ e poi $\delta = f(\alpha)$ da cui infine ottengo $x = f(\alpha)$

23.5.3 Punto 3

Il p-v è il valore osservato della probabilità di commettere un errore di primo tipo. Se dal punto 2 $x < 0.23$ con $\alpha = 5\%$, vuol dire che con un $x = 0.15$ il p-value sarà più piccolo.

$x = 0.15 \quad p - v = 1 - e^{-0.15^2} \sim 2\%$

il p-value è quanto rischio di rifiutare H_0 con i dati osservati, viene a posteriori, mentre α è un parametro che si imposta prima di eseguire il test. α è la probabilità di commettere un errore di primo tipo a prescindere dai dati che avrò, perchè oltre è un costo che non posso permettermi, il p-v è la probabilità di commettere un errore di primo tipo in base ai dati osservati.

$x = 0.49 \quad p - v = 1 - e^{-0.49^2} \sim 20\%$ difficilmente rifiuterò H_0 perchè è troppo rischioso rifiutarla. Se rifiuto H_0 avrò un errore del 20%.

A monte di tutto il ragionamento c'è se il rischio è basso, che a priori è α, mentre a posteriori è il p-v.

23.5.4 Punto 4

- $H_0 : f = f_b \quad H_1 : f = f_a$
- $\alpha \in [0, 1]$

$\alpha = P_{H_0}(X \in G)$

- $\Lambda = \frac{20xe^{-20x^2}}{2xe^{-x^2}} = 10e^{-9x^2}$
- Rifiuto H_0 se $10e^{-9x^2} < \delta \iff x > \gamma$
- $\alpha = P_{H_0}(X > \gamma) = 1 - F_x(\gamma) = e^{-10\gamma^2} \iff \gamma = \sqrt{-\frac{1}{10}\ln\alpha}$
- Rifiuto H_0 se $x > \sqrt{-\frac{1}{10}\ln\alpha}$

Osservare come scambiando l'ipotesi nulla con quella alternativa i risultati finali siano differenti.

23.5.5 Riepilogo

Esempio con i due velocisti nei 100 metri piani. L'ipotesi nulla è molto favorita

Esempio della mucca pazza, allevatore e veteriano che deve controllare la salute della mucca.

24 Verifica d'ipotesi

Un secondo modo per ottenere il test di verifica d'ipotesi è tramite gli intervalli di confidenza.

Campione di osservazioni unidimensionali contro campioni causali di dati che sono bidimensionali (sondaggi su aziende).

24.1 p-value

- $H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$
- Statistica test T
- Rifiuto H_0 se $T \geq k$, un k tale che il test di significatività abbia valore α
- [FIG 1]
- Sia $t = T(x_1, ..., x_n)$
- p-value: $\bar{P} = P_{H_0}(T \geq t)$ il valore del p-value è riportata nel formulario nel caso di test notevoli. Quando avrei esercizi sul calcolo del p-value bisogna ricavarlo con queste formule.

Il p-v è una statistica.

$\bar{P} =$? come si distribuisce sotto H_0? Devo calcolare $F_{\bar{P}_{H_0}}(p) = P_{H_0}\left(\bar{P} \leq p\right) = ...$

Se i dati sono continui allora

$$F_{\bar{P}_{H_0}}(p) = \begin{cases} 0 & p \leq 0 \\ p & 0 < p < 1 \\ 1 & p \geq 1 \end{cases} \Longrightarrow f_{\bar{P}_{H_0}}(p) = \begin{cases} 1 & p < 1 \\ 0 & \text{altrove} \end{cases}$$

- $H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_1$
- $P_{\theta_0}\left(\bar{P} \geq \underbrace{0.999}_{\text{accetto H0}}\right) = 1 - 0.999 = 0.001$
- $P_{\theta_0}\left(\bar{P} \leq \underbrace{0.001}_{\text{rifiuto H0}}\right) = 1 - 0.999 = 0.001$

Ragionare sulla statistica test (α) è uguale al ragionare con il $\bar{P}$, dal punto di vista grafico uno ragiona sulle x l'altro sulle y.

24.2 [IC] Test sulla varianza di popolazioni gaussiana

- $X_1, ..., X_n \sim N(\mu, \sigma^2)$
- $H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \sigma^2 \neq \sigma_0^2$
- Considero il caso μ incognito
- La tecnica già presentata negli intervalli di confidenza mostra che sulla media p robusta, non lo è per quanto riguarda la varianza (?)
- Caso Normale

$IC(\sigma^2)$ bilatero $\gamma\,(90\%)$

$$\frac{S^2(n-1)}{\chi^2_{n-1;\frac{1+\gamma}{2}}} < \sigma^2 < \frac{S^2(n-1)}{\chi^2_{n-1;\frac{1-\gamma}{2}}}$$

Al 90% la varianza appartiene a quell'intervallo.

Se $\sigma^2 \notin IC(\sigma^2)$ può essere indicatore del fatto che σ_0^2 non è il vero valore della varianza, posso utilizzare questa regola per rifiutare H_0. Rifiuta $H_0 : \sigma^2 = \sigma_0^2$ a favore di $\sigma^2 \neq \sigma_0^2$ se

$$\sigma_0^2 \notin IC\left(\sigma^2\right) \Longleftrightarrow G : \left\{(x_1, ..., x_n) : \frac{s^2(n-1)}{\chi^2_{n-1;\frac{1+\gamma}{2}}} \geq \sigma_0^2 \text{ oppure } \frac{s^2(n-1)}{\chi^2_{n-1;\frac{1-\gamma}{2}}} \leq \sigma_0^2\right\}$$

$$\sigma_0^2 \notin IC\left(\sigma^2\right) \Longleftrightarrow G : \left\{(x_1, ..., x_n) : \frac{s^2(n-1)}{\sigma_0^2} \leq \chi^2_{n-1;\frac{1+\gamma}{2}} \text{ oppure } \frac{s^2(n-1)}{\sigma_0^2} \geq \chi^2_{n-1;\frac{1-\gamma}{2}}\right\}$$

Ho scoperto che la statistica test del problema è $\frac{S^2(n-1)}{\sigma_0^2}$.

$$\begin{aligned} \alpha &= P_{\sigma_0^2}(G) \\ &= P_{\sigma_0^2}\left(\sigma_0^2 \notin IC\right) \\ &= 1 - P_{\sigma_0^2}\left(\sigma_0^2 \in IC\right) \\ &= 1 - \gamma \end{aligned}$$

24.2.1 Formalizzazione

- $H_0 : k = k_0 \quad H_1 : k \neq k_0$
- $IC(k)$ bilatero $\gamma : T_1 < k < T_2$
- Rifiuto k_0 se $k_0 \leq T_1$ oppure $k_0 \geq T_2$ e livello $\alpha = 1 - \gamma$

24.3 ...

1. $H_0 : \sigma^2 \leq \sigma_0^2 \quad H_1 : \sigma^2 > \sigma_0^2$
2. $H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \sigma^2 > \sigma_0^2$
3. $H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \sigma^2 = \sigma_0^2$ con $(\sigma_1^2 > \sigma_0^2)$

$IC(\sigma^2), \gamma = 1 - \alpha = 95\%$ del tipo $(c, +\infty) = \left(\frac{S^2(n-1)}{\chi^2_{n-1;\gamma}}, +\infty\right)$ Sono sicuro al 95% che il "vero" valore di $\sigma^2 > \frac{S^2(n-1)}{\chi^2_{n-1;95\%}}$

Se $\sigma_0^2 \leq \frac{S^2(n-1)}{\chi^2_{n-1;\gamma}}$ cioè $\sigma_0^2 \notin IC$ del tipo $(c, +\infty)$ allora

1. ogni σ^2specificato da H

[FIG 2]

Questo test ha livello $\alpha = 1 - \gamma$ (5%) e la $ST = \frac{S^2(n-1)}{\sigma_0^2}$

$$G = \left\{ (x_1, ..., x_n) : \frac{S^2 (n-1)}{\sigma_0^2} \geq \chi^2_{n-1;(1-\alpha)} \right\}$$

valido per i punti 1. 2. e 3.

24.4 Caso μ nota

- $H_0 : \sigma^2 \in S_0 \quad H_1 : \sigma^2 \in S_1$
- $S_0 \cap S_1 = 0$
- Vedere dispense per il seguito

25 Verifica di ipotesi per dati gaussiani accoppiati

- Potrebbe essere un questionario e X, Y sono le risposte
- Le coppie sono indipendenti ed hanno la stessa densità congiunta $(X_1, Y_1), ..., (X_n, Y_n) \sim f(x, y, \theta)$
- Voglio confrontare x con y per vedere se seguono lo stesso modello unidimensionale (test di omogeneità) oppure il problema che mi pongo è se x ed y sono indipendenti o meno. Se lo sono posso prima chiedere informazioni su x e poi chiedi informazioni su y ad un altro campione ottengo alla fine gli stessi risultati come se avessi chiesto x ed y allo stesso campione.
- F_x e G_y ripartizione delle due variabili aleatorie.
- Indipendenti significa che la densità congiunta fattorizza nel prodotto delle densità marginali.
  - $f(x, y, \theta) = f_x(x, \theta) \cdot f_y(y, \theta)$

25.1 Omogeneità dei dati

Due modi di procedere

25.1.1 Confrontare in modo esaustivo le funzioni di ripartizioni

- Elaborare un test statistico qualunque sia il modello sotto F e G
- $H_0 : F = G \quad H_1 = F \neq G$
- $H_0 : F(w) \leq G(w) \quad \forall w \in R$
  - Se per qualche valore $F(w) < G(w)$
  - $P(X \leq w) < P(Y \leq w)$ significa che X "tende" ad assumere valori più grandi di Y
- $H_1 : F(w) > G(w) \quad \forall w$

Esempio $X \sim N(\mu_x, \sigma^2) \quad Y \sim N(\mu_y, \sigma^2)$ e $F_x \leq G_y$

$F_x(w) = \Phi\left(\frac{w-\mu_x}{\sigma^2}\right) < \Phi\left(\frac{x-\mu_y}{\sigma^2}\right) \iff \frac{w-\mu_x}{\sigma^2} \leq \frac{x-\mu_y}{\sigma^2} \iff \mu_x \geq \mu_y$

Sono riuscito a ricondurmi ad un confronto tra due funzioni ad un confronto tra due numeri (le medie)

25.1.2 Test di omogeneità sulle medie

- $H_0 : F = G \quad H_1 : F \neq G \Longrightarrow H_0 : \mu_x = \mu_y \quad H_1 : \mu_x \neq \mu_y$
- $H_0 : F \leq G \quad H_1 : F > G \Longrightarrow H_0 : \mu_x \geq \mu_y \quad H_1 : \mu_x < \mu_y$

25.2 Test di omogeneità sulle medie per dati accoppiati

Dalla differenza delle medie utilizzo la differenza delle osservazioni $D_1, ..., D_n$ che sono iid e alla fine mi sono ricondotto ad un caso unidimensionale $\mu_D = \mu_x - \mu_y$. Il problema si trasforma nel seguente modo:

- $H_0 : \mu_D = \Delta \quad H_1 : \mu_D \neq \Delta$

Se i dati sono **numerosi** (una trentina?) uso il seguente test

$$\text{Rif } H_0 \text{ se } \frac{\left|\bar{D} - \Delta\right|}{\sqrt{\frac{S_D^2}{n}}} \geq z_{1-\frac{\alpha}{2}}$$

Se il campione è **piccolo**, quindi mi serve un test esatto, e le differenze sono normali $D_j \sim N$ allora

$$\text{Rif } H_0 \text{ se } \frac{\left|\bar{D} - \Delta\right|}{\sqrt{\frac{S_D^2}{n}}} \geq t_{n-1;1-\frac{\alpha}{2}}$$

Il modello tipico è la normale bidimensionale.

26 Esercitazioni

26.1 Esercizio 4.1.4be

$$X \sim f_x(x) = \frac{1}{2\theta\sqrt{x}} e^{-\frac{\sqrt{x}}{\theta}} \quad x > 0$$

- H_0 : durata media della batteria è 3.92
- H_1 : durata media della batteria è 8

26.1.1 Punto 1

Dobbiamo calcolare

$$\begin{aligned} E[X] &= E[f_x \cdot x] \\ &= \int_0^\infty \frac{\sqrt{x}}{2\theta} e^{-\frac{\sqrt{x}}{\theta}} dx \end{aligned}$$

Procedo integrale per sostituzione $\sqrt{x} = t \rightarrow x = t^2$ e $dx = 2t \cdot dt$

$$= \int_0^\infty \frac{t}{2\theta} e^{-\frac{t}{\theta}} 2t \cdot dt$$
$$= \int_0^\infty \frac{t^2}{\theta} e^{-\frac{t}{\theta}} dt$$

Sappiamo che $\int_0^\infty \frac{1}{\Gamma(3)\theta} t^2 e^{-\frac{t}{\theta}} dt = 1$

$$= \Gamma(3)\,\theta^2$$

$E[x] = 2\theta^2$

Quindi

- $H_0 \iff \theta = 1.4$
- $H_1 \iff \theta = 2$

26.1.2 Punto 2

$Y = \sqrt{X} \sim \varepsilon(\theta)$

26.1.3 Punto 3

Costruiamo il test, con n osservazioni

$$L = \left(\frac{1}{2\theta}\right)^n \frac{1}{\sqrt{\prod_i x_i}} e^{-\frac{1}{\theta}\sum_i \sqrt{x_i}}$$

$$\Lambda = \frac{\left(\frac{1}{2.8}\right)^n \frac{1}{\sqrt{\prod_i x_i}} e^{-\frac{1}{1.4}\sum_i \sqrt{x_i}}}{\left(\frac{1}{4}\right)^n \frac{1}{\sqrt{\prod_i x_i}} e^{-\frac{1}{2}\sum_i \sqrt{x_i}}} = \left(\frac{4}{2.8}\right)^n e^{\left(\frac{1}{2}-\frac{1}{1.4}\right)\sum_i \sqrt{x_i}}$$

Rifiuto H_0 se $\left(\frac{4}{2.8}\right)^n e^{\left(\frac{1}{2}-\frac{1}{1.4}\right)\sum_i \sqrt{x_i}} < \delta \rightarrow \sum_i \sqrt{x_i} > \gamma$

"Sappiamo" che $\sum_i \sqrt{x_i} \sim \Gamma(n, \theta)$ e siccome il calcolo non è semplice dovremo fare riferimento alle tabelle, ma prima dobbiamo eseguire qualche trasformazione

$$\Gamma(n, 2) = \chi^2_{2n}$$

$\left(\frac{2}{\theta}\sqrt{X}\right) \sim \varepsilon(2) \quad \rightarrow \quad \frac{2}{\theta}\sum\sqrt{X} \sim \chi^2_{2n}$

Rifiuterò H_0 se $\sum_i \sqrt{x_i} > \frac{\theta}{2}\chi^2_{2n;(1-\alpha)}$

26.1.4 Punto: funzione potenza del test

$\Pi = 1 - P(\text{non rif } H_0 | H_0 \text{ falsa}) = 1 - P\left(\sum_i \sqrt{x_i} < \chi^2_{2n;(1-\alpha)}\right)$

26.2 Esercizio

Un amico ci viene incontro dicendoci che un tizio lo ha avvicinato chiedendogli se voleva giocare a testa o croce e dopo 10min è riuscito a perdere tutti i soldi nel portafoglio. Dato che siamo esperti di statistica ci chiede, presentando la monetina con la quale ha giocato, se la moneta è truccata.

Utilizzo la distribuzione bernulliana. Ed ipotizzo che la moneta sia equa fino a prova contraria, significa che se la moneta è non equa ne saà convinto al di là di ogni ragionevole dubbio.

- $H_0 : p = \frac{1}{2}$
- $H_1 : p \neq \frac{1}{2}$

[FIG 3]

Se n è molto alto lo **z-test** è asintotico (?)

26.2.1 Statistica test

$$u = \frac{\bar{X}_0 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overbrace{\sim}^{H_0} Z$$

26.2.2 Risoluzione

- 200 lanci, 120 testa
- $u = \frac{\frac{120}{200} - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{200}}} = 2.81$
- $\alpha = 5\%$ rifiuto H_0 se $|u| > z_{0.95} = 1.96$
- Conclusione: il nostro amico si è fatto fregare

26.3 Esercizio

26.3.1 Testo

Uovo Kinder: è uscita un nuovo set di sorprese, gli elefanti blu. La pubblicità dice che una sorpresa su cinque è un elefante blu. Noi vogliamo un elefante blu. Compriamo 10 uova e non c'è neanche un elefante. Domanda: val la pena di andare dal giudice per dire che la pubblicità è ingannevole? Calcolare il p-value di questo test.

26.3.2 Soluzione

- $p_o = \frac{1}{5} \quad n = 10$
- $H_0 : p = \frac{1}{5} \quad H_1 : p < \frac{1}{5}$
- Con 10 uova non è possibile utilizzare lo z-test perchè n troppo piccolo.
- Il p-value è legato in qualche modo alla regione critica
- $pvalue = P_{H_0}\left(\bar{X} \leq 0\right) = \left(\frac{4}{5}\right)^{10} \backsimeq 0.1$

27 Esercitazioni

27.1 Esercizio 1

- $X \sim N(\mu = 30, \sigma^2 = ?)$
- $\sum_{i=1}^{100} x_i = 3006 \quad \sum_{i=1}^{100} x_i^2 = 90711$
- $H_0 : \sigma \geq 2 \quad H_1 : \sigma < 2$

Stima di σ^2, utilizzo la varianza campionaria

$$\begin{cases} \text{media incognita} & S^2_{n-1} = \frac{1}{n-1} \sum \left(x_i - \bar{X}\right)^2 = \frac{1}{n-1} \left(\sum_i x_i^2 - n\left(\bar{X}\right)^2\right) \\ \text{media nota} & S^2_n = \frac{1}{n} \sum \left(x_i - \mu\right)^2 = \frac{\sum_i x_i^2}{n} + \mu^2 - 2\mu\bar{X} \end{cases}$$

$S^2_n = \frac{90711}{100} + 30^2 - 230\frac{3006}{100} = 3.51$

Specificare una **regione critica** con $\alpha = 2.5\%$. Il test sulla varianza cade sempre sulla χ^2.

La statistica test $u = \frac{nS_0^2}{\sigma_0^2} = \frac{100 \cdot 3.51}{4} = 87.75$. Per questo test, quando rifiutiamo H_0? u è distribuita come una chi quadro di n gradi di libertà e rifiuto H_0 se finisco nella coda sinistra della distribuzione: Rif H_0 se $x < \chi^2_{100;\alpha}$

Fornire la **funzione di potenza** $\Pi(\sigma^2)$. La funzione di potenza non dice che H_1 è vera, la funzione dipende dalla variabile che stiamo testando. $= 1 - P_\sigma$ (non rif H_0), in base al valore di σ ha significato diverso, di per sè non dice la probabilità di fare la cosa giusta o sbagliata.

$$\begin{aligned} \Pi\left(\sigma^2\right) &= 1 - P\left(\frac{100S_0^2}{4} > \chi^2_{100;(\alpha)}\right) \\ &= 1 - P\left(\underbrace{\frac{100S^2}{\sigma^2}}_{\chi^2_{100}} > \frac{4}{\sigma^2}\chi^2_{100;(\alpha)}\right) \\ &= 1 - P\left(\Gamma(50, 2) > \frac{4}{\sigma^2}\chi^2_{100;(\alpha)}\right) \end{aligned}$$

$\chi^2_{100;(\alpha)}$ non è presente sulle tavole. Dato che n è molto grande posso utilizzare le tavole della normale.

$$\chi^2_n \approx N(n, 2n)$$

- $\chi^2_{n;\alpha} = n + z_\alpha \cdot \sqrt{2n}$
- $\chi^2_{100;(0.025)} = 100 - \underbrace{z_{0.975}}_{1.96}\sqrt{200}$

27.2 Esercizio

Abbiamo n osservazioni della geometrica di parametro p. Se la roulette non è truccata

- $H_0 : p = \frac{18}{37} \quad H_1 : p \neq \frac{18}{37}$
- Dobbiamo ricavare gli stimatori di massima verosomiglianza di $p, E[X], Var[X]$.
  - Lo stimatore di p è $\hat{p}_L = \frac{1}{\bar{X}}$
  - Lo stimatore $\hat{E}[X]_L = \bar{X}$
  - Lo stimatore della varianza $\hat{Var}[X]_L = \bar{X}^2 - \bar{X}$

Uno stimatore è efficiente se raggiunge il limite di FCR. Se uno è efficiente l'altro non lo è.

Q: Verificare l'ipotesi utilizzando un test per grandi campioni.

- $H_0 : E[X] = \frac{37}{18} \quad H_1 : E[X] \neq \frac{37}{18}$
- $\alpha = 5\%$, Rifiuto H_0 se $|u| > 1.96$
- $u = \frac{\bar{X} - \frac{37}{18}}{\sqrt{\frac{\left(\frac{37}{18}\right)^2 - \left(\frac{37}{18}\right)}{36}}}$, rifiuto H_0 se $|u| > z_{1-\frac{\alpha}{2}}$, p-v$= 2(1 - \Phi(|u|))$

27.3 Esercizio 5.3.4 (punto 1)

28 Test d'indipendenza dati accoppiati gaussiani

28.1 Introduzione

Dei dati accoppiati noi possiamo essere interessati a studiarne due caratteristiche:

1. Osservare se hanno la stessa **funzione di ripartizione** (quindi stessa media e varianza)
2. Stabilire se i dati sono **indipendenti**.

28.2 Inferenza dati non accoppiati

$(X_1, Y_1), (X_2, Y_2), \ldots$ (sono due rilevazioni sullo stesso individuo) iid distribuzione gaussiana bivariata $\sim N\left(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\right)$ (i dati sono congiuntamente gaussiani). ρ è il coefficiente di correlazione lineare che

$$\rho = \frac{Cov\left(X, Y\right)}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$$

- $|\rho| = 1 \iff Y = a + Xb$ con probabilità 1
- $\rho = 0 \iff$ X ed Y sono scorrelate (indipendenza, sempre vero)
- Se X ed Y indipendenti ($f_{x,y} = f_x f_y$) $\implies E\left[XY\right] = E\left[X\right] \cdot E\left[Y\right] \iff \rho = 0$
- $\rho = 0$ in generale non implica indipendenza
- ma $f_{x,y}\left(x, y\right)$ proporzionale a $\exp\left\{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{x-\mu_y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{x-\mu_y}{\sigma_y}\right)\right]\right\}$
- con $\rho = 0 \implies \exp\left\{\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{x-\mu_y}{\sigma_Y}\right)^2\right]\right\} \sim N\left(\mu_x, \sigma_x^2\right) \cdot N\left(\mu_y, \sigma_y^2\right)$
- Riassumendo: nel caso congiuntamente gaussiano, $\rho = 0 \implies$ **indipendenza**.

l'altro problema dei dati accopiati (il primo è osservare se hanno la stessa **ripartizione**) è controllare **l'indipendenza**. Fare allora un test parametrico su ρ. *Il test di indipendenza è un test su* ρ.

28.3 Test indipendenza dati gaussiani

Posso fare un test per escludere indipendenza

- $H_0 : \rho = 0$
  - $H_1 : \rho \neq 0$
  - $H_1 : \rho > 0$
  - $H_1 : \rho < 0$
- La dipendenza di una variabile dall'altra è lineare nel caso gaussiano

Qualunque problema affronteremo useremo la stessa statistica test. Le ST (statistiche sulla base delle quali prendiamo una decisione) vediamo come si stima ρ.

28.4 Stima di ρ

Stima numeratore e denominatore e poi metti insieme secondo la definizione di ρ

$$\gamma = Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \quad \rightarrow \quad \hat{Cov} = \frac{\sum_{j=1}^{N} (X_j - \bar{X})(Y_j - \bar{Y})}{n-1}$$

$$\hat{\sigma_x^2} = S_x^2 \quad \hat{\sigma_y^2} = S_y^2$$

(nota personale) La stima della covarianza ha come denominatore $n-1$ perchè quando in generale si stima una varianza il denominatore è $n-1$, vedi per esempio le stime delle varianze. Quindi

$$\hat{\rho} = R = \frac{\sum_{j=1}^{N} (X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{N} (X_j - \bar{X})^2 \sum_{j=1}^{N} (Y_j - \bar{Y})^2}}$$

- **Proprietà di R**:
  - $|R| \leq 1$ è una proprietà valida sempre indipendentemente dalla gaussianità
  - Altra proprietà: con le ipotesi del problema iniziale + $\rho = 0$

28.5 Statistica test

Ora che ho individuato un modo per stimare ρ a partire dai dati ricavati, posso sviluppare una statistica test che mi permetta di valutare quando rifiutare o non rifiutare H_0. In particolare alla fine interessa sempre ricondurre la statistica ad una distribuzione nota al fine di poter utilizzare le tabelle.

$$ST = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \sim t_{n-2} \quad n \geq 3$$

[FIG 1] ST è una funzione strettamente crescente su R, pertanto la verifica d'ipotesi posso farla sulla statistica test piuttosto che sul coefficiente di correlazione, perchè posso utilizzare le tabelle (in particolare quella della t-student)

28.6 Riformulazione del test d'ipotesi

- $H_0 : \rho = 0 \quad H_1 : \rho \neq 0$
- Rifiuto H_0 quando ST è lontano da zero, ovvero sono al di fuori di un certo intervallo che contiene lo 0, formalmente rifiuto H_0 se $|ST| \geq k$ con $k : \alpha = P_{\rho=0}\left(|ST| \geq k\right) \Longrightarrow k = t_{n-2;\left(1-\frac{\alpha}{2}\right)}$
  - $|| \Longrightarrow \frac{\alpha}{2}$
  - $\geq k \Longrightarrow (1 - ...)$ come quantile

Altro test

- $H_0 : \rho = 0 \quad H_1 : \rho > 0$
- Rifiuto H_0 quando ST è lontano da zero ma positivo, ovvero sono al di fuori di un certo intervallo che contiene lo 0, formalmente rifiuto H_0 se $ST \geq k$ con $k : \alpha = P_{\rho=0}\left(ST \geq k\right) \Longrightarrow k = t_{n-2;(1-\alpha)}$

28.7 Ricapitolo

1. Voglio eseguire un **test d'indipendenza** su dati gaussiani.
2. L'elemento matematico che permette di esprimere l'indipendenza di due variabili è il **coefficiente di correlazione** ρ (attenzione! l'indipendenza implica sempre $\rho = 0$, ma $\rho = 0$ in generale non implica indipendenza e questo è vero se i dati sono gaussiani)
3. Eseguo pertanto una **verifica d'ipotesi** su ρ
4. Dati dati ottengo una **stima** di ρ
5. $\hat{\rho}$ da sola non è utile perchè non riesco a legarla ad α (errore I tipo)
6. Elaboro una **statistica test** in funzione di $\hat{\rho}$ ed n che può essere approssimata ad una t-student
7. Dato che la ST è strettamente crescente allora posso riconsiderare la verifica d'ipotesi in funzione di ST, invece che di ρ.
8. Valuto se rifiutare H_0 o no, e questa volta posso considerare α.

29 Inferenza non parametrica

Nulla sappiamo della distribuzione congiunti, nel caso dei dati accoppiati (X,Y)

29.1 Problema di omogeneità

29.1.1 Parte 1

$(X_1, Y_1), (X_2, Y_2)$,... rimuoviamo l'ipotesi di normalità, quindi si eseguono test sulle distribuzioni

- $H_0 : F = G \quad H_1 : "F \leq G"$
- $X \sim F, Y \sim G$
- "$F \leq G$" X tende ad essere più grande di Y

- Ipotesi di lavoro in (X, Y)
  - Non osservo dati ripetuti, nè tra le X, nè tra le Y, nè a coppie (*no ties*), formalmente $P(X_i = Y_j) = P(X_i = X_j) = P(Y_i = Y_j) \quad \forall i = 1, ..., n \quad \forall j = 1, ..., n$ questa proprietà la ottengo se la funzione di ripartizione è la seguente

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{x,y}(s, t)\, dsdt$$

    Funzione di ripartizione **continua**.

Traduco $H_0 : F = G \rightarrow H_0 : P(X > Y) = P(X < Y) = \frac{1}{2} \quad (P(X = Y) = 0$ vedi ipotesi)

- $H_0 : \underbrace{P(X > Y)}_{P} = \frac{1}{2} \quad H_1 : P(X > Y) > \frac{1}{2}$ [FIG 2]
- ST numero di coppie con $X > Y$. $ST \sim Bin(n, p)$
  - Sotto H_0 ho che $ST \sim Bin\left(n, \frac{1}{2}\right)$ e rifiuto H_0 se $ST > q_{Bin\left(n,\frac{1}{2}\right);(1-\alpha)}$ Questo è un test esatto
  - Osserviamo ora il p-value (usiamo la regola empirica): se s è il valore della ST,

$$\begin{aligned} pv &= P\left(Bin\left(n, \frac{1}{2}\right) \geq s\right) \\ &= 1 - P\left(Bin\left(n, \frac{1}{2}\right) \leq s\right) \\ &= 1 - \sum_{k=0}^{s} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{n-k} \\ &= 1 - \frac{1}{2^n} \sum_{k=0}^{s} \binom{n}{k} \end{aligned}$$

  - Se n grande allora $ST \sim N(np, np(1-p))$, sotto $H_0 \rightarrow ST \sim N\left(\frac{n}{2}, \frac{n}{4}\right) \Longrightarrow q_{Bin\left(n,\frac{1}{2}\right);(1-\alpha)} = \sqrt{\frac{n}{4}} \cdot z_{1-\alpha} + \frac{n}{2}$
  - **Test di Wilkson dei segni per dati accoppiati**

29.1.2 Parte 2

$(X_1, Y_1), (X_2, Y_2), \ldots$ quindi

- $H_0 : F = G \quad H_1 : "F \leq G"$
- $X \sim F, Y \sim G$
- $"F \leq G"$ X tende ad essere più piccola di Y

Allora rifiuto H_0 se $ST < q_{Bin\left(n,\frac{1}{2}\right)(\alpha)}$ ed il p-value$= P\left(Bin\left(n, \frac{1}{2}\right) < s\right) = \sum_{k=0}^{s-1} \binom{n}{k} \left(\frac{1}{2^n}\right)$

29.2 Test di Wilcoxon-Mann-Wintney

29.2.1 Ipotesi

- $X_1, ..., X_m$ iid $\sim F$
- $Y_1, ..., Y_n$ iid $\sim G$
- F e G sono indipendenti
- $H_0 : F = G \quad H_1 : F \leq G$ (X tende ad essere più grande di Y)
- Prima ipotesi che impongo: non ci sono ripetizione nei dati (devo ipotizzare **F e G continue**)

29.2.2 Risoluzione

Per eseguire il test ho bisogno di una statistica su cui lavorare. Il ragionamento alla base del test **WNW** consiste nel contare il numero di coppie tale per cui $X > Y$ che è un modo per rappresentare lo **scostamento** di una F.d.r dall'altra. Se questo numero è troppo alto allora rifiuto l'ipotesi che $F = G$ (F e G sono funzioni di ripartizione).

- $U =$numero di volte in cui $X > Y$ sulle $m \times n$ coppie possibili
- Se H_0 **vera** (ovvero X ed Y sono regolate dallo stesso modello)

  - $E_{H_0}(U) = \frac{m \cdot n}{2}$
  - $Var_{H_0}(U) = \frac{mn(m+n+1)}{12}$
  - Sono stati messi in evidenza solo perchè servono a costruire T_x
- Valori estremi della statistica, $T_x = R_1 + R_2 + ... + R_m$
  - $\min(T_x) = 1 + 2 + ... + m = \frac{m(m+1)}{2}$ (quando le prime m posizioni sono occupate solo da X)
  - $\max(T_x) = (n+1) + (n+2) + ... + (n+m) = n \cdot m + \frac{m(m+1)}{2} = m\left(n + \frac{m+1}{2}\right)$
  - vere ind da H0

Rifiuto H_0 se $T_x > k$. Osserviamo ora il legame tra U ed T_x

- $sort(x) = x_1, ..., x_m$
- $\#Y_j < X_{(1)} = R_1 - 1$ se $R_1 = 10 \Longleftrightarrow$prima di $X_{(1)}$ ci sono 9 y
- $\#Y_j < X_{(2)} = R_2 - 2$
- $\#Y_j < X_{(m)} = R_m - m$

29.2.3 Conclusione

$$\begin{aligned} U &= (R_1 - 1) + (R_2 - 2) + ... + (R_m - m) \\ &= T_x - (1 + 2 + ... + m) \\ &= T_x - \frac{m(m+1)}{2} \end{aligned}$$

- $E_0(T_x) = E_0(U) + \frac{m(m+1)}{2} = \frac{m(m+n+1)}{2}$
- $Var(T_x) = \frac{mn(m+n+1)}{12}$
- Verifica del test
  - $T_x > k \sim w_{m,n}(1-\alpha)$ e m ed n piccoli ed utilizzo le tavole
  - se ($m \geq 7$ e $n > 7$) sotto H_0, $T_x \sim N\left(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right)$

29.2.4 Osservazione

1. Se due persone differenti **invertono X ed Y** stanno comunque facendo lo **stesso test**. Un test basato su T_y porta alle stesse conclusioni di quello basato di T_x. $T_x + T_y = 1 + 2 + ... + m + n = \frac{(m+1)(m+n+1)}{2}$ vero perchè non ho ripetizioni. Se ci sono ripetizioni, ma in numero ridotto questo non è un problema, ma devo calcolare il rango medio

2. Se eseguo l'operazione $w_p = m(m+n+1) - w_{1-p}$ allora H_0 e H_1 si invertono ed anche la condizione di rifiuto di H_0 si ribalta. Solitamente non è necessario eseguire un'operazione del genere, ma è sufficiente fare delle considerazioni tra il p-value e α ; infatti è sufficiente ricavare che il p-value è inferiore ad α per rifiutare H_0 (i valori esatti non interessano)

30 Test omogeneità campioni gaussiani indipendenti

Finiamo la verifica d'ipotesi parametrica

- $X_1, ..., X_m$ iid $\sim N(\mu_x, \sigma_x^2)$
- $Y_1, ..., Y_n$ iid $\sim N(\mu_y, \sigma_y^2)$
- Fare un test di omogeneità significa confrontare i due campioni ed osservare se le distribuzioni sono le stesse o no.
- $H_0 : \mu_x = \mu_y, \sigma_x^2 = \sigma_y^2 \quad H_1 : \mu_x \neq \mu_y | \sigma_x^2 \neq \sigma_y^2$

Il test sarà sequenziale. Confrontiamo le varianze e se accettiamo l'ipotesi H_0 faremo poi un confronto sulle medie. Il test sarà in due passi

1. Test di confronto delle **varianze**: $H_0 : \sigma_x^2 = \sigma_y^2 \quad H_1 : \sigma_x^2 \neq \sigma_y^2$, fissando un livello α_1
2. Test di confronto delle **medie**: $H_0 : \mu_x = \mu_y \quad H_1 : \mu_x \neq \mu_y$ se al passo 1 a livello α_1 non ho rifiutato H_0 varianze uguali, ed effetto il test a livello α_2. Se ho invece rifiutato il test 1 allora mi fermo.

Livello di significativià complessivo, ovvero il problema originario che livello di significatività ha? La significatività totale del test a due passi è (non dimostrato)

$$\begin{aligned} 1-(1-\alpha_1)(1-\alpha_2) &= 1-[1-\alpha_2-\alpha_1+\alpha_1\alpha_2] \\ &= (\alpha_1+\alpha_2)- \underbrace{\alpha_1\alpha_2}_{\text{trascurabile}} \\ &\cong \alpha_1+\alpha_2 \end{aligned}$$

Attenzione quindi quando si svolgono più test sullo stesso set di dati, perchè il livello di significatività aumenta.

30.1 Problema

- $X_1, ..., X_m$ iid $\sim N(\mu_x, \sigma_x^2)$
- $Y_1, ..., Y_n$ iid $\sim N(\mu_y, \sigma_y^2)$
- Test:
  - $H_0 : \sigma_x^2 = \sigma_y^2 \quad H_1 : \sigma_x^2 \neq \sigma_y^2$
  - $H_0 : \sigma_x^2 \leq \sigma_y^2 \quad H_1 : \sigma_x^2 > \sigma_y^2$
  - $H_0 : \sigma_x^2 \geq \sigma_y^2 \quad H_1 : \sigma_x^2 < \sigma_y^2$
  - La statistica test sarà sempre la stessa. Il risultato sarebbe stato lo stesso se avessimo utilizzato lo stimatore di massima verosomiglianza
- α
- μ_x e μ_y incognite (potrebbero non esserlo)

30.1.1 Risoluzione

Distribuzione congiunta di ST sotto H_0 ? Da cui deriva la regione critica (perchè la approssimo ad una 'tabella')

- $H_0 : \frac{\sigma_x^2}{\sigma_y^2} = 1 \quad H_1 : \frac{\sigma_x^2}{\sigma_y^2} \neq 1$

- Non posso scrivere $\sigma_x^2 = \sigma_y^2$ perchè al secondo membro non posso avere un'incognita, al più posso avere una variabile come nel caso di $\theta = \theta_0$. Per questo motivo sono costretto a testare un rapporto.

- Stima di $\frac{\sigma_x^2}{\sigma_y^2}$? che posso ottenere con il rapporto delle varianze campionarie, che è anche la statistica test $ST = \frac{S_x^2}{S_y^2}$
- Rifiuto H_0 se questo rapporto è lontano da 1, cosa che devo formalizzare nel seguente modo: rifiuto H_0 se $ST \geq k_2$ oppure $ST \leq k_1$ con $(k_1 < k_2)$.

$$k_1, k_2 : \alpha = P_{H_0}\left(ST \geq k_2 | ST \leq k_1\right)$$

- **Nota**: non posso utilizzare $|ST| > k$ perchè la statistica test non si può approssimare ad una normale, oppure ad una funzione di distribuzione simmetrica. Infatti se le X_i seguono la legge normale (μ, σ), **lo stimatore S_{n-1}^2 segue la legge del χ^2** , che ha un andamento non simmetrico

30.2 Densità F di Fischer

30.2.1 Definizione

- W_1, W_2 indipendenti
- $W_1 \sim \chi_a^2, W_2 \sim \chi_b^2$
- Poniamo $Z := \frac{\frac{W_1}{a}}{\frac{W_2}{b}}$ (Z non indica la normale) allora
  - $Z \geq 0$
  - Z continua
  - la sua densità è detta **densità F di Fischer** con a gradi di libertà al numeratore e b al denominatore

$$Z \sim F_{a,b}$$

  - I quantili sono indicati $F_{a,b}\left(p\right)$

- Dalle tavole conosco $F_{a,b}(1-\alpha)$, come determinare $F_{a,b}(\alpha)$?

$$\begin{aligned}
\alpha &= P\left(Z \geq F_{a,b}(\alpha)\right) \\
&= P\left(\frac{\frac{W_1}{a}}{\frac{W_2}{b}} \geq F_{a,b}(\alpha)\right) \\
&= P\left(\underbrace{\frac{\frac{W_2}{b}}{\frac{W_1}{a}}}_{\sim F_{b,a}} \leq \frac{1}{F_{a,b}(\alpha)}\right) \\
&= P\left(F_{b,a} \leq \frac{1}{F_{a,b}(\alpha)}\right) \\
&= 1 - P\left(F_{b,a} > \frac{1}{F_{a,b}(\alpha)}\right) \\
&= \ldots \text{ FABIO}
\end{aligned}$$

30.2.2 Situazione (test-bilatero)

- $\frac{S_x^2(m-1)}{\sigma_x^2} \sim \chi^2_{m-1}$
- $\frac{S_y^2(n-1)}{\sigma_y^2} \sim \chi^2_{n-1}$
- χ^2_{m-1} indipendente da χ^2_{n-1}
- $\frac{\frac{\chi^2_{m-1}}{m-1}}{\frac{\chi^2_{n-1}}{n-1}} \sim F_{m-1,n-1} \rightarrow \frac{\frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}}}{} \sim F_{m-1,n-1}$
- In particolare se è vera H_0 allora il rapporto sopra è la statistica test che stavo cercando.
- $ST = \frac{S_x^2}{S_y^2} \overset{H_0}{\sim} F_{m-1,n-1}$
  - $k_1 = F_{m-1,n-1}\left(\frac{\alpha}{2}\right)$
  - $k_2 = F_{m-1,n-1}\left(1-\frac{\alpha}{2}\right)$

30.2.3 Situazione (test unilatero)

- $H_0 : \frac{\sigma_x^2}{\sigma_y^2} \leq 1 \quad H_1 : \frac{\sigma_x^2}{\sigma_y^2} > 1$
- Rifiuto H_0 se $\frac{S_x^2}{S_y^2} \geq k$
- $k = F_{m-1,n-1}(1-\alpha)$

30.2.4 Osservazione per gli esercizi

- $\pi\left(\sigma_x^2 = \sigma_y^2\right) = P_0\left(\frac{S_x^2}{S_y^2} \geq F_{m-1,n-1}(1-\alpha)\right) = P\left(\frac{1}{2}\frac{S_x^2}{S_y^2} \geq \frac{1}{2}F_{m-1,n-1}(1-\alpha)\right)$
- Nota: con le distribuzioni F di Fischer sono capace di ricavare **analiticamente** la funzione di potenza.

30.2.5 Differenze

μ_x,μ_y inc	medie note
S_x^2, S_y^2	S_{OX}^2, S_{OY}^2
$F_{m-1,n-1}$	$F_{m,n}$

30.2.6 Riassunto

La densità di Fischer permette di mettere in relazione il rapporto tra due distribuzioni χ^2 e un α che indica l'errore di I° grado in un test verifica d'ipotesi. Questo tipo di densità è utilizzato nell'inferenza non parametrica nel confronto tra due varianze campionarie, che è espresso come rapporto tra queste.

La statistica test con la quale si lavora è

$$ST = \frac{\frac{S_x^2}{m-1}}{\frac{S_y^2}{n-1}}$$

che si comporta come una **distribuzione di Fischer**. Il valore della statistica test che ottengo a partire dai dati lo posso confrontare con limiti inferiori e superiori, ricavabili a partire dalle tabelle di Fischer.

$$\alpha = P\left(ST > k_2 | ST < k_1\right)$$

con $k_1 = F_{m-1,n-1}\left(\frac{\alpha}{2}\right)$ e $k_2 = F_{m-1,n-1}\left(1-\frac{\alpha}{2}\right)$

Infine il caso appena guardato considera un test bilatero con medie incognite, se queste fossero invece conosciute allora bisogna adattare le varianze campionarie e gli indici di conseguenza. Se il test è unilaterio, ulteriori modifiche devono essere apportate.

30.3 Confronto tra varianze se le medie sono note

- $X_1, ..., X_m$ iid $\sim N(\mu_x, \sigma_x^2)$
- $Y_1, ..., Y_n$ iid $\sim N\left(\mu_y, \sigma_y^2\right)$
- Campioni indipendenti
- Posso avere varianze note od incognite
- $H_0 : \mu_x - \mu_y = \Delta \quad H_1 : \mu_x - \mu_y \neq \Delta$ nessuna specifica su Δ
- Il punto di partenza è:
  1. $\left(\bar{X} - \bar{Y}\right)$ stima $(\mu_x - \mu_y)$
  2. $\left(\bar{X} - \bar{Y}\right) \sim N\left(\mu_x - \mu_y; \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right)$
  3. $Var\left(\bar{X} - \bar{Y}\right) = Var\left(\bar{X}\right) + Var\left(\bar{Y}\right) - \underbrace{2Cov\left(\bar{X}, \bar{Y}\right)}_{0}$
  4. $\frac{\left(\bar{X}-\bar{Y}\right)-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}} \sim N(0,1)$

30.3.1 Se le varianze sono note (z-test di confronto di medie)

1. $ST = \frac{\bar{X}-\bar{Y}-\Delta}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}} \sim N(0,1)$ sotto H_0, quindi rifiuto H_0 se $|ST| \geq z\left(1-\frac{\alpha}{2}\right)$
2. Questo test lo posso utilizzare quando le varianze sono note e ho grandi campioni. I precedenti test sulle varianze non funzionano se i dati non sono gaussiani.

30.3.2 Se le varianze sono incognite

- $ST_1 = \frac{\bar{X}-\bar{Y}-\Delta}{\sqrt{\frac{S_x^2}{m}+\frac{S_y^2}{n}}}$ va bene se i campioni m ed n sono grandi e
- rifiuto H_0 se $|ST_1| \geq z_{1-\frac{\alpha}{2}}$
- Non va assolutamente bene se m ed n sono piccoli, anche se $X \sim N$ e $Y \sim N$
- Questo problema è ancora aperto e chiamato problema di **Behrens-Fisher ?**

30.4 Confronto media con $\sigma_x^2 = \sigma_y^2 = \sigma^2$ incognito

- Sotto H_0 $\frac{\left(\bar{X}-\bar{Y}\right)-\Delta}{\sqrt{\sigma^2\left(\frac{1}{m}+\frac{1}{n}\right)}} \sim N(0,1)$

30.4.1 Stima di σ^2 comune

Ripasso

- $S_x^2 = \frac{\sum_{i=1}^n \left(x_i-\bar{X}\right)^2}{n-1} \sim \Gamma(\cdot,\cdot)$
- $\frac{S_x^2(n-1)}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$

Stima di S_p^2

1. $\sum_{i=1}^m \left(X_i-\bar{X}\right)^2 + \sum_{j=1}^n \left(Y_j-\bar{Y}\right)^2$ dato che la varianza è comune (per ipotesi) posso sfruttare i dati di entrambe le variabili aleatorie per ottenere un più preciso valore della varianza
2. Affinchè io riesca ad ottenere uno stimatore non distorto $\hat{\sigma}^2$ devo calcolare la media delle somme al punto 1

$$\begin{aligned} E\left[\sum_{i=1}^m \left(X_i-\bar{X}\right)^2 + \sum_{j=1}^n \left(Y_j-\bar{Y}\right)^2\right] &= E\left[(m-1)S_x^2 + (n-1)S_y^2\right] \\ &=^* (m-1)\sigma^2 + (n-1)\sigma^2 \\ &= \sigma^2(m+n-2) \end{aligned}$$

(a) * ricordando che la varianza campionaria è uno stimatore indistorto per n grandi

3. $E\left[\frac{\sum_{i=1}^{m}\left(X_i-\bar{X}\right)^2+\sum_{j=1}^{n}\left(Y_j-\bar{Y}\right)^2}{m+n-2}\right]=\sigma^2 \quad \Longrightarrow \quad S_p^2=\frac{(m-1)S_x^2+(n-1)S_y^2}{m+n-2}$

S_p e χ^2 corrispondente

$$\frac{S_p^2\,(m+n-2)}{\sigma^2}=\underbrace{\frac{(m-1)\,S_x^2}{\sigma^2}+\frac{(n-1)\,S_y^2}{\sigma^2}}_{\text{indipendenti}}\sim\chi^2_{m-1}+\chi^2_{n-1}=\chi^2_{m+n-2}$$

30.4.2 Statistica Test

$ST=\frac{\left(\bar{X}-\bar{Y}\right)-\Delta}{\sqrt{S_p^2\left(\frac{1}{m}+\frac{1}{n}\right)}}\overbrace{\sim}^{H_0}t_{m+n-2}$

Media campionaria di variabili normali e **varianza campionaria** di variabili normali sono **indipendenti**.

Sotto H_0 $ST\sim t_{m+n-2}$ infatti ... (ce lo ricostruiamo a casa). A questo punto

$$G=\left[(x_1,...,x_m,y_1,...,y_n):\frac{\left|\left(\bar{X}-\bar{Y}\right)-\Delta\right|}{\sqrt{S_p^2\left(\frac{1}{m}+\frac{1}{n}\right)}}\geq t_{m+n-2}\left(1-\frac{\alpha}{2}\right)\right]$$

30.4.3 Osservazione ai fini del compito

1. Test sulle medie con varianza incognite non posso fare domande sulla potenza, perchè non ne siamo in grado. A meno che non abbiamo tanti dati che possiamo approssimare ad una normale. Con pochi dati ci servirebbe una t-student non c'entrata, con la quale non sappiamo operare.
2. Posso chiedere la funzione di potenza con: test sulle varianze, test di confronto sulle varianze.
3. Se abbiamo dubbi sulla normalità utilizzare il test di Wilcoxon-Mann-W...

31 Esercitazione

31.1 Esercizio 1

31.1.1 Punto 1

- $X \sim N(\mu_x, \sigma_x^2) \quad Y \sim N(\mu_y, \sigma_y^2)$
- $n = 100$ del tipo (x_i, y_i)
- $\sum x_i = 2371.2, \quad \sum y_i = 2456.5, \quad \sum x_i y_i = 59601.8$
  - $\sum x_i y_i = 59601.8$ è interessante perchè dice che le VA non sono indipendenti, considerazione confermata nel punto due dell'esercizio dove si trova che $\rho \neq 0$.
- $\sum x_i^2 = 57682.1 \quad \sum y_i^2 = 62261.8$
- $H_0 : \mu_x - \mu_y = 0 \quad H_1 : \mu_x - \mu_y < 0$

Per un test di questo genere la statistica test è

$$ST = \frac{\bar{D}}{\sqrt{\frac{S_n^2}{n}}} \overbrace{\sim}^{H_0} t_{(n-1)}$$

Rifiuto H_0 se $-u > t_{n-1}(1-\alpha)$

$$t_{99;0.95} \sim z_{0.95} = 1.645$$

- $\bar{d} = 23712 - 24565$

- $S_i^2 = \frac{\sum_i^n (x_i - \mu_x)^2}{n-1} = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{X}^2\right)$

$$\begin{aligned} S_D^2 &= \frac{\sum_i^n (d_i - \mu_D)^2}{n-1} \\ &= \frac{\sum_i^n d_i^2 - n\bar{d}^2}{n-1} \\ &= \frac{\sum_i^n (x_i - y_i)^2 - n\bar{d}^2}{n-1} \\ &= \frac{1}{n-1}\left(\sum x_i^2 + \sum y_i^2 - 2\sum x_i y_i - n\left(\bar{d}\right)^2\right) \\ &= 8.34 \end{aligned}$$

- $u = \frac{2371.2 - 2456.5}{\sqrt{\frac{8.34}{100}}} = -2.95$

$-u > t_{n-1}(1-\alpha) \rightarrow$ Rifiuto H_0 perchè sono abbastanza lontano da 1.645

31.1.2 Punto 2

È lecito pensare che X ed Y siano indipendenti?

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

- $H_0 : \rho = 0 \quad H_1 : \rho \neq 0$

$$\hat{\rho} = R = \frac{\sum \left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{\sqrt{\underbrace{\sum \left(x_i - \bar{X}\right)^2}_{\sum x_i^2 - n\bar{X}} \cdot \underbrace{\sum \left(y_i - \bar{Y}\right)^2}_{\sum y_i^2 - n\bar{Y}}}} = \frac{59601.8 - \frac{23712 \cdot 24645}{100}}{\sqrt{\left(57682.1 - \frac{(2371.2)^2}{100}\right)(...)}} = 0.81$$

$$ST = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \overbrace{\sim}^{H_0} t_{n-2} = 13.76$$

Con un n grande i quantili della t-studenti sono approssimativamente quantili della normale. Siccome già con un quantile di 3 ho raggiunto il 100, con un quantile di 13.76 sono nell'estremità della campana.

31.2 Esercizio

- $X, Y \sim N$
- Voglio fare un test sulla varianza
- $n_x = 5 \quad n_y = 5$
- $S_x^2 = 3.2 \quad S_y^2 = 15.8$
- $H_0 : \sigma_x^2 = \sigma_y^2 \quad H_1 : \sigma_x^2 > \sigma_y^2$
- $u = \frac{S_x^2}{S_y^2} \overset{H_0}{\sim} F\left(n_x - 1, n_y - 1\right)$
- Rifiuto H_0 se la statistica test è minore di $u < q_{F_{n-1,m-1}}\left(\alpha\right)$ oppure $u > q_{F_{n-1,m-1}}\left(1-\alpha\right)$
- In questa situazione **chi è il p-value**?

$$\begin{aligned} pv &= P_{H_0}\left(F_{4,4} < u\right) \\ &= P_{H_0}\left(F_{4,4} < \frac{3.2}{15.8}\right) \\ &= \int_0^{0.21} f_F\left(v\right) dv \\ &= \int_0^{0.21} 6\frac{v}{\left(1+v\right)^4} dv \\ &= 0.08 \end{aligned}$$

$f_F\left(\theta\right) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}} \frac{v^{\frac{n}{2}-1}}{\left(1+\frac{n}{m}v\right)^{\frac{n+m}{2}}}$

31.3 Esercizio 5.2.2 (Test Wilcoxon)

$$34.6_y, 49.8_y, 49_y, 54_y, 66_x, 77.4_y, 88.5_x, 116.4_y, 120.2_y, 121.3_y$$
$$122.3_x, 125_x, 127.8_y, 132_x, 162_x, 211.9_x$$

In particolare ci interessa l'ordine di comparizione delle variabili

$$y, y, y, y, x, y, x, y, y, y, x, x, y, x, x, x$$

Le x si trovano nella posizione (ranghi) 5,7,11,12,14,15,16.

$$T_x = \sum r_x = 80$$

- Scelgo l'ipotesi $F_x > F_y$ se $T_x < \underbrace{w_\alpha}_{\text{tabella}}$
- Scelgo l'ipotesi $F_x < F_y$ (H_1) se $\mathbf{T_x > w_{1-\alpha}}$
- Scelgo l'ipotesi $F_x \neq F_y$ se $T_x < w_{\frac{\alpha}{2}} | T_x > w_{1-\frac{\alpha}{2}} = n_x (n_x + n_y + 1) - w_{\frac{\alpha}{2}}$
- Rifiuto H_0 se $80 > w_{0.975}$
  - Considerando un $m_x = 7$ ed un $n_y = 9$, ottengo che $w_{0.975} = m (m + n + 1) - w_{0.025}$
  - $w_{0.025}$ si legge dalle tabelle di **WNW**.
  - $w_{0.975} = 78$
  - $80 > 78$, rifiuto H_0

31.4 Esercizio 5.2.5

Le ipotesi da sottolineare (tra le altre?) sono

- Non ripetizione dei dati

31.4.1 Punto 1

- $(X_1, Y_1) \dots (X_6, Y_6)$
- $H_0 : F = G \quad H_1 : F < G$
  - F è la distribuzione dei dati in assenza di attività fisica

  - G è la distribuzione dei dati con intensa attività fisica

- $k = \#$ coppie in cui $x_i > y_i$, mi è sufficiente sapere k per calcolare il p-value.

$$\begin{cases} pv = \sum_{j=k+1}^{n} \binom{n}{j} \frac{1}{2^n} & H_1 : X > Y \\ pv = \sum_{j=0}^{k-1} \binom{n}{j} \frac{1}{2^n} & H_1 : X < Y \end{cases}$$

$$pv = \sum_{j=5}^{6} \binom{n}{j} \frac{1}{2^6} = \frac{7}{64} \sim 0.11$$

Rifiuto H_0 perchè $12\% > 5\%$

31.4.2 Punto 2

$$\begin{aligned} \bar{X}_d &= \frac{(x_1 - y_1) + (x_2 - y_2) + ... + (x_n - y_n)}{6} \\ &= \frac{\sum_{i=1}^{6} x_i - \sum_{i=1}^{6} y_i}{6} \\ &= \frac{116.1 - 101.9}{6} = 2.367 \end{aligned}$$

$$\begin{aligned} S_d &= \frac{\sum_{i=1}^{6} \left[x_{d_i} - \bar{X}_d\right]^2}{5} \\ &= \frac{\sum_{i=1}^{6} \left(x_{d_i}^2\right) - 5\bar{X}_d^2}{5} \\ &= \frac{\sum_{i=1}^{6} (x_i - y_i)^2 - 5\bar{X}_d^2}{2} \\ &= \frac{\sum_{i=1}^{6} x_i^2 + \sum_{i=1}^{6} y_i^2 - 2\sum_{i=1}^{6} x_i y_i - n\bar{X}_d^2}{5} \\ &= \frac{2992.81 + 1782.99 - 2 \cdot 2130.82 - 6 \cdot 5.6}{5} \\ &= \frac{480.56}{5} \\ &= 96.112 \end{aligned}$$

31.4.3 Punto 3

- $H_0 : \mu_x - \mu_y = 0 \quad H_1 : \mu_x - \mu_y > 0$
  - μ_x ormone della crescita medio in assenza di attività
  - μ_y ormone della crescita media in presenza di forte attività fisica

$ST : \frac{\bar{x}_d - 0}{\sqrt{\frac{S_d^2}{n}}} = \frac{1.58}{4.08} = 0.39$ con un p-value del 39% non rifiuto H_0

31.5 Esercizio 5.2.7

31.5.1 Punto 1

$$\begin{aligned} \mathbf{S_p^2} &= \frac{(\mathbf{m-1})\,\mathbf{S_x^2} + (\mathbf{n-1})\,\mathbf{S_y^2}}{\mathbf{m+n-2}} \\ &= \frac{43 \cdot 4.9 + 52 \cdot 5.2}{44 + 53 - 2} \\ &= 5.06421 \end{aligned}$$

31.5.2 Punto 2

$$\begin{aligned} P\left(T_1 < \sigma_p^2 < T_2\right) &= 0.9 \\ P\left(q_1 < \frac{(m+n-2)\,S_p^2}{\sigma_p^2} < q_2\right) &= 0.9 \end{aligned}$$

L'intervallo di confidenza è il seguente: $\left(\frac{(m+n-2)S_p^2}{\chi^2_{m+n-2}\left(\frac{1+\gamma}{2}\right)}; \frac{(m+n-2)S_p^2}{\chi^2_{m+n-2}\left(\frac{1-\gamma}{2}\right)}\right)$

- $(m+n-2) = 95$
- $S_p^2 = 5.064211$

- $\chi^2_{95}\left(\frac{1+\gamma}{2}\right) \sim z_{\frac{1+\gamma}{2}}$ perchè i grandi di libertà sono tanti
  - $\chi^2_{m+n-2}\left(\frac{1+\gamma}{2}\right) \cong 1.65$
  - $\chi^2_{m+n-2}\left(\frac{1-\gamma}{2}\right) \cong 1 - 1.65 = 0.65$

Con questi calcoli l'intervallo di confidenza diventa $(291.58; 740.15)$ e signica che il 90% delle volte la varianza cade in questo intervallo.

31.5.3 Punto 3

- $H_0 : \mu_x - \mu_y = 0, \quad H_1 : \mu_x - \mu_y > 0$
- Dal formulario utilizzo il *Test per il confronto di medie di due popolazioni gaussiane*, con varianze incognite ma uguali [t-test]
  - Si rifiuta H_0 se $\frac{\bar{x}-\bar{y}}{\sqrt{S_p^2\left(\frac{1}{m}+\frac{1}{n}\right)}} \geq t_{m+n-2}\left(1-\alpha\right)$
  - p-value: $1 - P\left(t_{m+n-2} \leq \frac{\bar{x}-\bar{y}}{\sqrt{S_p^2\left(\frac{1}{m}+\frac{1}{n}\right)}}\right)$

Si rifiuta H_0 se $\frac{7.2-6.9}{\sqrt{5.064211\left(\frac{1}{44}+\frac{1}{53}\right)}} \geq t_{95}\left(1-0.05\right) \sim z_{(1-0.05)} \rightarrow 0.6536 \geq 1.65$ pertanto non rifiuto H_0

Calcolo del p-value: $1 - P\left(t_{95} \geq 0.6536\right) = 1 - 0.7389 = 0.2611 = 26.11\%$

31.6 Esercizio 5.2.10

- $H_0 : \mu_x - \mu_y = 0, \quad H_1 : \mu_x - \mu_y < 0$
- Considero le popolazioni X e Y distribute come due gaussiane
- Regione critica

$$G : \left\{ \left(x_1, ..., x_7\right), \left(y_1, ..., y_6\right) : \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{7}+\frac{1}{6}\right)}} \leq -t_{11}\left(1-\alpha\right) \right\}$$

- Calcoli

- $\bar{X} = 1.896$
- $\bar{Y} = 2.933$
- $S_x^2 = \frac{\sum_{i=1}^{7}\left(x_i - \bar{X}\right)^2}{7-1} = 0.27$
- $S_y^2 = \frac{\sum_{i=1}^{6}\left(y_i - \bar{Y}\right)^2}{6-1} = 1.344$
- $S_p^2 = \frac{(7-1)S_x^2 + (6-1)S_y^2}{7+6-2} = \frac{8.34}{11} = 0.758$

$$\begin{aligned} \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{7} + \frac{1}{6}\right)}} &\leq -t_{11}\left(1-\alpha\right) \\ \frac{-1.037}{0.484} &\leq -t_{11}\left(0.95\right) \\ -2.143 &\leq -1.796 \\ 2.143 &\geq 1.796 \end{aligned}$$

Rifiuto H_0 e significa che il nuovo protocollo al di là di ogni ragionevole dubbio trasmette i dati in minore tempo rispetto al vecchio protocollo

32 Test di buon adattamento χ^2 di Pearson

$$\sum_{i=0}^{k} \frac{\left(N_i - np_{0i}\right)^2}{np_{0i}} \geq \chi^2_{k-n-2}\left(1-\alpha\right)$$

Dati categorici: individui raggruppati per caratteristica (razza, sesso, ecc..).

32.1 Test d'indipendenza

Utilizzo questo test quando non posso utilizzare quello per dati accoppiati perchè non ho i dati grezzi ma raggruppati.

- $H_0 : X, Y$ sono indipendenti
- $H_1 : X, Y$ non sono indipendenti

Dati

X\Y	b_1	b_2	...	b_{r_2}	$N_{i,\cdot}$ num delle X con coordinata i
a_1	$N_{1,1}$				$\sum_{j=1}^{r_2} N_{i,j}$
a_2	$N_{2,1}$				
...					
a_{r_1}					
$N_{\cdot,j}$	$N_{\cdot,1}$				

Per ogni coppia di tipo (a,b) ho contato quante sono le X con...

A monte c'è un campione accoppiato bidimensionale $(X_1, Y_1), ..., (X_n, Y_n)$ iid, ora vado a contare per ogni possibile coppia assumono la modalita $X = a_i$ e $Y = b_j$, ovvero $N_{i,j} = \#coppie\,(X, Y)$ con $X = a_i, Y = b_j, i = 1, ..., r_1, j = 1, ..., r_2$

$\sum_i \sum_j N_{i,j} = n$ ripartisco quindi le n coppie nella tabella in base ai valori.

Indipendenza: $P(X, Y) = P(X) \cdot P(Y)$

Come posso allora tradurre le ipotesi?

- $H_0 : P(X = a_i, Y = b_j) = p_i q_j$ con $p_i = P(X = a_i), q_j = P(Y = b_j) \quad \forall i, \forall j$
- $H_0 : f(a, b) = p_i q_j$

Per decidere se rifiutare o meno l'ipotesi conto quante coppie effettive ho e quante mi aspettavo se l'indipendenza fosse rispettata, ovvero n° atteso di coppie (X, Y) su n con $X = a_i$ e $Y = b_j$ se H_0 è vera $= n \cdot p_i \cdot q_j$ per ogni i e j.

Se p_i è incognito stimo $\hat{p}_i$ contando quante volte la X su n X

$$\hat{p}_i = \frac{n^\circ \text{ di } X = a_i}{n} = \frac{N_{i,\cdot}}{n}$$

, allo stesso modo

$$\hat{q}_j = \frac{n^\circ \text{ di } Y = b_j}{n} = \frac{N_{\cdot,j}}{n}$$

in conclusione

$$n \cdot p_i \cdot q_j = n \frac{N_{i,\cdot}}{n} \frac{N_{\cdot,,j}}{n}$$

Distanza fra modello specificato da H_0 per (X, Y) e dati:

$$\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \frac{\left(N_{i,j} - \frac{N_{i,\cdot} N_{\cdot,j}}{n}\right)^2}{\frac{N_{i,\cdot} N_{\cdot,j}}{n}}$$

è la statistica test ST.

Se ho n grandi dati, rifiuto H_0 a livello di α se $ST \geq q_{\chi^2_?}(1-\alpha)$

i gradi della χ^2 è il numero delle classi $(r_1 \cdot r_2)$ - 1 - (numero parametri stimati nel modello sotto H_0)

$$= (r_1 - 1)(r_2 - 1)$$

- $H_0 : f(a, b) = p_i q_j$
- Abbiamo stimato tutte le p e tutte le q $(p_1, ..., p_{r_1})$ e $(q_1, ..., q_{r_2})$, ma sono correlate tra loro, ovvero
  - $p_{r_1 - 1} = 1 - \sum_{i=1}^{r_1 - 1} p_i$
  - $q_{r_2 - 1} = 1 - \sum_{j=1}^{r_2 - 1} q_j$

$ST \geq q_{\chi^2_{(r_1-1)(r_2-2)}}(1-\alpha)$

HP: Quante osservazioni devo avere? Almeno 30 osservazione ed in ogni casella devo avere almeno 5 osservazioni.

Finora ho ragionato come se le X e Y fossero variabili categoriche discrete con 5 e 7 modalità (?)

Distribution-free perchè non dipende dalla distribuzione in ipotesi H0

32.2 Test di buon adattamento Kolmogorov-Smirnov

- Valido quando ho pochi dati
- Test per **dati grezzi**.
- $H_0 : X \sim F_0 \quad H_1 : X \not\sim F_0$
- dati: $X_1, ..., X_n$ iid $\sim F$
- HP su F_0 : F_0è continua, completamente specificata (non ci sono parametri da stimare)

Il test si basa sull'uso di uno stimatore che funziona bene anche quando abbiamo solo i dati. In questa sitauzione di completa ignoranza del sistema generatore di dati, posso utilizzare la funzione di ripartizione empirica. Quindi il test di KS utilizza la funzione di ripartizione empirica o campionaria, che cosa è?

- $X_1, ..., X_n$
- $\hat{F}_n(x) = \frac{\#X_j \leq x}{n} \quad \forall x \in R$ x deve essere nel reale perchè

$$F(x) = P(X \leq x) \quad \forall x \in R$$

- Proprietà di $\hat{F}_n$
  - numeratore di $\hat{F}_n(x*) = \#X_j \leq x*$ su n repliche di un esperimento iid conto quante volte si è verificato quel fatto. Che distribuzione ha questa statistica? Binomiale,

$$\#X_j \leq x \sim Bin(n, F(x*))$$

  - $\Longrightarrow E\left[\hat{F}_n(x*)\right] = \frac{nF(n*)}{n} = F(n*)$ la ripartizione empirica **puntualmente** è non distorta.
  - $\Longrightarrow Var\left[\hat{F}_n(x*)\right] = \frac{nF(x*)(1-F(x*))}{n^2}$, quindi con $n \to \infty$ la varianza va 0, quindi lo stimatore è consistente in media quadratica. Questa proprietà mi dice che la ripartizione empirica si avvicina a quella teorica
  - Qual'è la distribuzione asintotica? Una normale $\sim N\left(F(x*), \frac{F(x*)(1-F(x*))}{n}\right)$
- Abbiano notato la convergenza puntuale, quella globale? sì all'aumentare del numero di osservazioni.

$$D_n = \sup_{x \in R} \left|\hat{F}_n(x) - F(x)\right|$$

Il teorema di Glivenco-Cantelli

$$\lim_{n \to +\infty} D_n = 0 \text{ con probabilità 1}$$

32.2.1 Esempio (Es 2.2 dispensa non parametrica)

[FIG 1]

- I punti di salto della ripartizione empirica sono in corrispondenza dei dati (quindi quali dati ho osservato).
- Il salto è la frequenza relativa della modalità.
- Dalla Fdr posso ricavare la media (e varianza) campionaria.

$$\bar{X} = \sum x_i \cdot salto_i$$

Cos'ha in meno la Fdr empirica? L'ordine dei dati, informazione che non ho mai usato in tutto il corso perchè ho utilizzato dati iid.

Capire bene la **distribution-free**.

Rifiuto H_0 se $D_n > q_{KS}\left(1-\alpha\right)$

[FIG 2]

- A destra x_i : $\left|\hat{F}_n\left(x_i\right) - F_0\left(x_i\right)\right|$
- A sinistra di x_i : $\left|\hat{F}_n\left(x_{i-1}\right) - F_0\left(x_i\right)\right|$

33 Esercitazioni

33.1 Esercizio

Tipi di cioccolato

- 70% prima qualità
- 20% seconda qualità

- 10% terza qualità

Abbiamo $n = 200$

- 130 prima qualità
- 45 seconda qualità
- 25 terza qualità

Questa distribuzione rappresenta l'andamento tipico?

$$X = \begin{cases} 1 & p = 0.7 \\ 2 & p = 0.2 \\ 3 & p = 0.1 \end{cases}$$ Il campione corrisponde a questa distribuzione teorica?

classi	I qualità	II qualità	III qualità
N_i	130	65	25
$n\theta_i$	140	40	20

- N_i = numerosità osservate
- $n\theta_i$ = numerosità teoriche

La ST è un numero che dice quanto queste coppie di valori sono simili o diverse, come è costruita? tutti i test \chi^2 il criterio è lo stesso: ho una fila di numerosità osservate ed una fila di num teoriche, come si comportano?

$$Q = \sum_i^n \frac{(N_i - n\theta_i)^2}{n\theta_i}$$

In questo esercizio

$$\begin{aligned} Q &= \frac{(130-140)^2}{140} + \frac{(45-40)^2}{40} + \frac{(25-20)^2}{20} \\ &= 2.7 \end{aligned}$$

Come funziona ora il test \chi^2? L'idea qual'è? Se c'è adattamento le due numerosità devono coincidere, altrimenti ci sarà differenza. Se sono simili Q sarà piccolo. Morale: rifiuto H_0 se Q è troppo grande, ovvero se il valore osservato della ST finisce della coda destra della χ^2_{k-1-m}

- k = numero di classi (3)
- m = numero di parametri che abbiamo dovuto stimare (0)

Al livello α rifiuto H_0 se $2.7 > \chi^2_2(1-\alpha)$

Qual'è il p-value? Generalmente è $P\left(\underbrace{\chi^2}_{\text{quella coinvolta}} > q\right)$, nel nostro caso $P(\chi^2_2 > 2.7)$. Non è possibile ricavare il p-value dalle tavole, quindi devo ricordare la parentela della \chi^2 con le gamma

$$\begin{aligned} P\left(\chi^2_2 \geq 2.7\right) &= P(\Gamma(1,2) > 2.7) \\ &= P(\varepsilon(2) > 2.7) \\ &= 1 - F_\varepsilon(2.7) \\ &= 1 - \left[1 - e^{-\frac{2.7}{2}}\right] \end{aligned}$$

33.2 Esercizio

[FIG 2] tabella delle osservazioni

Eseguo il test d'indipendenza χ^2. Nei test d'indipendenza le ipotesi sono già fatte H_0 : c'è indipendenza, contro H_1 : non c'è indipendenza. Nei test di buon adattamento più o meno siamo nella stessa situazione.

Se sono indipendenti $P(XY) = P(X) \cdot P(Y)$

Alla luce di questa tabella la migliore stima di essere sufficienti al secondo compitino è $\frac{89}{116}$

$$q = \frac{(19-8.38)^2}{8.38} + \frac{(17-27.62)^2}{27.62} + \frac{(8-18.62)^2}{18.62} + \frac{(72-61.38)^2}{61.38} \cong 22 \pm 1$$

Rifiuto H_0 se $25.44 > \chi^2(1-\alpha)$

I gradi di libertà di χ^2 sono $1 = (col-1)(row-1)$ se osserviamo come la tabella teorica è stata costruita osserviamo che effettivamente abbiamo calcolato un solo valore, gli altri li abbiamo derivati.

33.3 Esercizio

[FIG 3]

$ST = \frac{(34-28.27)^2}{28.27} + ... + \frac{(16-10.27)^2}{10.27} = 8.5$

Lo confronto con la χ^2 ad un grado di libertà. All'1% rifiuto H_0 se $8.5 > \underbrace{\chi^2(0.99)}_{66}$ quindi rifiuto H_0

Attenzione a non confondere "non indipendenza" con "causa effetto" (esempio scuola elementare)

p-value:

$$\begin{aligned} P\left(\chi_1^2 > 8.5\right) \sim \Gamma\left(\frac{1}{2}, 2\right) &= P\left(Z^2 > 8.5\right) \\ &= P\left(|Z| > 2.91\right) \\ &= 2\left(1 - \Phi\left(2.91\right)\right) \end{aligned}$$

33.4 Esercizio

Centralino

- 113 volte abbiamo registrato 0 telefonate
- 170 volte 1 telefonata
- 180 volte 2 telefonate
- 68 volte 3 telefonate
- 32 volte 4 telefonate
- 5 volte 5 telefonate
- 1 volta 6 telefonate
- 1 volta 7 telefonate (nell'arco di un minuto)

Se X = numero di telefonate in 1 minuto, può essere che $X \sim P(\lambda = 1.4)$

Non posso usare il test di KS perchè tra le ipotesi è richiesto che la X sia una funzione continua

classi	0	1	2	3	4	5	6	7
N_i	113	170	190	68	32	5	1	1
	$520 \cdot \underbrace{P(X=0)}_{e^{-1.4}}$	$520 \cdot \underbrace{P(X=1)}_{1.4e^{-1.4}}$						

- $n = 520$

Ma c'è un problema, come regola empirica abbiamo che la numerosità **teorica** non deve essere inferiore a 5, in questa situazione si raggruppa

classi	0	1	2	3	4	5 o più
N_i	113	170	190	68	32	7
	$520 \cdot \underbrace{P(X=0)}_{e^{-1.4}}$	$520 \cdot \underbrace{P(X=1)}_{1.4e^{-1.4}}$				

- $q = 10.4$
- Rifiuto H_0 se $10.4 > \chi^2_5(1-\alpha)$

Esercizio finito, ma lo rifacciamo cambiando la domanda in $X \sim P$? La differenza è che devo ricavare λ sulla base dei dati che abbiamo. Quanto vale λ ? Con le osservazioni a disposizione, quanto vale λ ?

$$\hat{\lambda} = \bar{X} = \frac{113 \cdot 0 + 170 \cdot 1 + 190 \cdot 2 + ... + 1 \cdot 7}{520} = 1.54$$

classi	0	1	2	3	4	5 o più
N_i	113	170	190	68	32	7
	$520 \cdot \underbrace{P(X=0)}_{1.54e^{-1.54}}$	$520 \cdot \underbrace{P(X=1)}_{1.54e^{-1.54}}$				

- $q = 4.67$
- Rifiuto H_0 se $4.67 > \chi^2_4(1-\alpha)$

- p-value:

$$\begin{aligned} P\left(\chi_4^2 > 4.67\right) &= P\left(\underbrace{\Gamma\left(2,2\right)}_{Y} > 4.67\right) \\ &= \int_{4.67}^{+\infty} f_y\left(y\right) dy \\ &= \int_{4.67}^{+\infty} \frac{1}{4} y e^{-\frac{y}{2}} \end{aligned}$$

  - $\Gamma\left(2\right) = 1$, perchè $\Gamma\left(\alpha\right) = \left(\alpha - 1\right)!$

33.5 Esercizio

- X_i : 0.9, 0.7, 1.2, 0.1, 1
- $X \sim \varepsilon\left(\frac{1}{2}\right)$?
- Si nota che è un test di buon adattamento; ne conosciamo solo due.
- Va bene KS perchè soddisfa le ipotesi, comunque il test della χ^2 non posso farlo perchè ho bisogno di un campione numeroso
  - Dai dati costruisco la Fdr campionaria, ovvero una Fdr composta da 5 scalini messi in corrispondenza dei valori ordinati dei dati ricevuti. [FIG 4]
  - C'è buon adattamento se la Fdr empirica e quella teorica sono vicine.
  - Devo calcolare le distanze. E ricavare il sup
    - Il sup non può essere raggiunto in un punto intermedio del gradino, ma solo in corrispondenza del gradino perchè la F è una funzione crescente.

x_i	F_x	* $\hat{F}_-$	* $\hat{F}_+$	$\lvert diff\rvert_-$	$\lvert diff\rvert_+$
0.1	$1 - e^{-2(0.1)} = 0.18$	0	0.2	0.18	0.02
0.7	0.75	0.2	0.4	**0.55**	0.35
0.9	0.83	0.4	0.6	0.43	0.23
1	0.86	0.6	0.8	0.26	0.06
1.2	0.91	0.8	1	0.11	0.09

* calcoliamo il limite destro e sinistro della $\hat{F}$

- $q = 0.55$ rifiuto H_0 se q è troppo grande. Con $n = 5, \alpha = 0.1 \rightarrow 0.5094$ allora rifiuto H_0 cioè non è un esponenziale di parametro $\frac{1}{2}$

34 Schema

1. Test di buon adattamento
   (a) χ^2 di Pearson (grandi campioni, F può avere parametri)
   (b) K-S (F deve essere continua)
2. Test di indipendenza
   (a) χ^2 di Pearson
   (b) ρ (Gaussiana bivariata)
3. Test di omogeneità non parametrico
   (a) Con binomiale (Wilcoxon)
   (b) Ranghi (Mann-Whitney)
4. Test di omogeneità parametrico
   (a) Due passi per Normale ($\sigma^2 - \mu$)

35 Test di Lilliefors

- $H_0 : X \sim N \quad H_1 : X \not\sim N$
- $X_1, ..., X_n$ iid
- Non posso utilizzare il test di adattamento di χ^2 perchè ho pochi dati. Non sono specificati i parametri μ e σ^2 , pertanto non posso usare K-S, ma ne uso una variante.

35.1 Procedimento

1. Stimo media e varianza (parametri della normale) e li stimo con media e varianza campionaria.
2. Trasformo i dati nel seguente modo

$$Z_1 = \frac{X_1 - \bar{X}}{\sqrt{S^2}}, Z_2 = \frac{X_2 - \bar{X}}{\sqrt{S^2}}, ...$$

   è un campione **non** indipendente
3. Se avessi operato la vera standardizzazione Z sarebbe una normale (0,1). Questo non è ora più vero, ma la normale(0,1) può essere una buona approssimazione sotto H0.
4. $H_0 : Z = \frac{X_j - \bar{X}}{\sqrt{S^2}} \sim N(0,1)$ allora calcolo la distanza tra Z e la normale con il K-S

$$D^* = \sup_{z \in R} \left| \hat{F}_{n,z}(z) - \Phi(z) \right|$$

5. Quali tavole devo consultare? Non posso usare K-S perchè il campione non è indipendente
6. Rifiuto H_0 se $D^* \geq q_L(1-\alpha)$ quantile che vado a leggere sulle tavole di Lillifors

36 Esercizi

Ricevimento studenti: email

36.1 Tema 0910 - esercizio 5.3

36.1.1 Punto 1

- Nella tavola di L ragiono come K-S
- Rifiuto H_0 cioè X non è normale

36.1.2 Punto 2

- H_0 : X,Y seguono lo stesso modello
- H_1 : X,Y non seguono lo stesso modello
- (sapendo che X ed Y non sono normali perchè prima l'ho confutato)

36.2 Eserciziario 5.2.3

36.2.1 Esercizio per casa

- $E\left[\frac{\bar{X}}{\bar{Y}}\right]$?
- $Var\left[\frac{\bar{X}}{\bar{Y}}\right]$?

36.3 Eserciziario 5.1.11

36.4 Eserciziario 4.2.7

Part II

Probabilità e Statistica (Ross)

37 La distribuzione delle statistiche campionarie

37.1 Introduzione

La statistica è la scienza che si occupa di trarre conclusioni dai dati sperimentali. Una situazione tipica con la quale bisogna spesso confrontarsi negli ambiti tecnologici, è quella in cui si studia un insieme molto grande,

detto *popolazione*, di oggetti a cui sono associate delle quantità misurabili. L'approccio statistico consiste nel selezionare un sottoinsieme ridotto di oggetti, che viene detto *campione*, e analizzarlo sperando di essere in grado di trarre da esso delle conclusioni valide per la popolazione nel suo insieme.

Per basare sui dati del campione delle inferenze che riguardino l'intera popolazione, è necessario assumere qualche condizione sulle relazioni che legano questi due insiemi. Un'ipotesi fondamentale (in molti casi del tutto ragionevole) è che vi sia una (implicita) distribuzione di probabilità della popolazione, nel senso che se da essa si estraggono degli oggetti in maniera casuale, le quantità numeriche loro associate possono essere pensate come variabili aleatorie indipendenti, tutte con tale distribuzione. Se tutto il campione viene selezionato in maniera casuale, sembra ragionevole supporre che i suoi dati siano valori indipendenti provenienti da tale distribuzione.

Definizione. Un insieme $X_1, ..., X_n$ di variabili aleatorie indipendenti, tute con la stessa distribuzione F, si dice *campione* o *campione aleatorio* della distribuzione F.

In pratica la distribuzione F non è mai completamente nota, però è possibile usare i dati per fare dell'*inferenza* su F. In alcuni casi è possibile che F sia nota eccetto che per dei parametri incogniti (si potrebbe ad esempio sapere che F è una distribuzione normale, ma non conoscerne la media e la varianza; oppure F potrebbe essere di Poisson, ma con parametro incognito); in altri casi potremmo non sapere praticamente nulla di F (tranne forse assumere che essa sia continua, oppure discreta). I problemi in cui la distribuzione F è nota a meno di un insieme di parametri incogniti sono detti problemi di inferenza *parametrica*; quelli in cui nulla si sa della distribuzione F sono invece problemi di inferenza *non parametrica*.

37.2 Media Campionaria

$$\bar{X} := \frac{X_1 + X_2 + ... + X_n}{n}$$

$$\begin{aligned} E\left(\bar{X}\right) &= E\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) \\ &\overbrace{=}^{\text{indip}} \frac{E\left(X_1\right) + E\left(X_2\right) + ... + E\left(X_n\right)}{n} \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned}$$

$$
\begin{aligned}
Var\left(\bar{X}\right) &= Var\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) \\
&\overbrace{=}^{\text{indip}} \frac{Var\left(X_1\right) + Var\left(X_2\right) + ... + Var\left(X_n\right)}{n^2} \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

La media campionaria ha quindi lo stesso valore atteso della distribuzione da stimare, mentre la sua varianza risulta ridotta di un fattore n. Da questo possiamo dedurre che $\bar{X}$ è centrata in μ, e la sua variabilità si riduce sempre di più all'aumentare di n.

Part III

Esercizi

38 Calcolo delle Probabilità

38.1 Variabili aleatorie gaussiane

38.1.1 Esercizio 1.1.1 eserciziario

Soluzione punto 1 $P\left(Z \leq 1.51\right) \rightarrow$ si guarda direttamente il quantile sulle tavole $= 0.9345$

$P\left(Z \leq -1.51\right) = 1 - P\left(Z \leq 1.51\right) = 1 - 0.9345 = 0.0655$

$P\left(Z > 0.31\right) = 1 - P\left(Z < 0.31\right) = 1 - 0.6217 = 0.5783$

$P\left(|Z| \leq 0.81\right) = P\left(-0.81 \leq Z \leq 0.81\right) = P\left(Z \leq 0.81\right) - P\left(Z \leq -0.81\right) = 0.7910 - \left(1 - P\left(Z \leq 0.81\right)\right) = 2 \cdot 0.7910 - 1 = 0.582$

$P\left(0.31 < Z < 1.51\right) = P\left(Z < 1.51\right) - P\left(Z < 0.31\right) = 0.9345 - 0.6217 = 0.3128$

Soluzione punto 2 Data la simmetricità della normale $z_{0.95} = z_{0.05} \to 1.64 \sim 1.65$

Soluzione punto 3 (?)

Soluzione punto 4 (? a,b) $P(X < x_{0.95}) = 0.95 \to P\left(Z < \frac{x_{0.95}-20}{4}\right) = 0.95 \to \frac{x_{0.95}-20}{4} = z_{0.95} \to x_{0.95} = 4 \cdot 1.65 + 20 = 26.6$

$P(X < x_{0.05}) = 0.05 \to P\left(Z < \frac{x_{0.05}-20}{4}\right) = 0.05 \to \frac{x_{0.05}-20}{4} = z_{0.05} \to x_{0.95} = 4 \cdot 0.5199 + 20 = 22.0796$

$P(X > 6) = P\left(Z > \frac{6-20}{4}\right) = P(Z > -3.5) = P(Z < 3.5) = 0.9998$

38.1.2 Esercizio 1.1.2 eserciziario

Soluzione punto 1 Secondo il teorema 5.1 on page 5 la soluzione alla prima parte è

$$W \sim N\left(\mu_X - \mu_Y, \underbrace{25}_{\sigma_1^2+\sigma_2^2}\right)$$

da notare che la varianza aumenta sempre.

Soluzione punto 2 Nel caso $\mu_X = \mu_Y \to W \sim N(0, 25)$

$$P(W > -2) = P\left(Z > \underbrace{-\frac{2-0}{5}}_{=-0.4}\right) = P(Z < 0.4) = 0.6554$$

39 Intervalli di confidenza

39.0.3 Esercizio 2.1.1

1. punto 3: $z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma_0^2}{2}} < 0.01 \to n > \left(\frac{1.96}{0.01}\right)^2 \cdot 0.01 = 384$
2. punto 4: $z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma_0^2}{2}} < 0.02 \to n > \left(\frac{1.96}{0.02}\right)^2 \cdot 0.01 = 96$

39.0.4 Esercizio 2.1.2

Punto 2

- $M_2 = \frac{\sum_j^n X_j^2}{n}$
- $S^2 = \frac{\sum_j^n \left(X_j - \bar{X}\right)^2}{n-1} = \frac{\sum_j^n X_j^2 - n\bar{X}^2}{n-1}$
- $E\left[X^2\right] = Var\left[X\right] - E\left[X\right]^2 \rightarrow \frac{\overbrace{\sum_j X_j^2}^{nM_2} - n\bar{X}^2}{n-1} = \frac{nM_2 - n\bar{X}^2}{n-1} = \frac{n}{n-1}\left(M_2 - \bar{X}^2\right)$

Punto 3 La domanda scritta in testo può essere tradotta in $P\left(\theta \geq 57435\right) = \gamma$? Che è un IC (c, ∞) di livello γ per $\theta = \frac{\mu}{2}$ con σ^2 incognita e dati gaussiani. Devo partire da una statistica test che comprenda quella variabile θ .

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{24}$$

$$P\left(\frac{\bar{X} - 2\theta}{\sqrt{\frac{S^2}{n}}} \overbrace{\leq}^{*} q_{t_{24}}\left(\gamma\right)\right) = \gamma$$

* ho messo $\leq$ perchè devo girare quella diseguaglianza fino a farmi venire un $\theta > \square$, in modo tale da poter utilizzare l'informazione della domanda che $\theta \geq 57435$

$$\theta \geq \underbrace{\frac{\bar{X} - q_{t_{24}}\left(\gamma\right)\sqrt{\frac{S^2}{25}}}{2}}_{=57435}$$

da cui ricavo $\gamma = 0.95$

39.0.5 Esercizio 2.1.3

Lunghezza di un IC simmetrico col campione proveniente da una popolazione normale

$$2z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\sigma^2}{n}}$$

Secondo quanto riportato dal testo: $2z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\sigma^2}{n}} < \sigma \rightarrow 2z_{\frac{1+\gamma}{2}} \cdot \frac{1}{\sqrt{20}} < 1$ da cui ricavo γ

39.0.6 Esercizio 2.1.4

Punto 1

- $-z_{\frac{1+\gamma}{2}} < \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}} < z_{\frac{1+\gamma}{2}} \rightarrow \bar{X} - 1.96 \cdot 0.8 < \mu < \bar{X} + 1.96 \cdot 0.8$
- $\mu \in [171.43; 174.57]$
- Il motivo risiede nel fatto che il campione è composto prevalentemente da soggetti appartenenti alla parte destra dela funzione di distribuzione.

Punto 2

- $170 - 1.96 \cdot 0.8 < \bar{X} < 170 + 1.96 \cdot 0.8$
- $\bar{X} \in [168.43; 171.57]$

39.0.7 Esercizio 2.1.5

Punto 1

- $t_{n-1}\left(\frac{1+\gamma}{2}\right) \cdot \sqrt{\frac{S^2}{n}} = t_{19}\left(0.975\right)\sqrt{\frac{0.4}{20}} = 0.296$
- $\mu \in [0.934; 1.526]$

Punto 2

- $z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\sigma_0^2}{n}} = z_{0.975}\sqrt{\frac{0.4}{20}} = 0.277$
- $\mu \in [0.953; 1.507]$

39.0.8 Esercizio 2.1.6

- $\bar{X} = 2.1 \cdot 10^{-4}$
- $S^2 = 5.2 \cdot 10^{-5}$
- **Hp**: campioni indipendenti ed identicamente distribuiti (iid); varianza e media non note

Punto 1

- $\frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1+\gamma}{2}\right)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1-\gamma}{2}\right)} \rightarrow \frac{9 \cdot 5.2 \cdot 10^{-5}}{19.023} < \sigma^2 < \frac{9 \cdot 5.2 \cdot 10^{-5}}{2.700}$
- $\sigma^2 \in [2.46 \cdot 10^{-5}; 1.73 \cdot 10^{-4}]$

Punto 2

- $\sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1}(1-\gamma)} \rightarrow \sigma^2 < \frac{9 \cdot 5.2 \cdot 10^{-5}}{3.325}$
- $\sigma^2 \in [0; 1.41 \cdot 10^{-4}]$

Punto 3

- $\sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1}(\gamma)} \rightarrow \sigma^2 < \frac{9 \cdot 5.2 \cdot 10^{-5}}{16.919}$
- $\sigma^2 \in [2.77 \cdot 10^{-5}; \infty]$

40 Verifica d'ipotesi

40.1 Test di omogeneità

40.1.1 Esercizio 5.2.1 (esercizio del Ross)

Ipotesi da imporre

- F e G continue
- X ed Y indipendenti e iid
- I dati non presentano ripetizioni

Risoluzione

- $X \sim F, \quad Y \sim G$
- $H_0 : F = G, \quad H_1 : F > G$
  - F: fdr delle città con campagnia pubblicitaria
  - G: fdr delle città senza campagna pubblicitaria
  - $\mu_X < \mu_Y \iff F > G$

1	2	3	4	5	6	7	8	9	10	11	12	13	14
19 x	28 y	31 x	36 y	39 x	44 y	45 x	47 x	49 y	52 y	60 y	66 x	74 x	81 x

- $T_x = \sum_i r_i = 63$ ipotizzando $m = 8$ e $n = 6$
- p-value: $P(T_x < 63)$ e rifiuto H_0 se $\underbrace{P(T_x < 63)}_{\alpha \text{ reale}} < \underbrace{P(T_x < w_{8,6}(10\%))}_{\alpha \text{ accettabile}}$
  - $P(T_x < 63)$ non è presente sulle tavole, posso solo dire che è maggiore del 10%
  - $P(T_x < w_{8,6}(10\%)) = P(T_x < 50) = 10\%$
- Dato che il p-value è maggiore dell'errore massimo accettabile (α) allora accetto H_0 .

40.1.2 Esercizio 5.2.3

Punto 1

- $H_0 : F_A = F_B \quad H_1 : F_A < F_B$
- $\alpha = 0.05$
- $T_A = \sum r_a = 98$
- Rifiuto H_0 se $T_A > w_{1-0.05} = w_{0.95}, \quad m = 8, n = 10$
  - $w_{0.95} = m\,(m + n + 1) - w_{0.05} = 95$
  - $T_A > 95$ pertanto rifiuto H_0

40.1.3 Esercizio 5.2.6

Soluzione: sulla base dei dati, con un livello di significatività del 10% accettiamo l'ipotesi alternativa, che le due diverse macchine abbiano precisione diversa.