Contents

1 Introduzione

In probabilità quello che facciamo noi è quello di supporre che le nostre distribuzioni siano **note** in statistica facciamo il contrario, ossia dire qualcosa (anche detto *fare dell'inferenza*) su **parametri sconosciuti**.
Dato che i parametri sono sconosciuti il massimo che possiamo fare è quello di ottenere *una stima* dei parametri *incogniti*.

Questi sono chiamati **stimatori puntuali** e sono indicati con il simbolo $\hat{\boldsymbol{\theta}}$ (in questo caso stiamo parlando di uno stimatore del parametro incognito $\boldsymbol{\theta}$)

Esisono anche gli *stimatori non puntuali*, noti come **intervalli di confidenza**, ossia un intervallo di valori in cui può essere contenuto il *dato incognito*.

Esempio $\hat{\theta}?$ Altezza della popolazione

$X_1 = 1.7$
$X_2 = 1.82$
$X_3 = 1.73$
$X_4 = 1.7$
$X_5 = 1.8$

Possibile soluzione :

$$\hat{\theta_a} = \frac{1}{n}\sum_{4}^{5} x_i = \frac{1.7 + 1.82 + 1.73 + 1.7 + 1.8}{5} = \frac{8.75}{5} = 1.75$$

$$\hat{\theta_b} = \frac{min(x_i) + \max(x_i)}{2} = \frac{3.52}{2} = 1.76$$

$$\hat{\theta_c} = \frac{1}{3}\sum_{2}^{4} x_i = \frac{1}{3}(1.8 + 1.73 + 1.7) = \frac{5.23}{3} = 1.743$$

Scartiamo il più *piccolo* e il *massimo*, calcolando poi la **media** dei rimanenti

2 MLE

Definizione: Stima a Massima Verosomiglianza (Maximum Likelihood Estimation)
Questa classe di stimatori sono molto usati in statistica, servono per determinare i migliori parametri del modello che si adattano ai dati e comparare molteplici modelli per *determinare* quello che si adatta di più ai dati.
Ad esempio la stima di massima verosomiglianza $\hat{\theta}$ è definita come il valore di θ che rende massima $\boldsymbol{f(x_1, x_2, \ldots, x_n|\theta)}$ $\rightarrow$ anche detta *funziona di likelihood*

Likelihood: avendo dei dati quale è la probabilità che un certo modello descriva al meglio la natura dei nostri dati

$$\hat{\theta} = argmaxL(\theta) = argmax[f(X_1 \ldots X_n/\theta)]$$

Stima parametrica (Point) Parametric Estimation

Ipotesi: - Esiste un parametro θ incognito n dati a disposizione $\{X_1, X_2, X_n\}$
Legge di probabilità che descrive il fenomeno che ha generato i dati

Formula generica: Bayes

$$P(\theta/X_1 \ldots X_n) = \frac{P(X_1 \ldots X_n/\theta)P(\theta)}{P(X_1 \ldots X_n)}$$

Verosomiglianza (likelihood)

2.1 MLE di una Bernoulliana

Vengono realizzate *n* prove indipendenti con probabilità p di successo

$$X_i = \begin{cases} 1 & \text{se la prova i-esima ha successo} \\ 0 & \text{altrimenti} \end{cases}$$

La distribuzione dell X_i è la seguente:

$$P(X_i = k) = p^k(1-p)^{1-k}, \qquad k \in \{0, 1\}$$

La likelihood (ossia la *funzione di massa congiunta*) è:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|p) &:= P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n|p) \\ &= p^{x1}(1-p)^{1-x1} \ldots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_i x_1}(1-p)^{n-\sum_i x_1} \qquad x_1 = 0, 1 \qquad i = 1, \ldots, n \end{aligned}$$

Possiamo derivare rispetto a p:

$$\frac{d}{dp} \log f(x_1, x_2, \ldots, x_n|p) = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_1 \right)$$

Da questo bro possiamo ottenere un'espressione per la stima $\hat{\boldsymbol{p}}$:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2.2 MLE di una Poisson

La funzione di *likelihood* è data da:

$$f(x_1, x_2 \dots x_n/\lambda) = \frac{\lambda^{x_1} e^{-y}}{x_1!} \dots \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$
$$= \frac{\lambda^{\sum_i x_i} e^{-\lambda}}{x_1! \dots x_n!}$$

Come sempre deriviamo e otteniamo:

$$\frac{d}{d\lambda} \log f(x_1, x_2, \dots, x_n|\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

Da questo bro possiamo ottenere un'espressione per la stima $\hat{\lambda}$:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La stessa formula può essere applicata al campione $X_1, X_2, \dots, X_n$:

$$P\{X_i = 1\} = 1 - P\{X_i = 0\}$$

Esempio Numero di incidenti stradali in 10 giornate senza pioggia
Dataset: { 4 0 6 5 2 1 2 0 4 3 }
Si vuole stimare per quell'anno la frazione di giornate senza pioggia con *2 incidenti o meno*

$$\overline{X} = \frac{1}{10} \sum_{i=1}^{10} X_i = \mathbf{2.7}$$

Cosi otteniamo che la media della poissoniana è 2.7, la stima desiderata è data da:

$$(1 + 2.7 + (2.7)^2/2)e^{-2.7} \approx 0.4936$$

2.3 MLE distribuzione Uniforme

$$f(X_1, \dots X_n|\theta) = \begin{cases} \frac{1}{\theta} & 0 < x_1 < \theta \\ 0 & \text{altrimenti} \end{cases}$$

La formula per la stima di θ:

$$\hat{\theta} = \max\{X_1, \dots, X_n\}$$

$$\hat{\theta}_{\frac{\text{MLE}}{2}} = \textit{media}$$

2.4 MLE distribuzione Normale

Definizione: La distribuzione normale ha media μ e dev. st. σ **incognite**
La densità congiunta (la likelihood) è data da:

$$f(x_1, x_2, \dots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\mu\sigma}} \exp\left\{-\frac{(x_1 - \mu)^2}{2\sigma^2}\right\}$$

La log-likelihood è data da:

$$\log f(x_1, x_2, \dots, x_n|\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

La risoluzione ci porta alle formule per le stime:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

3 Teorema del limite centrale

3.1 Definizione

Questo teorema afferma che la somma di un numero elevato di **var. aleatorie indipendenti** tende ad avere una distribuzione approssimativamente normale.
Quindi un campione(insieme di var. aleatorie da $X_1, X_2 \ldots, X_n$) può essere trasformato in una Normale Standard:

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

4 Intervalli di confidenza

4.1 Distribuzione normale

4.1.1 μ incognita e varianza σ^2 nota

Sia $\boldsymbol{X_1, X_2, \ldots, X_n}$ un campione di una popolazione normale con μ *incognita* e varianza σ^2 *nota*

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

Intervallo di confidenza per la media:

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Il 95% circa delle volte μ starà a una distanza non superiore a 1.96 $\sigma/\sqrt{n}$ dalla media aritmetica dei dati. Se osserviamo il campione, e registriamo che $\overline{X} = \overline{x}$, allora possiamo dire che "con il 95% di confidenza"

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Questo intervallo è detto *intervallo di confidenza* ad un livello del 95%

intervallo destro:

$$(\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \quad +\infty)$$

intervallo sinistro:

$$(-\infty, \quad \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}})$$

Esempio segnale elettrico di valore μ
i valori registrati sono i seguenti: 5 8.5 12 15 7 9 7.5 6.5 10.5
Otteniamo $\overline{\boldsymbol{x}}$:

$$\overline{x} = \frac{81}{9} = 9$$

Un intervallo di confidenza al 95% per μ è

$$\left(9 - 1.96\frac{2}{3}, \quad 9 + 1.96\frac{2}{3}\right) = (7.69, 10.31)$$

Otteniamo quindi il 95% di fiducia che il messaggio fosse **compreso** tra 7.69 e 10.31

4.1.2 μ incognita e varianza σ^2 incognita

Dato che tutti i nostri parametri sono ignoti, non possiamo basarci sul fatto che $\sqrt{n}(\overline{X} - \mu)/\sigma$ è una *normale standard*, dobbiamo quindi ricorrere a una varianza campionaria come segue:

$$S^2 := \frac{1}{n-1}\sum_i (X_i - \overline{X})^2 \longrightarrow \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Alla fine otteniamo una variabile aleatoria di tipo t con n-1 gradi di libertà

Per Bilaterale

$$P\left\{\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\frac{\alpha}{2},n_1}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha$$

4.2 Intervalli di confidenza per Bernoulli

Nel caso avessimo n oggetti con una quantita X di oggetti che soddisfano i requisiti, possiamo dire che X ha distribuzione *binomiale* di parametri n e p

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1)$$

Per ottenere un intervallo per p denotiamo con $\hat{p} := X/n$ la frazione degli oggetti del campione che soddisfano i requisiti, quindi:

$$\frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} \sim \mathcal{N}(0,1)$$

Da questa formula possiamo ottenere cosi un intervallo di confidenza

Per caso Bilaterale

$$1 - \alpha = \left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}\right)$$

Per caso Unilaterale

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n} < p\right) \quad \left(p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}\right)$$

Esempio Un campione di *100* transitor viene testato. *80* pezzi sono *adeguati* Volendo trovare un intervallo del 95% per la percentuale p scriviamo:

$$\left(0.8 - 1.96\sqrt{0.8 \cdot 0.2/100}, \quad 0.8 + 1.96\sqrt{0.8 \cdot 0.2/100}\right) = (0.7216, \quad 0.8784)$$

Possiamo dire quindi con il 95% di confidenza che sarà *accettabile* una percentuale compresa tra il **72.16%** e il **87.84%**

Tipo di intervallo	Intervallo di confidenza
Bilaterale	$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}$
Unilaterale sinistro	$\left(-\infty, \quad \hat{p} + z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}\right)$
Unilaterale destro	$\left(\hat{p} - z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}, \quad \infty\right)$

4.3 Metodo Montecarlo

supponendo di avere una funzione f da $\mathbb{R}^r$ in $\mathbb{R}$ e vogliamo stimare la quantità θ:

$$\theta := \int_0^1 \int_0^1 \cdots \int_0^1 f(y_1, y_2, \ldots, y_n)\, dy_1\, dy_2 \ldots dy_n$$

Possiamo notare che $U_1, U_2, \ldots, U_r$ sono var. al. *uniformi* su 0,1 quindi:

$$\mathbb{E}\left[f(U_1, U_2, \ldots, U_r)\right] = \theta$$

Se produciamo un numero casuale distribuito come la funzione e lo ripetiamo n volte, possiamo stimare $\boldsymbol{\theta}$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Esempio pensiamo alla stima di questo integrale:

$$\theta := \int_0^1 \sqrt{1-y^2}\,dy = \mathbb{E}\left[\sqrt{1-U^2}\right]$$

Se $\boldsymbol{U_1, U_2, \ldots, U_{100}}$ sono variabili aleatorie con tale distribuzione e *indipendenti* ponendo

$$X_i := \sqrt{1-U_i^2} \qquad i = 1, 2, \ldots, 100$$

Otteniamo un campione di **100** variabili aleatorie di media θ. Calcoliamo ora la *media campionaria*:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_1 = 0.786$$

e successivamente la *deviazione standard campionaria*:

$$S = 0.23$$

dato che $t_{0.025,99} \approx 1.985$ otteniamo che un intervallo di confidenza al 95% per θ è il seguente:

$$0.786 \pm 1.985 \cdot 0.023$$

Quindi il valore è compreso tra 0.740 e 0.832

5 Intervalli di predizione

5.1 Predizione di un elemento del mio campione

Supponiamo che $X_1, X_2, \dots, X_n, X_{n+1}$ sia un campione normale di media μ e varianza σ^2 entrambe *incognite*, dobbiamo prevedere l'elemento X_{n+1}:

ha come distribuzione:

$$\frac{X_{n+1} - \overline{X}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$$

Dato che σ è incognita dobbiamo sostituirla col suo stimatore (scegliendo la *deviazione standard campionaria* quindi poniamo:

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

quindi otteniamo

$$X_{n+1} \in \left(\overline{X}_n - t_{\frac{\alpha}{2}, n-1} S_n \sqrt{1 + \frac{1}{n}}, \overline{X}_n + t_{\frac{\alpha}{2}, n-1} S_n \sqrt{1 + \frac{1}{n}} \right)$$

Esempio prendiamo in campione i valori rilevati da un contapassi negli ultimi 7 giorni
Dataset: 6822 5333 7420 6252 7005 6752
Si trovi l'intervallo di predizione al 95% di confidenza
Risoluzione: le statistiche del campione sono:

$$\overline{X}_7 \approx 6716.57 \qquad S_7 \approx 733.97$$

Dalle tabelle ricaviamo che $t_{0.025,6} \approx 2.447$ (+ altri passaggi) concludiamo col dire che il 95% di confidenza che X_8 cadrà nell'intervallo [4796, 8637]

6 Intervalli di confidenza per la varianza

Se $X_1, X_2, \ldots, X_n$ è un campione di una distribuzione *normale* con parametri μ σ^2 **incogniti** ci possiamo basare sul fatto che

Formula generica:

$$(n-1)\frac{S^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}$$

Per caso Bilaterale :

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \quad \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right) \tag{1}$$

Per caso Unilaterale :

$$\left(\sigma^2 < \frac{(n-1)S^2}{\mathcal{X}^2_{1-\alpha,n-1}} \right) \quad \left(\frac{(n-1)S^2}{\mathcal{X}^2_{\alpha,n-1}} < \sigma^2 \right) \tag{2}$$

Tabella 7.1 Intervalli con livello di confidenza $1-\alpha$ per campioni normali.

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S := \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{1/2}$$

Ipotesi	θ	**Intervallo bilaterale**	**Intervallo sinistro**	**Intervallo destro**
σ^2 nota	μ	$\overline{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	$\left(-\infty, \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$	$\left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right)$
σ^2 non nota	μ	$\overline{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$	$\left(-\infty, \overline{X} + t_{\alpha, n-1} \frac{S}{\sqrt{n}}\right)$	$\left(\overline{X} - t_{\alpha, n-1} \frac{S}{\sqrt{n}}, \infty\right)$
μ non nota	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}\right)$	$\left(0, \quad \frac{(n-1)S^2}{\chi^2_{1-\alpha, n-1}}\right)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha, n-1}}, \quad \infty\right)$

6.1 Stime per la differenza tra le medie di due popolazioni normali

Siano $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ due campioni normali e differenti e denotiamo con μ_1 e σ_1^2 e con μ_2 e σ_2^2
$\boldsymbol{\overline{X} - \overline{Y}}$ è lo stimatore di massima verosomiglianza $\mu_1 - \mu_2$ Per ottenere uno *stimatore non puntuale*, dobbiamo conoscere la distribuzione di $\overline{X} - \overline{Y}$ poiche:

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n}\right) \qquad \text{e} \qquad \overline{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

Possiamo dedurre che:

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Ipotizzando di conoscere σ_1^2 e σ_2^2 abbiamo che:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0,1)$$

e possiamo dedurre, con i passaggi che ci sono ormai familiari, che

Per caso Bilaterale

$$\begin{aligned}
1 - \alpha &= \left(-z_{\frac{\alpha}{2}} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} < z_{\frac{\alpha}{2}}\right) \\
&= \left(\overline{X} - \overline{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)
\end{aligned}$$

Per caso Unilaterale

$$\left(\overline{X} - \overline{Y} - z_\alpha\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2\right)$$

$$\left(\mu_1 - \mu_2 < \overline{X} - \overline{Y} - z_\alpha\sqrt{\frac{\sigma_2^2}{n} + \frac{\sigma_2^2}{m}}\right)$$

Tabella 7.2 Intervalli di confidenza ad un livello di $1-\alpha$ per $\mu_1-\mu_2$, cioè la differenza tra le medie di due popolazioni normali.

$$X_i \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), i = 1, \dots, n \qquad Y_j \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right), j = 1, \dots, m$$

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \overline{Y} := \frac{1}{m}\sum_{j=1}^{m} Y_j$$

$$S_1^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \qquad S_2^2 := \frac{1}{m-1}\sum_{j=1}^{m}(Y_j-\overline{Y})^2$$

$$N := n+m-2 \qquad S_p := \sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{N}}$$

Si assume	Intervallo bilaterale	Intervallo sinistro
σ_1 e σ_2 note	$\overline{X}-\overline{Y} \pm z_{\frac{\alpha}{2}}\sqrt{\sigma_1^2/n+\sigma_2^2/m}$	$\left(-\infty, \quad \overline{X}-\overline{Y}+z_{\alpha}\sqrt{\sigma_1^2/n+\sigma_2^2/m}\right)$
σ_1 e σ_2 non note ma uguali	$\overline{X}-\overline{Y} \pm t_{\frac{\alpha}{2},N} \cdot S_p\sqrt{\frac{1}{n}+\frac{1}{m}}$	$\left(-\infty, \quad \overline{X}-\overline{Y}+t_{\alpha,N} \cdot S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right)$

Nota: gli intervalli unilaterali destri per $\mu_1-\mu_2$ si possono ricavare da quelli sinistri per $\mu_2-\mu_1$.

7 Qualità ed efficienza degli stimatori

Sia $X := (X_1, X_2, \dots, X_n)$ un campionare di una distribuzione *nota* tranne per il parametro θ che è incognito e $d(X)$ uno stimatore di θ
Come possiamo valutare la sua efficacia? un criterio può essere quello dell'*errore*

quadratico medio ossia:

$$r(d,\theta) := \mathbb{E}\left[(d(X)-\theta)^2\right]$$

e sarà questo il nostro indicatore del valore di d come stimatore di θ

7.0.1 Bias e Polarizzazione

Definizione: Sia $d = d(X)$ uno stimatore del parametro θ allora:

$$b_\theta(d) := \mathbb{E}\left[d(X)\right] - \theta$$

Questo viene detto *bias* di d come stimatore di θ

Se il bias è nullo(quindi $\mathbb{E}\left[d(X)\right] = \theta$), si dice che è uno stimatore *corretto* o *non distorto*

Esempio Sia $X_1, \dots, X_n$ un campione con media *incognita* θ quindi:

$$d_1(X_1, X_2, \dots, X_n) = X_1$$

$$d_2(X_1, X_2, ., X_n) = \frac{X_1 + X_2 + \dots + X_n}{n}$$

sono entrambi *stimatori non distorti di* θ

perchè entrambi:

$$\mathbb{E}\left[X_1\right] = \mathbb{E}\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \theta$$

Definizione Se $d = d(X)$ è uno *stimatore corretto*, il suo errore quadratico medio diventa:

$$\begin{aligned} r(d,\theta) &= \mathbb{E}\left[(d-\theta)^2\right] \\ &= \mathbb{E}\left[(d-\mathbb{E}\left[d\right])^2\right] \\ &= Var(d) \end{aligned}$$

Quindi l'errore quadratico medio di uno stimatore corretto è pari alla sua varianza

7.0.2 Combinazioni di stimatori corretti

Consideriamo due stimatori *corretti* e *indipendenti* di parametro θ (denotati con d_1 e d_2) con varianze rispettivamente σ_1^2 σ_2^2

$$\mathbb{E}\left[d_i\right] = \theta \quad Var(d_i) = \sigma_i^2 \qquad i = 1, 2$$

uno stimatore corretto di θ è il seguente

$$d := \lambda d_1 + (1 - \lambda) d_2$$

Successivamente vogliamo trovare anche il valore di λ che produce lo stimatore d con il *minore errore quadratico medio*

$$\begin{aligned} r(d, \theta) &= Var(d) \\ &= \lambda^2 Var(d_1) + (1 - \lambda)^2 Var(d_2) \qquad \text{per l'indipendenza di } d_1 \text{ e } d_2 \\ &= \lambda^2 \sigma_1^2 + (1 - \lambda)^2 \sigma_2^2 \end{aligned}$$

ayo bro what's this shit, le me calculate the derivata with latti:

$$\frac{d}{d\lambda} r(d, \theta) = 2\lambda\sigma_1^2 - 2(1 - \lambda)\sigma_2^2$$

e belin lo studiamo sto segno o no? denotiamo con $\hat{\lambda}$ il valore di θ che produce il minimo

$$2\hat{\lambda}\sigma_1^2 - 2(1 - \hat{\lambda})\sigma_2^2 = 0$$

da cui otteniamo:

$$\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}$$

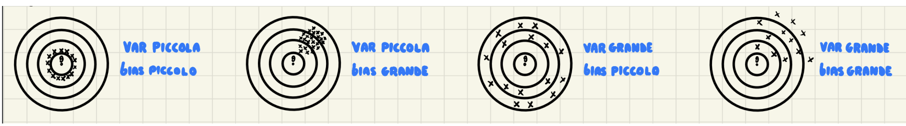
il peso ottimale da dare a uno stimatore deve essere **inversamente** proporzionale alla sua varianza
La migliore combinazione lineare delle d_i per l'errore quadratico medio è:

$$\begin{aligned} r(d,\theta) &= Var(d) \\ &= \left(\frac{1}{\sum_{i=1}^n} 1/\sigma_i^2\right)^2 \sum_{i=1}^n \left(\frac{1}{\sigma_i^2}\sigma_i^2\right) \\ &= \frac{1}{\sum_{i=1}^n 1/\sigma_i^2} \end{aligned}$$

$$r(d,\theta) = Var(d)$$

Bias/Polarizzazione Se $d(X)$ è **distorto**:

$$\begin{aligned} r(d,\theta) &= \mathbb{E}\left[(d-\theta)^2\right] \\ &= Var(d) + 0 + \mathbb{E}\left[b_\theta(d)^2\right] \\ &= Var(d) + b_\theta(d)^2 \end{aligned}$$



7.1 Stimatore della media di una distribuzione uniforme

Siano $X_1, X_2, \ldots, X_n$ un campione estratto da una popolazione con distribuzione *uniforme* su $(0,\theta)$ dove θ è un parametro incognito.

Dato che (non si sa come) $\mathbb{E}\left[\boldsymbol{X_i}\right] = \boldsymbol{\theta/2}$ è uno stimatore naturale per θ è dato da

$$d_1 = d_1(X) := 2\overline{X} := \frac{2}{n}\sum_{i=1}^{n} X_i$$

Siccome $\mathbb{E}\left[\boldsymbol{d_1}\right] = \boldsymbol{\theta}$, otteniamo che:

$$\begin{aligned}
\boldsymbol{r(d_1, \theta)} &= Var(d_1) \\
&= \frac{4}{n} Var(X_i) \\
&= \frac{4}{n} \cdot \frac{\theta^2}{12} \\
&= \frac{\theta^2}{3n}
\end{aligned}$$

Un secondo stimatore che abbiamo è quello di **massima verosomiglianza** (d_2):

$$d_2 = d_2(X) = MLE := \max(X_i)$$

Per trovare l'errore quadratico medio di d_2 dobbiamo prima conoscere la sua *media* e la sua *varianza*

$$\mathbb{E}\left[d_2\right] = \int_0^\theta x \frac{nx^{n-1}}{\theta^n}\, dx = \frac{n}{n+1}\theta$$

$$\mathbb{E}\left[d_2^2\right] = \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n}\, dx = \frac{n}{n+2}\theta^2$$

$$Var(d_2) = \mathbb{E}\left[d_2^2\right] - \mathbb{E}\left[d_2\right]^2 = \frac{n\theta^2}{(n+2)(n+1)^2}$$

Quindi ora calcoliamo la $\boldsymbol{r(d_2, \theta)}$:

$$\begin{aligned}
\boldsymbol{r\left(d_2, \theta\right)} &= \operatorname{Var}\left(d_2\right) + \left(E\left[d_2\right] - \theta\right)^2 \\
&= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{\theta^2}{(n+1)^2}\left[\frac{n}{n+2} + 1\right] \\
&= \frac{2\theta^2}{(n+1)(n+2)}
\end{aligned} \tag{3}$$

Confrontando gli errori quadratici medi notiamo che d_2 è **migliore** di d_1 per θ

$$\frac{2\theta^2}{(n+1)(n+2)} \leq \frac{\theta^2}{3n} \qquad d_2 \text{ migliore}$$

- SI DEPOLARIZZA $d_2 \longrightarrow d_3 = \frac{m+1}{m} d_2$ CORRETTO $\longrightarrow r(d_3, \theta) = VAR(d_3) = \frac{(m+1)^2}{m^2} VAR(d_2) = \frac{\theta^2}{m^2+2m}$ $\boxed{\frac{\theta^2}{m^2+2m} < \frac{2\theta^2}{m^2+3m+2} \Rightarrow d_3 \text{ MIGLIORE}}$
- ESISTE UNO MIGLIORE DI $d_3 \longrightarrow d_c = \overset{\text{COSTANTE}}{c} \cdot d_2 \longrightarrow r(d_c, \theta) = VAR(d_c) + (\mathbb{E}[d] - \theta)^2 = c^2 VAR(d_2) + (c\,\mathbb{E}[d_2] - \theta)^2 = \frac{2cm\theta^2}{(m+2)(m+1)^2} + \theta^2\left(c\frac{m}{m+1} - 1\right)^2$

(MINOR ERRORE: $\frac{dr(d_c,\theta)}{dc} = \frac{2m\theta^2}{(m+1)^2}\left[\frac{c}{m+2} + cm - (m+1)\right] \longrightarrow \frac{dr}{dc} = 0 \longrightarrow \frac{c}{m+2} + cm - (m+1) = 0 \to c = \frac{(m+1)(m+2)}{m^2+2m+1} = \frac{m+2}{m+1}$) $r(\overset{\text{POLARIZZATO}}{d_c}, \theta) = \frac{\theta^2}{(m+1)^2} = \frac{\theta^2}{m+2m+1}$ $\boxed{\frac{\theta^2}{m+2m+1} < \frac{\theta^2}{m^2+2m} \Rightarrow d_c \text{ MIGLIORE}}$

8 Stimatori Bayesiani

Definizione: Quando il parametro incognito $\boldsymbol{\theta}$ possiamo considerarlo come una variabile aleatoria, questo approccio viene detto *bayesiano*.
Se abbiamo delle informazioni su quelli che possono essere assunti i valori da θ ed esse assumono la forma di distribuzione di probabilità si dice che abbiamo **una**

distribuzione a priori per θ

Se i valori che osserviamo sono $X_i = x_i \qquad i = 1, 2, \ldots, n$ la *densità di probabilità condizionale di* θ è data da:

$$\begin{aligned} \boldsymbol{f(\theta|x_1, x_2, \ldots, x_n)} &= \frac{f(x_1, x_2, \ldots, x_n, \theta)}{\boldsymbol{f(x_1, x_2, \ldots, x_n)}} \\ &= \frac{\boldsymbol{f(x_1, x_2, \ldots, x_n|\theta)p(\theta)}}{\int f(x_1, x_2, \ldots, x_n|\theta')p(\theta')d\theta'} \end{aligned}$$

Dove:

- $\boldsymbol{f(\theta|x_1, x_2, \ldots, x_n)}$ Viene detta *probabilità a posteriori*
- $\boldsymbol{f(x_1, x_2, \ldots, x_n)}$ è la *MLE Marginale*
- $\boldsymbol{f(x_1, x_2, \ldots, x_n|\theta)}$ è la *MLE*
- $\boldsymbol{p(\theta)}$ è la *distribuzione a priori*

Una buona stima per θ può essere la **media** perciò:

$$\mathbb{E}\left[\theta|X_1 = x_1, \ldots, X_n = x_n\right] = \int_{-\infty}^{\infty} \theta f(\theta|x_1, x_2, \ldots, x_n)\, d\theta \quad \textit{nel caso continuo}$$

8.1 Stimatore di θ per Bernoulli

Se abbiamo $X_1, X_2, \ldots, X_n$ Bernoulliane, con massa di probabilità:

$$f(x|\theta) = \theta^x(1-\theta)^{1-x} \qquad x = 0, 1$$

Dove θ è un parametro sconosciuto

Supponiamo quindi che la *distribuzioni a priori* di θ sia uniforme su (0,1), denotiamo con p la densità a propri di θ

$$p(\theta) = 1 \qquad 0 < \theta < 1$$

La densità condizionale di θ date $x_1, x_2, \ldots, x_n$ è

$$\begin{aligned} f(\theta \mid x_1, x_2, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n, \theta)}{f(x_1, x_2, \ldots, x_n)} \\ &= \frac{f(x_1, x_2, \ldots, x_n \mid \theta)\, p(\theta)}{\int_0^1 f(x_1, x_2, \ldots, x_n \mid \vartheta)\, p(\vartheta) d\vartheta} \\ &= \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}}{\int_0^1 \vartheta^{\sum_i x_i}(1-\vartheta)^{n-\sum_i x_i} d\vartheta} \end{aligned} \tag{4}$$

Non è difficile provare (e invece lo è) integrando per parti un certo numero di volte che per ogni valore di m e r:

$$\int_0^1 \theta^m(1-\theta)^r \, d\theta = \frac{m!r!}{(m+r+1)!}$$

poniamo ora $\boldsymbol{x := \sum_{i=1}^n x_i}$

$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} \qquad 0 < \theta < 1$$

Ora siamo in grado di calcolare *la stima bayesiana*

$$\begin{aligned} E\left[\theta \mid x_1, x_2, \ldots, x_n\right] &= \frac{(n+1)!}{x!(n-x)!} \int_0^1 \theta^{1+x}(1-\theta)^{n-x} d\theta \\ &= \frac{(n+1)!}{x!(n-x)!} \frac{(1+x)!(n-x)!}{(n+2)!} \\ &= \frac{x+1}{n+2} \\ &= \boldsymbol{\frac{1+\sum_{i=1}^{n} X_i}{n+2}} \end{aligned} \tag{5}$$

Esempio Se raccogliamo un campione di *10 bernoulliane* e trovassimo **6 successi**, lo stimatore bayesiano di θ fornirebbe un valore di **7 / 12**
Lo stimatore di massima verosomiglianza vale invece **6 / 10**

8.2 Stimatore di θ per una Normale

Supponiamo che $X_1, X_2, \ldots, X_n$ sia una distribuzione normale con media θ *incognita* e varianza σ_0^2 **nota**

Calcoliamo la *densità condizionale* di θ:

$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n|\theta)p(\theta)}{f(x_1, x_2, \ldots, x_n)}$$

Ora calcoliamo la *media*:

$$\mathbb{E}\left[\theta|X_1, X_2, \ldots, X_n\right] = \boldsymbol{\frac{n/\sigma_0^2}{n/\sigma_0^2 + 1/\sigma^2}\overline{X} + \frac{1/\sigma^2}{n/\sigma_0^2 + 1/\sigma^2}\mu}$$

e successivamente la *varianza*:

$$Var(\theta|X_1, X_2, \ldots, X_n) = \frac{\mathbf{1}}{\boldsymbol{n/\sigma_0^2 + 1/\sigma^2}}$$

8.3 Stimatore di θ per Uniformi

Avendo una funzione di likelihood $f(x_1, x_2, \ldots, x_n|\theta)$ e sapendo che la distribuzione è *uniforme*
$\theta \in [a, b]$
$d_b = d_{MLE}$

9 Verifica delle ipotesi

Un ipotesi statistica è un'affermazione su uno o più parametri della distribuzione, si chiama ipotesi perchè non sappiamo a priori se sia vera oppure no.

9.1 Livelli di significatività

Consideriamo una popolazione con distribuzione F_θ che dipende da θ incognito e vogliamo verificare una qualche ipotesi su θ.

1. $H_0 : \theta = 1 \rightarrow$ *Ipotesi nulla* ***semplice***

2. $H_0 : \theta \leq n \rightarrow$ *Ipotesi nulla* ***composta***

Quando la prima ipotesi è vera, caratterizza l'intera distribuzione, mentre questo non è vero per la seconda ipotesi.

Esiste una regione critica **C** per cui se il campione aleatorio vi appartiene l'ipotesi *non viene accettata*. Esiste un livello di tolleranza specificato all'interno della regione critica per cui un'ipotesi può essere *ancora accettata*. Questa tolleranza è definita dal **livello di significatività**, ovvero viene definito α tale che se l'ipotesi è vera la probabilità di rifiutarla non superi α

$$\text{accetta } H_0 \text{ se}(X_1, X_2, \ldots, X_n) \notin C$$

e

$$\text{rifiuta } H_0 \text{ se}(X_1, X_2, \ldots, X_n) \in C$$

Esistono **due tipi di errori**:

1. **Prima Specie**: Si rifiuta H_0 anche se è vera

2. **Seconda Specie**: si accetta H_0 anche se è falsa

9.2 Verifica di ipotesi sulla media di una popolazione normale

Supponiamo di avere $X_1, X_2, \ldots, X_n$ sia un campione aleatorio di una popolazione normale di parametri μ σ^2 con *varianza* nota e *media* incognita, vogliamo verificare le seguenti ipotesi:

$$H_0 : \mu = \mu_0$$

e l'ipotesi *alternativa*

$$H_1 : \mu \neq \mu_0$$

Lo **stimatore puntuale** per μ è:

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

La **regione critica del test** invece è:

$$C := \{(X_1, X_2, \ldots, X_n) : |\overline{X} - \mu_0| > c\}$$

Dove c rappresenta la *tolleranza*

Quando $\mu = \mu_0$ sappiamo che $\overline{X}$ ha distribuzione **normale** con media μ_0 e varianza σ^2/n allora:

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim Z$$

Dove la relazione $\sim$ è **condizionata** all'ipotesi $H_0 : \mu = \mu_0$

c deve soddisfare la seguente relazione:

$$\alpha = P(\text{errore di I specie}) = \boldsymbol{P_{\mu_0}(|\overline{X} - \mu_0) > c}$$

Possiamo scrivere l'equazione di sopra in questo modo:

$$\alpha = 2P\left(Z > \frac{c\sqrt{n}}{\sigma}\right)$$

Per $P(Z > c\sqrt{n}/\sigma)$ per la definzione $z_{\frac{\alpha}{2}}$ vale:

$$P\left(Z > z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \longrightarrow c = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Il test con livello di significatività ha **due esiti**:

$$\text{si rifiuta } H_0 \text{ se} \left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\frac{\alpha}{2}}$$

$$\text{si accetta } H_0 \text{ se}\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| \leq z_{\frac{\alpha}{2}}$$

p dei dati $= \sqrt{n}\frac{\overline{X}-\mu_0}{\sigma}$

H_0 si **accetta** se $2P(Z > z_{\frac{\alpha}{2}})$ è elevata
H_0 si **rifiuta** se $2P(Z > z_{\frac{\alpha}{2}})$ è bassa
Perche se la probabilità che Z sia $> z_{\frac{\alpha}{2}}$ è *alta* allora il mio valore sarà vicino al mezzo e va bene. Se è basso allora è *lontano* dal mezzo e non va bene.

Esempio supponiamo una media campionaria dei *5* segnali ricevuti fosse *8.5*:

$$\frac{\sqrt{n}}{\sigma}|\overline{X}-\mu_0| = \frac{\sqrt{5}}{2}\cdot 0.5 \approx 0.559$$

Dato che:

$$P(|Z| > 0.559) = 2P(Z > 0.559) \approx 2X0.288 = 0.576$$

Otteniamo che il *p-dei-dati* è 0.576 e quindi l'ipotesi nulla che il segnale inviato fosse **8**, che viene accettata per ogni $\alpha < 0.576$
Se avessimo ottenuto che $\overline{X} = 11.5$ il valore del *p-dei-dati* sarebbe:

$$P\left(|Z| > \frac{\sqrt{5}}{2}\cdot 3.5\right) \approx 2P(Z > 3.913) \approx 0.00005$$

Con un valore così piccolo, l'ipotesi che il messaggio fosse stato 8, **va rifiutata**.

Riprendendo il discorso degli errori di specie andiamo a vedere ora *gli errori di* ***seconda specie***.

Rinfreschiamo la memoria, l'errore di seconda specie è quando *si accetta* H_0 *anche se è falsa*, quindi:

$$\boldsymbol{\beta(\mu)} := P_\mu(\text{accettare } H_0) = \boldsymbol{P_\mu}\left(-\boldsymbol{z_{\frac{\alpha}{2}}} \leq \frac{\boldsymbol{\overline{X} - \mu_0}}{\boldsymbol{\sigma/\sqrt{n}}} \leq \boldsymbol{z_{\frac{\alpha}{2}}}\right)$$

La funzione $\beta(\mu)$ è detta **curva OC** (*curva operativa caratteristica*) e rappresenta la probabilità di accettare H_0 quando la media reale è μ.

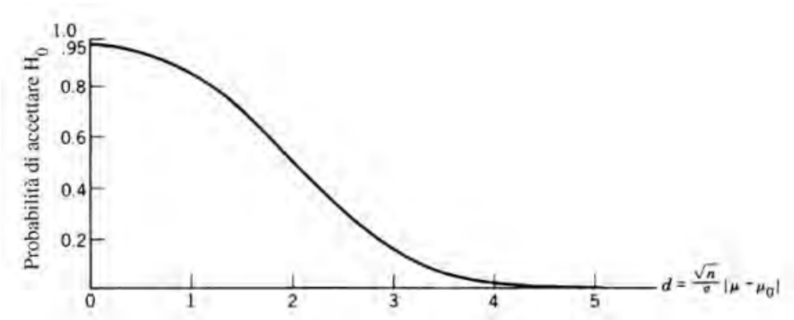


Figure 1: Curva OC di un test *bilaterale* per la media di una popolazione normale, con $\alpha = 0.05$

Per calcolare la probabilità ricordiamoci il fatto che $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$:

$$Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Quindi:

$$\begin{aligned}\beta(\mu) &= P_\mu\left(-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}\right)\\ &= \boldsymbol{\Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}}+z_{\frac{\alpha}{2}}\right)-\Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}}-z_{\frac{\alpha}{2}}\right)=1-\alpha}\end{aligned}$$

Dove Φ indica la *funzione di ripartizione* della distribuzione normale standard

Esempio quanto vale la probabilità di accettare $\mu = 8$ quando in realtà $\mu = 10$:

$$\frac{\sqrt{n}}{\sigma}(\mu_0-\mu) = \frac{\sqrt{5}}{2}(-2) = -\sqrt{5}$$

Dato che $\boldsymbol{z_{0.025} \approx 1.96}$ ricaviamo la probabilità cercata:

$$\begin{aligned}\beta(10) &\approx \Phi(-\sqrt{5}+1.96)-\Phi(-\sqrt{5}-1.96)\\ &= 1-\Phi(0.276)-1+\Phi(4.196)\\ &\approx -0.609+1 = \mathbf{0.391}\end{aligned}$$

Riprendendo il discorso della curva OC, ci permette di dimensionare il campione in modo che *l'errore di seconda specie* soddisfi le condizioni specifiche.

Come facciamo a trovare n tale che la probabilità di *accettare* $H_0 : \mu = \mu_0$ quando il vero valore è μ_1 sia un valore fissato β per n tale che $\boldsymbol{\beta(\mu_1) \approx \beta}$

$$n \approx \left[\frac{(z_{\frac{\alpha}{2}}+z_\beta)\sigma}{\mu_1-\mu_0}\right]^2$$

Notiamo che anche nel caso in cui $\boldsymbol{\mu_1 < \mu_0}$ troviamo sempre la stessa formula

Esempio Quante volte è necessario inviare il segnale con verifica dell'ipotesi $H_0 : \mu = 8$ con livello di significatività 0.05 con almeno il 75% di probabilità di rifiutare l'ipotesi nulla quando $\boldsymbol{\mu = 9.2}$
Dato che $z_{0.025 \approx 1.96}$ e $z_{0.25} \approx 0.67$

$$n \approx \left(\frac{1.96 + 0.67}{1.2}\right)^2 \cdot 4 \approx 19.21$$

Come vediamo per il risultato è necessario un *campione di 20 segnali*, quindi con $n = 20$

$$\begin{aligned} \beta(9.2) &\approx \Phi\left(-\frac{1.2\sqrt{20}}{2} + 1.96\right) - \Phi\left(-\frac{1.2\sqrt{20}}{2} - 1.96\right) \\ &\approx \Phi(-0.723) - \Phi(-4.643) \\ &\approx 1 - \Phi(0.723) \approx \mathbf{0.235} \end{aligned}$$

Quindi ricapitolando, se il segnale viene trasmesso 20 volte c'è il 76.5% di probabilità che l'ipotesi nulla $\mu = 8$ sia **rifiutata** se la media reale è **9.2**

9.3 Test unilaterali

Introduzione bla bla bla
Verifichiamo due ipotesi:

$$H_0 : \mu = \mu_0 \quad \text{contro} \quad H_1 : \mu > \mu_0$$

Dovremmo rifiutare l'ipotesi nulla quando lo stimatore di μ è molto più grande di μ_0, la regione critica è quindi:

$$C := \{(X_1, X_2, \ldots, X_n) : \overline{X} - \mu_0 > c\}$$

la probabilità di rifiuto dovrebbe essere α quando H_0 è vera, occorre però che c soddisfi la relazione:

$$P_{\mu_0}(\overline{X} - \mu_0 > c) = \alpha$$

Il test *con livello di significatività* α dovrà rifiutare $\boldsymbol{H_0}$ se $\boldsymbol{\overline{X} - \mu_0 > z_\alpha \cdot \sigma/\sqrt{n}}$

$$\text{si rifiuta } H_0 \text{ se} \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

$$\text{si accetta } H_0 \text{ se} \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha$$

Quella trovata è detta *regione criticia* ***unilaterale*** o a una coda, quindi il problema di verificare le ipotesi alternative

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Si dice problema di **test unilaterale**

poniamo $Z := \sqrt{n}(\overline{X} - \mu)/\sigma$ questa statistica è una *normale standard* quindi:

$$\begin{aligned}\beta(\mu) &:= P_\mu(\text{accettare } H_0) \\ &= P_\mu\left(Z \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \\ &= \boldsymbol{\Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) = 1 - \alpha}\end{aligned}$$

Dato che Φ in quanto *funzione di ripartizione* è ***crescente*** però $\beta(\mu)$ è una funzione **decrescente**
L'ipotesi *unilaterale*

$$H_0 : \mu \leq \mu_0$$

contro l'alternativa

$$H_1 : \mu > \mu_0$$

Per accertarci che il *livello di significatività* sia rimasto α
Al variare di μ la probabilità di rifiuto è data da $\mathbf{1 - \boldsymbol{\beta}(\boldsymbol{\mu})}$
Dobbiamo verificare che per ogni μ compatibile con H_0 per ogni $\boldsymbol{\mu \leq \mu_0}$

$$1 - \beta(\mu) \leq \alpha, \qquad \text{per ogni } \mu \leq \mu_0$$

Quindi:

$$\beta(\mu) \geq 1 - \alpha, \qquad \text{per ogni } \mu \leq \mu_0$$

Osservazione è possibile verificare l'ipotesi

$$H_0 : \mu = \mu_0$$

contro l'ipotesi *alternativa*

$$H_1 : \mu < \mu_0$$

ad un livello di significatività $\boldsymbol{\alpha}$, decidendo che:

$$\text{si rifiuta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$$

$$\text{si accetta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq -z_\alpha$$

9.4 Il test t

Fino ad ora abbiamo supposto che l'unico parametro incognito fosse la *media*, in questo caso la nostra varianza $\boldsymbol{\sigma^2}$ **non è nota**

In questa situazione consideriamo che si possa verificare l'ipotesi nulla che μ sia uguale ad un valore assegnato μ_0 contro l'ipotesi alternativa $\mu \neq \mu_0$

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

(Cambiare poi con desc più corta) Come in precedenza, sembra ragionevole rifiutare l'ipotesi nulla quando $\overline{X}$ cade lontano da μ_0 tuttavia la distanza a cui deve essere da μ_0 per giustificare questo rifiuto, dipende dalla deviazione standard σ che in quella sede era nota; in particolare $|\overline{X} - \mu 0|$ doveva essere maggiore di $z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}$ o equivalentemente

$$\left[\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right] > z_{\frac{\alpha}{2}}$$

Qui $\boldsymbol{\sigma}$ non è più conosciuta, sostituiamola quindi con la *deviazione standard campionaria* S

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

rifiutando l'ipotesi nulla quando

$$|\frac{\overline{X} - \mu_0}{S/\sqrt{n}}|$$

Quindi alla fine noi dobbiamo ottenere una distribuzione t

$$t_{n-1} \sim \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Se si denota con T la statistica di questo test, ovvero

$$T := \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

allora quando H_0 è vera (visto che $\mu = \mu_0$) ha distribuzione t **con** $n-1$ **gradi di libertà.**

$$P_{\mu_0}\left(-c \leq \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \leq c\right) = \mathbf{1-\boldsymbol{\alpha}}$$

Se vogliamo ricavare c:

$$\begin{aligned}\alpha &= 1 - P(-c \leq T < c) \\ &= P(T \leq -c) + P(T \geq c) \\ &= 2P(T \geq c)\end{aligned}$$

Per cui $P(T > c) = \frac{\alpha}{2}$, e quindi deve valere $c = t_{\frac{\alpha}{2},n-1}$, quindi in fin dei conti:

$$\text{si rifiuta } H_0 \text{ se } \left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| > t_{\frac{\alpha}{2},n-1}$$

$$\text{si accetta } H_0 \text{ se } \left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| \leq t_{\frac{\alpha}{2},n-1}$$

Vedere tabella sotto per tutt'e cose

Figure 2: $X_1, X_2, \ldots, X_n$ è un campionare estratto da una popolazione $\mathcal{N}(\mu, \sigma^2)$

σ^2 nota $\qquad\qquad \overline{X} := \frac{1}{n}\sum_{1=1}^{n} X_i$

H_0	H_1	Statistica del test, X_{ts}	Si rifiuta H_0 con livello di significatività α se...	p-dei-dati se $X_{ts} = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\ldots \lvert X_{ts} \rvert > z_{\frac{\alpha}{2}}$	$2P(Z > \lvert t \rvert)$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\ldots X_{ts} > z_\alpha$	$P(Z > t)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\ldots X_{ts} < -z_\alpha$	$P(Z < t)$

Esempio Vogliamo verificare l'ipotesi che il consumo *medio* di acqua sia *350* galloni al giorno.
Si misurano i consumi medi di un campione di *20 campioni* che seguono:

340	356	332	362	318	344	386	402	322	360
362	354	340	372	338	375	364	355	324	370

Dobbiamo verificare le due ipotesi seguenti:

$$H_0 : \mu = 350 \quad \text{contro} \quad H_1 : \mu \neq 350$$

Calcoliamo ora la **media** e la **deviazione standard campionaria**

$$\overline{X} = 353.8 \qquad S \approx 21.85$$

troviamo ora il valore della statistica del test:

$$T \approx \frac{\sqrt{20} \cdot 3.8}{21.85} \approx 0.778$$

il valore che abbiamo trovato è minore di $t_{0.05,19} \approx 1.729$ l'ipotesi nulla è *accettata* ad un livello del *5%* TODO FINIRE PAGINA PDF 324 / 342 TOTALE

Figure 3: $X_1, X_2, \ldots, X_n$ è un campionare estratto da una popolazione $\mathcal{N}(\mu, \sigma^2)$ e σ^2 non è nota

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

H_0	H_1	**Statistica del test, X_{ts}**	**Si rifiuta H_0 con livello di significatività α se...**	**p-dei-dati se $X_{ts} = t$**
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\ldots \lvert X_{ts} \rvert > t_{\frac{\alpha}{2},n-1}$	$2P(T_{n-1} > \lvert t \rvert)$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\ldots X_{ts} > t_{\alpha,n-1}$	$P(T_{n-1} > t)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\ldots X_{ts} < -t_{\alpha,n-1}$	$P(T_{n-1} < t)$

Nota: T_{n-1} ha distribuzione t con $n-1$ gradi di libertà. Inoltre $P(T_{n-1} > t_{\alpha,n-1}) = \alpha$.

9.5 Verifica se due popolazioni hanno la stessa media

Una situazione che accade spesso è decidere se *vari approcci* portano allo stesso risultato, oppure no.
Questa problematica si ricordune spesso alla verifica dell'ipotesi che due popolazioni normali *abbiano la stessa media.*

9.5.1 Il caso in cui le varianze sono note

Supponiamo di avere $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ sono due campioni di due popolazioni *normali* di medie μ_x μ_y e varianze *note* σ_x^2 e σ_y^2
Come sempre verifichiamo le due ipotesi

$$H_0 : \mu_x = \mu_y$$

$$H_1 : \mu_x \neq \mu_y$$

Dato che $\overline{X}$ e $\overline{Y}$ sono rispettivamente *stimatori* di μ_x e μ_y
Possiamo dire che $\overline{X} - \overline{Y}$ può essere **usato come stimatore** di $\mu_x - \mu_y$

$$\text{si rifiuta } H_0 \text{ se } \left|\overline{X} - \overline{Y}\right| > c$$

$$\text{si accetta } H_0 \text{ se } \left|\overline{X} - \overline{Y}\right| \leq c$$

Come facciamo sempre noi possiamo trovare il valore di c che rende questo test di livello di significatività α in questo modo:

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \sim \mathcal{N}(0,1)$$

Per verificare l'ipotesi nulla $H_0 : \mu_x = \mu_y$ contro $H_1 : \mu_x \neq \mu_y$ facciamo cosi:

$$\text{si rifiuta } H_0 \text{ se } \frac{|\overline{X} - \overline{Y}|}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} > z_{\frac{\alpha}{2}}$$

$$\text{si accetta } H_0 \text{ se } \frac{|\overline{X}-\overline{Y}|}{\sqrt{\sigma_x^2/n+\sigma_y^2/m}} \leq z_{\frac{\alpha}{2}}$$

9.5.2 Il caso in cui le varianze non sono note ma supponiamo siano uguali

Prendiamo in considerazione i campioni di prima, tutti i nostri parametri sono *incogniti* e studiamo le due ipotesi

$$H_0 : \mu_x = \mu_y \quad \text{contro} \quad H_1 : \mu_x \neq \mu_y$$

Prima di far tutto possiamo supporre che le due varianze *incognite* siano uguali tra di loro quindi:

$$\sigma^2 := \sigma_x^2 = \sigma_y^2$$

Calcoliamo le due *varianze campionarie*

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

$$S_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(X_j-\overline{Y})^2$$

Equazione idk:

$$\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{S_p\sqrt{1/n+1/m}} \sim t_{n+m-2}$$

Dove $\boldsymbol{S_p^2}$ è lo *stimatore pooled* di σ^2 e viene definito in questo modo:

$$S_p^2 := \frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}$$

Quando H_0 è vera ($\mu_x - \mu_y = 0$):

$$T := \frac{\overline{X} - \overline{Y}}{S_p\sqrt{1/n + 1/m}}$$

ha distribuzione t con $n + m - 2$ gradi di libertà

Quindi possiamo verificare le ipotesi cosi:

si rifiuta H_0 se $|T| > t_{\frac{\alpha}{2};n+m-2}$

si accetta H_0 se $|T| \leq t_{\frac{\alpha}{2};n+m-2}$

possiamo eseguire il test determinando *il p-dei-dati*, denotando con υ il valore assunto da T

$$\begin{aligned} \boldsymbol{p - dei - dati} &= P(|T_{n+m-2}| \geq |\upsilon|) \\ &= 2P(T_{n+m-2} \geq |\upsilon|) \end{aligned}$$

Caso unilaterale Per l'ipotesi *unilaterale* abbiamo le due seguenti ipotesi:

$$\mu_0 : \mu_x \leq \mu_y \qquad \text{contro} \qquad H_1 : \mu_x > \mu_y$$

H_0 deve essere **rifiutata** per valore elevati di T, il test di significatività α è:

si rifiuta H_0 se $T > t_{\alpha,n+m-2}$

si accetta H_0 se $T \leq t_{\alpha,n+m-2}$

il *p-dei-dati* invece è il seguente (ricordando che υ è il valore assunto da T)

$$\boldsymbol{p - dei - dati} = P(T_{n+m-2} \geq \upsilon)$$

Esempio abbiamo $\overline{X} = 6.450$ e $\overline{Y} = 7.125$
Calcoliamo le due S, quindi $\boldsymbol{S_x^2 \approx 0.581}$ e $\boldsymbol{S_y^2 \approx 0.778}$
Calcoliamo ora lo stimatore $\boldsymbol{S_p^2}$:

$$S_p^2 = \frac{9}{20} S_x^2 + \frac{11}{20} S_y^2 \approx 0.689$$

e la statistica del test:

$$v = \frac{-0.675}{\sqrt{0.689(1/10 + 1/12)}} \approx -1.90$$

9.5.3 verifica di due popolazioni con stessa media

Si assume	Statistica del test, D_{ts}	Si rifiuta H_0 con livello di significatività α se...	p-dei-dati se $D_{ts} = t$
σ_x e σ_y note	$\frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$	$\ldots \lvert D_{ts} \rvert > z_{\frac{\alpha}{2}}$	$2P(Z > \lvert t \rvert)$
$\sigma_x = \sigma_y$ ignote	$\frac{\overline{X} - \overline{Y}}{S_p \sqrt{1/n + 1/m}}$	$\ldots \lvert D_{ts} \rvert > t_{\frac{\alpha}{2}, n+m-2}$	$2P(T_{n+m-2} > \lvert t \rvert)$
n e m grandi	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	$\ldots \lvert D_{ts} \rvert > z_{\frac{\alpha}{2}}$	$2P(Z > \lvert t \rvert)$

9.6 Il test t per campioni di coppie di dati

I dati che prendiamo in esempio sono descritti da n coppie di valori (X_i, Y_i) per $i = 1, 2, \ldots, n$

$X_1, X_2, \dots, X_n \qquad Y_1, Y_2, \dots, Y_m \qquad n$ **e** m **devono essere uguali**
Le nostre due variabili sono **dipendenti** quindi:

$$X \sim \mathcal{N}(\mu_x, \sigma_x)$$

$$Y \sim \mathcal{N}(\mu_y, \sigma_y)$$

Se poniamo $\boldsymbol{W_i := X_i - Y_i}$ per $i = 1, 2, \dots, n$ possiamo verificare queste due ipotesi

$$H_0 : \mu_W = 0 \quad \text{contro} \quad H_1 : \mu_W \neq 0$$

La nostre W provengono da un campione di popolazione $\mathcal{N}(\mu_W, \sigma_W^2)$, il test t quindi ci fornisce le seguenti regole:

$$\text{si accetta } H_0 \text{ se } -t_{\frac{\alpha}{2}, n-1} \leq \sqrt{n}\frac{\overline{W}}{S_W} \leq t_{\frac{\alpha}{2}, n-1}$$

si rifiuta H_0 negli altri casi

9.7 Verifica di ipotesi sulla varianza di una popolazione normale

Sia $X_1, X_2, \dots, X_n$ un campione di popolazione normale con media incognita μ e varianza incognita σ^2, verifichiamo le seguenti ipotesi:

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{contro l'alternativa} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

con un valore di $\boldsymbol{\sigma_0^2}$ prefissato

Otteniamo ora il test, abbiamo una distribuzione *chi-quadro* con $n - 1$ gradi di libertà, quindi quando H_0 è vera:

$$\frac{S^2}{\sigma_0^2}(n-1) \sim \mathcal{X}_{n-1}^2$$

e quindi otteniamo:

$$P_{H_0}\left(\mathcal{X}^2_{1-\frac{\alpha}{2},n-1} \leq \frac{S^2}{\sigma_0^2}(n-1) \leq \mathcal{X}^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

Queste sono infine le nostre regole da adottare

$$\text{si accetta } H_0 \text{ se } \mathcal{X}^2_{1-\frac{\alpha}{2},n-1} \leq \frac{S^2}{\sigma_0^2}(n-1) \leq \mathcal{X}^2_{\frac{\alpha}{2},n-1}$$

$$\text{si rifiuta } H_0 \text{ negli altri casi}$$

Il *p-dei-dati* del test è il seguente:

$$\boldsymbol{p-dei-dati} = 2\min\{P(\mathcal{X}^2_{n-1} \leq c), \quad 1-P(\mathcal{X}^2_{n-1} \leq c)\}$$

9.8 Verifica di due popolazione normali che hanno la stessa varianza

Abbiamo $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ sono due campioni *normali* **indipendenti**, con μ_x, σ_x^2 e μ_y, σ_y^2 incogniti, vediamo le verifiche dell'ipotesi:

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad \text{contro} \quad H_1 : \sigma_x^2 \neq \sigma_y^2$$

Le due *varianza campionarie* sono:

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

$$S_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

Abbiamo una distribuzione F con parametri $n-1$ e $m-1$ quando H_0 è vera:

$$\frac{S_x^2}{S_y^2} \sim F_{n-1,m-1}$$

e ne deduciamo che:

$$P_{H_0}\left(F_{1-\frac{\alpha}{2},n-1,m-1} \leq \frac{S_x^2}{S_y^2} \leq F_{\frac{\alpha}{2},n-1,m-1}\right) = 1-\alpha$$

Le nostre regole da adottare sono:

si accetta H_0 se $F_{1-\frac{\alpha}{2},n-1,m-1} \leq \dfrac{S_x^2}{S_y^2} \leq F_{\frac{\alpha}{2},n-1,m-1}$

si rifiuta H_0 negli altri casi

Il test del *p-dei-dati* è dato da:

$$\boldsymbol{p-dei-dati} = 2\min\{P(F_{n-1,m-1} \leq \upsilon), \quad 1-P(F_{n-1,m-1} \leq \upsilon)\}$$

Nota: il test **impone** di rifiutare H_0 ogni volta che il *livello di significatività* α è *maggiore o uguale* al p-dei-dati

Esempio Vengono eseguiti 10 esperimenti nel primo caso e 12 nel secondo, con le seguenti varianze campionarie $S_1^2 = 0.14$ e $S_2^2 = 0.28$, possiamo rifiutare ad un livello di significatività del 5% ?
Calcoliamo la funzione di ripartizione delle *distribuzioni F*, quindi:

$$P(F_{9,11} \leq 0.5) \approx 0.154$$

Quindi ora calcoliamo il *p-dei-dati*

$$p-dei-dati \approx 2\min(0.154, 0.846) = 0.308$$

L'ipotesi nulla **deve essere accettata.**

9.9 La verifica di ipotesi su una popolazione di Bernoulli

Il numero di difetti in un campione di n pezzi ha una distribuzione *binomiale* di parametri (n, p), le verifiche dell'ipotesi sono le seguenti:

$$H_0 : p \leq p_0 \quad \text{contro l'alternativa} \quad H_1 : p > p_0$$

p_0 è un *valore assegnato*

$$\mathbb{E}[X] = np$$

$$Var(X) = np(1-p)$$

$$\frac{X - np}{\sqrt{np_0(1-p_0)}} \sim \mathcal{N}(0,1)$$

———————

$$\text{si accetta } H_0 \text{ se } \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \leq z_\alpha$$

$$\text{si rifiuta } H_0 \text{ se } \frac{X - np_0}{\sqrt{np_0(1-p_0)}} > z_\alpha$$

10 AN.O.VA

Definizione: Analysis of variance, ci serve per confrontare più gruppi diversi per esempio per capire se hanno *medie uguali*

$$Z_i := \frac{X_i - \mu_i}{\sigma^2} \sim \mathcal{N}(0,1)$$

le seguenti variabili aleatorie sono *normali standard* e quindi:

Abbiamo m gruppi formati da n oggetti. ogni gruppo rappresenta una variabile aleatoria $X_i \sim \mathcal{N}(\mu, \sigma^2)$

$$\sum_{i=1}^{N} Z_i^2 = \sum_{i=1}^{N} \frac{(X_i - \mu_i)^2}{\sigma^2} \sim \mathcal{X}_N^2$$

Essa è una *chi-quadro* con N gradi di libertà, non stimiamo direttamente le μ_i ma usiamo il fatto che queste sono combinazione lineari di k *parametri incogniti*
In questa ipotesi possiamo dimostrare ciò:

$$\sum_{i=1}^{N} \frac{(X_i - \hat{\mu_i})^2}{\sigma^2} \sim \mathcal{X}_{N-k}^2$$

dove *N* sono gli oggetti totali mentre *k* sono i gruppi

Prendiamo μ come *unico parametro da stimare* cosi che $k = 1$ se sostituiamo μ con $\overline{X}$ che è il suo stimatore, troviamo questa espressione:

$$\sum_{i=1}^{N} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{N-1}{\sigma^2} \cdot \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})^2 = \boldsymbol{\frac{S^2}{\sigma^2}(N-1)}$$

10.1 Anova a 1 via

In questo caso noi abbiamo m campioni *indipendenti*, formati da n variabili aleatorie con media che **dipende** dal campione e varianza *fissata*
Denotiamo $X_{ij} \quad i = 1, \dots, m$ con quello che indica il campione mentre con $j = 1, \dots, n$ indichiamo la posizione all'interno del campione stesso

I parametri $\mu_1, \mu_2, \dots, \mu_m$ e σ sono incogniti, il nostro scopo è quello di verificare l'ipotesi nulla:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_m$$

e la sua controparte:

$$H_1 : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_m$$

Dato che ci sono nm *variabili aleatorie indipendenti* la *somma dei quadrati* è una distribuzione *chi-quadro* con nm gradi di libertà:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\mathbb{E}\left[X_{ij}\right])^2}{\sigma} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\mu_i)^2}{\sigma^2} \sim \mathcal{X}^2_{nm}$$

come stimatori degli m usiamo le medie campionarie dei singoli campioni di dati; in particolare $\boldsymbol{X_{i*}}$ denoterà quella del campione i-esimo:

$$X_{i*} := \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

Siccome X_{i*} è uno stimatore di μ_i lo **sostituiamo** nell'equazione di sopra, quindi:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-X_{i*})^2}{\sigma^2} = \frac{SS_W}{\sigma^2} \sim \mathcal{X}^2_{nm-m}$$

$$SS_W := \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-X_{i*})^2$$

Essa rappresenta una *chi-quadro* con $nm-m$ gradi di libertà

Calcoliamo ora la media di $\boldsymbol{SS_W}$ otteniamo che:

$$\mathbb{E}\left[\frac{SS_W}{\sigma^2}\right] = nm - m \quad \text{ovvero} \quad \mathbb{E}\left[\frac{SS_W}{nm-m}\right] = \sigma^2$$

Cosi abbiamo trovato il primo stimatore di σ^2

Fino ad ora abbiamo supposto che H_0 fosse vera o meno.

10.1.1 Stima di σ^2 valida solo quando $\mu_i = \mu$

In questi casi tutti gli stimatori $X_{1*}, X_{2*}, \ldots, X_{m*}$ sono normali di media μ e varianza σ^2/n, la loro somma dei quadrati è la seguente:

$$\sum_{i=1}^{m} \frac{(X_{i*} - \mathbb{E}[X_{i*}])^2}{Var(X_{i*})} = \sum_{i=1}^{m} \frac{(X_{i*} - \mu)^2}{\sigma^2/n} \sim \mathcal{X}_m^2$$

questa è una *chi-quadro* con m gradi di libertà

Abbiamo bisogno però di uno *stimatore* di μ, e la loro media campionaria risulta essere la scelta migliore, quindi:

$$X_{**} := \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} = \frac{1}{m} \sum_{i=1}^{m} X_{i*}$$

Nell'equazione sopra ora andiamo quindi a sostituire μ con X_{**} e otteniamo (quando H_0 è vera)

$$\sum_{i=1}^{m} \frac{(X_{i*} - X_{**})^2}{\sigma^2/n} = \frac{SS_b}{\sigma^2} \sim \mathcal{X}_{m-1}^2$$

Dove $\boldsymbol{SS_b}$ è:

$$SS_b := n \sum_{i=1}^{m} (X_{i*} - X_{**})^2$$

Quindi, riassumento, quando H_0 è vera:

$$\mathbb{E}\left[\frac{SS_b}{\sigma^2}\right] = m - 1 \quad \text{ovvero} \quad \mathbb{E}\left[\frac{SS_b}{m-1}\right] = \sigma^2$$

Di seguito la tabella che riassume tutta la merda che il libro spiega in 10 pagine:

Variazione	Somma di quadrati	Gradi di libertà
Tra i campioni	$SS_b := n\sum_i (X_{i*} - X_{**})^2$	$m-1$
Entro i campioni	$SS_W := \sum_i \sum_j (X_{ij} - X_{i*})^2$	$nm-m$

Ipotesi nulla	Statistica del test	Un test con significatività α deve	p-dei-dati se $D_{ts} = v$
Tutte le μ_i uguali	$D_{ts} := \dfrac{SS_b/(m-1)}{SS_W/(nm-m)}$	rifiutare H_0 se $D_{ts} > F_{\alpha,m-1,nm-m}$	$P(F_{m-1,nm-m} \geq v)$

11 Regressione lineare

Molti problemi di statistica prevedono una singola variabile Y di *risposta* e un certo numero di variabili $\boldsymbol{x_1, x_2, \ldots, x_r}$ di *ingresso*. La *risposta* è in funzione dei dati, Y è anche detta *variabile dipendente*, mentre le x_i sono le *variabili indipendenti*.
La più semplice relazione potrebbe essere quella lineare:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r$$

Dove $\beta_0, \beta_1, \ldots, \beta_r$ sono costanti.

Predire esattamente le β_i non è possibile, quindi all'equazione si aggiunge un *errore casuale* denominato e:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r + e$$

La variabile e ha distribuzione normale standard. $e \sim \mathcal{N}(0,1)$
L'equazione qui sopra è chiamata **equazione di regressione lineare**.
Questa esprime la regressione di Y rispetto alle variabili indipendenti $x_1, x_2, \ldots, x_r$, mentre le costanti $\beta_0, \beta_1, \ldots, \beta_r$ sono dette *coefficienti di regressione* e vanno

normalmente stimate. Un equazione di regressione si dice *semplice* se $r = 1$, e quindi c'è solo una variabile indipendente, negli altri casi si dice regressione *multipla*. Quindi la relazione diventa:

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

11.1 Stima di parametri di regressione

Indichiamo con A e B (variabili aleatorie) degli stimatori di α, β. L'equazione diventerà:

$$Y = A + Bx + e$$

Per avvicinarsi alla retta reale la quantità $\boldsymbol{(Y_i - A + Bx_i)^2}$ deve risultare minima. (rappresenta il quadrato della differenza tra predizione e valore osservato) Quindi:

$$SS := \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

Somma dei quadrati degli scarti tra risposte *stimate* e *reali*

11.1.1 Metodo dei minimi quadrati

Ricaviamo A e B tale per cui la SS risulta minima:

$$A = \overline{Y} - B\overline{x}$$
$$B = \frac{\sum_i x_i Y_i - \overline{x} \sum_i Y_i}{\sum_i x_i^2 - n\overline{x}^2}$$

La retta $Y = A + Bx + e$ è *la stima della retta di regressione*.

12 Distribuzione degli stimatori

$Y_1, Y_2, \ldots, Y_n$ sono indipendenti con distribuzione normale. $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$ B e A anch'esse hanno distribuzione normale. B è uno stimatore non distorto di β perché il suo valore atteso è uguale a β:

$$\mathbb{E}[B] = \beta$$

Quindi la sua varianza risulta essere:

$$\text{Var}(B) = \frac{\sigma^2}{\sum_i x_i^2 - n\overline{x}^2}$$

Anche A è uno stimatore non distorto di α perché il valore atteso è α:

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

Varianza di A:

$$\text{Var}(A) = \frac{\sigma^2 \sum_i x_i^2}{n(\sum_i x_i^2 - n\overline{x}^2)}$$

Somma dei quadrati dei residui è usata per stimare la varianza degli errori, σ^2:

$$SS_R := \sum_i^n (Y_i - A - Bx_i)^2$$

La SS_R ha distribuzione chi-quadro, con $n-2$ gradi di libertà:

$$\frac{SS_R}{\sigma^2} \sim \mathcal{X}^2_{n-2}$$

Il valore atteso della SS_R è uguale alla varianza, quindi è uno stimatore non distorto del parametro incognito σ^2:

$$\mathbb{E}\left[\frac{SS_R}{\sigma^2}\right] = n - 2 \quad \Rightarrow \quad \mathbb{E}\left[\frac{SS_R}{n-2}\right] = \sigma^2$$

$$S_{xY} := \sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y} \quad \text{dispersione di } x \text{ e } Y$$

$$S_{xx} := \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \quad \text{dispersione di } x \text{ e } x$$

$$S_{YY} := \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \quad \text{dispersione di } Y \text{ e } Y$$

Possiamo riscrivere B come:

$$B = \frac{\sum_i x_i Y_i - \overline{x} \sum_i Y_i}{\sum_i x_i^2 - n\overline{x}^2} \quad \Rightarrow \quad B = \frac{S_{xY}}{S_{xx}}$$

In generale: Nel caso in cui $Y_i, i = 1, 2, 3, \ldots, n$ siano normali indipendenti con media $\alpha + \beta x_i$ e varianza σ^2, gli stimatori dei minimi quadrati per β e α sono:

$$B = \frac{S_{xY}}{S_{xx}} \qquad A = \overline{Y} - B\overline{x}$$

e hanno distribuzione:

$$B \sim \mathcal{N}(\beta, \frac{\sigma^2}{S_{xx}}) \qquad A \sim \mathcal{N}(\alpha, \frac{\sigma^2 \sum_i x_i^2}{nS_{xx}})$$

La somma dei quadrati dei residui è calcolata tramite:

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

La SS_R ha distribuzione:

$$\frac{SS_R}{\sigma^2} \sim \mathcal{X}_{n-2}^2$$

13 Inferenza sui parametri della regressione

Quanto sono distanti A e B da α e β? Dobbiamo vedere l'intervallo di confidenza

13.1 Inferenza su β

Formula dell'intervallo di confidenza di β:

$$\beta \in B \pm \sqrt{\frac{SS_R}{(n-2)S_{xx}}} \sim t_{\frac{\alpha}{2},n-2}$$

Estesa:

$$P\Big(B - t_{\frac{\alpha}{2},n-2} \cdot \frac{\sqrt{SS_R}}{(n-2)S_{xx}} < \beta < B + t_{\frac{\alpha}{2},n-2} \cdot \frac{\sqrt{SS_R}}{(n-2)S_{xx}}\Big)$$

Importante: α **NON** è il parametro della regressione, ma è il livello di confidenza.

13.2 Inferenza su α

Formula dell'intervallo di confidenza di α:

$$\alpha \in A \pm \frac{SS_R \sum_i x_i^2}{\sqrt{n(n-2)S_{xx}}} \sim t_{\frac{\alpha}{2},n-2}$$

Dove la prima α è il coefficiente della retta mentre α nella t è il livello di confidenza.

13.3 Inferenza su $\alpha + \beta x_0$ (test su $\overline{Y}$)

Il valore atteso di $A + Bx_0$ è uguale a $\alpha + \beta x_0$ quindi è uno stimatore non distorto:

$$\mathbb{E}\left[A + Bx_0\right] = \mathbb{E}\left[A\right] + x_0\mathbb{E}\left[B\right] = \boldsymbol{\alpha + \beta x_0}$$

La varianza è:

$$\mathsf{Var}(A + Bx_0) = \sigma^2 \left[\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}} \right]$$

qual'è la distribuzione di $A + Bx_0$?

$$A + Bx_0 \sim \mathcal{N}\Big(\alpha + \beta x_0, \sigma^2 \Big[\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}} \Big]\Big)$$

13.3.1 Intervalli di confidenza

intervallo di confidenza di $\alpha + \beta x_0$:

$$\alpha + \beta x_0 \in A + Bx_0 \pm t_{\frac{\alpha}{n}, n-2} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \cdot \sqrt{\Big(\frac{SS_R}{n-2} \Big)}$$

S_{xx} risulta piccolo se i punti sono vicini alla media.

13.4 Inferenza di $Y_0 = Y(x_0) \to$ predittivo

Nel caso dovessimo prevedere un nuovo elemento della retta di regressione (utilizzando i dati già a disposizione) dobbiamo utilizzare la seguente formula:

$$A + Bx_0 \pm t_{\frac{\alpha}{2},n-2} \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\frac{SS_R}{n-2}}$$

Riassunto: Delle distribuzioni

Riassumiamo qui di seguito le distribuzioni ottenute nella sezione.

$$\textit{modello: } Y = \alpha + \beta x + e, \qquad e \sim \mathcal{N}\left(0, \sigma^2\right)$$
$$\textit{dati: } (x_i, Y_i), \qquad i = 1, 2, \ldots, n$$

Inferenze su	**Risultato da utilizzare**
β	$\sqrt{\frac{(n-2)S_{xx}}{SS_{\mathrm{R}}}}(B - \beta) \sim t_{n-2}$
α	$\sqrt{\frac{n(n-2)S_{xx}}{SS_{\mathrm{R}} \cdot \sum_i x_i^2}}(A - \alpha) \sim t_{n-2}$
$\alpha + \beta x_0$	$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_{\mathrm{R}}}{n-2}}} \sim t_{n-2}$
$Y(x_0)$	$\frac{Y - A - Bx_0}{\sqrt{\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_{\mathrm{R}}}{n-2}}} \sim t_{n-2}$

13.5 Coefficiente di determinazione

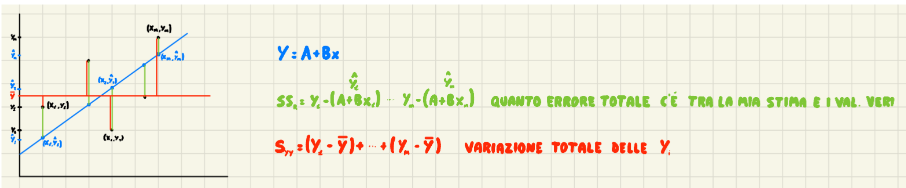
Come verifico i miei valori (della retta)? Tramite il *coefficiente di determinazione*. Formula del coefficiente di determinazione:

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}} \qquad 0 \leq R^2 \leq 1$$

Casi possibili:

1. Se $R^2 = 1$:
   (a) la dispersione è data solo dalla retta (regressione)
2. Se $R^2 = 0$:
   (a) la dispersione è dovuta solo dal rumore

La retta è migliore più R^2 è vicino a 1.



13.6 Coefficiente di correlazione

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}}$$

Dimostrazione matematica di R^2:

$$r^2 = \frac{S_{xY}^2}{S_{xx} S_{YY}} = \ldots = \mathbf{1} - \frac{\boldsymbol{SS_R}}{\boldsymbol{S_{YY}}}$$

Quindi:

$$|\boldsymbol{r}| = \sqrt{\boldsymbol{R^2}}$$

13.7 Analisi dei residui

Se il nostro modello non segue la forma di una "retta" non possiamo utilizzare la retta di regressione per rappresentare i nostri dati.

13.8 Trasformazione al lineare

Si può linearizzare tramite diverse funzioni, quella esponenziale in questo modo:

$$W(t) = ce^{-dt}$$

dove e, t sono parametri

Calcoliamo il log:

$$\log(W(t)) \approx \log(c) - dt$$

Se ora poniamo:

- $Y = \log W(t)$

- $\alpha = \log c$

- $\beta = -d$

La regressione lineare:

$$Y = \alpha + \beta t + e \quad \text{diventa} \quad W(t) \approx e^{A+Bt}$$

13.9 Rimedio al caso eteroschedastico

Nel modello eterschedastico la varianza è in funzione della x.
Ovvero l'errore cresce in base alle x.

Formula della *varianza degli errori*:

$$\text{Var}(e_i) = \frac{\sigma^2}{W_i}$$

La W_i è il peso nel caso eteroschedastico:

$$W_i = \frac{1}{x_i}$$

Formula della somma dei quadrati dei residui moltiplicato per il peso:

$$\sum_i W_i(Y - (A + Bx_0))^2$$

13.10 Regressione lineare multipla

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_k x_k + e$$

$$\min \sum_i (Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots + \beta_k x_{ik}))$$

13.11 Regressione (lineare) polinomiali

Nel caso in cui il nostro modello non può essere approssimato con un modelli lineari, si possono utilizzare relazioni polinomiali:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + e$$

Dobbiamo minimizzare:

$$\sum_i^n (Y_i - B_0 - B_1 x_1 - \ldots - B_r x_i^r)^2$$

Per determinare $\beta_0, \beta_1, \ldots, \beta_r$ dobbiamo

14 Stima di affidabilità dei sistemi

14.1 Introduzione

In questa sezione prendiamo in considerazione una popolazione di oggetti i cui tempi di vita sono *variabili aleatorie* con distribuzione comune.
L'obiettivo di questo capitolo è quello di usare tutti i dati che abbiamo per stimare **un parametro incognito**
Nella sezione *14.2* viene introdotto il concetto di *funzione di rischio (o intensità di rotture)*, mentre nella sezione *14.3* ci concentriamo sulla *legge esponenziale*

14.2 Funzione di intensità di rotture

Consideriamo una var. aleatoria X *continua e positiva*, e rappresenta il tempo di vita di un certo tipo di oggetti.
Se abbiamo come F la *funzione di ripartizione* e f la *densità di probabilità*
La sua **funzione di rischio / intensità di rotture** è la funzione λ definita da:

$$\lambda(t) := \frac{f(t)}{1 - F(t)}$$

Noi vogliamo studiare un elemento che è soggetto a *rotture*, che funziona *ininterrotamente* da un tempo t
Quindi noi vogliamo cercare una probabilità condizionata, ossia la seguente:

$$\begin{aligned} P(X \in (t, t+dt) | X > t) &:= \frac{P(X \in (t, t+dt), X > t)}{P(X > t)} \\ &= \frac{P(X \in (t, t+dt))}{1 - F(t)} \\ &\approx \frac{f(t)dt}{1 - F(t)} =: \boldsymbol{\lambda(t)dt} \end{aligned}$$

In questo caso quindi $\lambda(t)$ rappresenta la densità condizionale di probabilità, che un oggetti si guasti nel prossimo istante

In caso di distribuzione esponenziale In questo caso la distribuzione della vita residua di un oggetto di eta t è identica a quella di un oggetto nuovo, quindi dobbiamo avere un **valore costante**:

$$\lambda(t) := \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \boldsymbol{\lambda}$$

Il valore trovato è *l'intensità della distribuzione esponenziale*

La funzione λ determina **univocamente** la F, quindi per definizione:

$$\begin{aligned}\lambda(s) &:= \frac{f(s)}{1-F(s)} \\ &= \frac{F'(s)}{1-F(s)} \\ &= \boldsymbol{-\frac{d}{ds}\log(1-F(s))}\end{aligned} \tag{6}$$

Possiamo integrare sto mapazzone con i membri tra 0 e t ottenendo che:

$$\int_0^t \lambda(s)\,ds = -\log(1-F(t)) + \log(1-F(0)) = \boldsymbol{-\log(1-F(t))}$$

Ottenendo alla fine che

$$1-F(t) = \exp\left\{-\int_0^t \lambda(s)\,ds\right\}$$

La funzione di ripartizione di una var. aleatoria continua può essere specificata tramite la *corrispondente funzione di intensità di rotture*

14.3 Il ruolo della distribuzione esponenziale

14.3.1 Interruzione al fallimento r-esimo

in questa sezione vediamo l'esame simultaneo di un campione di n oggetti con *tempi di vita esponenziali* e *indipendenti* con media incognita θ e terminiamo il test non appena raggiungiamo un numero fissato $r \leq n$
I dati che abbiamo sono gli *r* tempi di vita registrati, nel seguente ordine:

$$x_1 \leq x_2 \leq \cdots \leq x_r$$

Se denotiamo con X_i il tempo di vita dell'oggetto possiamo riassumere la parte di sopra come segue:

$$X_{i1} = x_1, X_{i2} = x_2, \ldots, X_{ir} = x_r$$

La *densità di probabilità* delle X_{ij} è

$$f_{X_{i_j}}(x_j) = \frac{1}{\theta} e^{-x_j/\theta}, \qquad j = 1, 2, \ldots, r$$

La *densità congiunta* invece è la seguente:

$$f_{X_{i_j},\ldots,X_{i_r}}(x_1, \ldots, x_r) = \prod_{j=1}^{r} \frac{1}{\theta} e^{x_j/\theta}$$

Per verificare la probabilità che le altre $n - r$ siano tutte maggiori x_r è data dall'indipendenza:

$$P(X_j > x_r \text{ per } j \neq \{i_1, i_2, \ldots, i_r\}) = (e^{-x_r/\theta})^{n-r}$$

Di conseguenza la likelihood (o verosimiglianza) dei dati osservati, che viene denotata con $L(x_1, x_2, \ldots, x_r, i_1, i_2, \ldots, i_r|\theta)$, è data da

$$\begin{aligned}
&L(x_1, x_2, \ldots, x_r, i_1, i_2, \ldots, i_r|\theta) \\
&\quad = f_{X_{i_1},\ldots,X_{i_r}}(x_1, \ldots, x_r) P(X_j > x_r, j \notin \{i_1, i_2, \ldots, i_r\}) \\
&\quad = \frac{1}{\theta^r} e^{-x_1/\theta} e^{-x_2/\theta} \cdots e^{-x_r/\theta} (e^{-x_r/\theta})^{n-r} \\
&\quad = \frac{1}{\theta^r} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{r} x_i - \frac{(n-r)x_r}{\theta} \right\} \qquad (14.3.2)
\end{aligned}$$

Se abbiamo bisogno della verosomiglianza in *funzione* solo degli r tempi di rottura, la funzione di likelihood sarebbe:

$$f(x_1, x_2, \ldots, x_r) = \frac{n!}{(n-r)!\theta^r} \exp\left\{\frac{1}{\theta}\sum_{i=1}^{r} x_i - \frac{(n-r)x_r}{\theta}\right\}$$

Per calcolare lo stimatore di massima verosomiglianza di θ invece facciamo cosi:

$$\hat{\theta} := \frac{\sum_{i=1}^{r} X_{(i)} + (n-r)X_{(r)}}{r} =: \frac{\tau}{r}$$

Dove τ viene definito come **Total Time of Test statistic**

τ rappresenta la somma delle statistiche Y_i per $i = 1, 2, \ldots, r$ che indicano il tempo totale di funzionamento racchiuso tra la rottura dell'oggetti (i-1)-esimo e quella dell'i-esimo

Il calcolo per trovarlo è il seguente

$$\tau = \frac{j=1}{r} Y_j$$

Dato che la *somma* di variabili aleatorie esponenziali ha distribuzione *gamma* otteniamo che il nostro τ è una **gamma** con parametri r e $1/\theta$

Sfruttando questa relazione:

$$\frac{2\tau}{\theta} \sim \mathcal{X}^2_{2r}$$

Notiamo subito (ensomma) che:

$$P(\mathcal{X}^2_{1-\frac{\alpha}{2},2r} < 2\tau/\theta < \mathcal{X}^2_{1\frac{\alpha}{2},2r}) = \mathbf{1 - \boldsymbol{\alpha}}$$

E quindi sappiamo che abbiamo un livello di confidenza $\mathbf{1 - \boldsymbol{\alpha}}$ nell'affermare che:

$$\theta \in \left(\frac{2\tau}{\mathcal{X}^2_{\frac{\alpha}{2},2r}}, \quad \frac{2\tau}{\mathcal{X}^2_{1-\frac{\alpha}{2},2r}}\right)$$

Questo per il caso **bilaterale**

Esempio 50 transistor vengono messi in funzione simultaneamente. L'esperimento si conclude quando il 15-esimo(r) si rompe. Il Total time on test(TTT) è di 525 ore. Trovare un intervallo di confidenza del 95% per la vita media di un componente. La distribuzione è esponenziale.
Mediamente si rompono:

$$\hat{\theta} = \frac{TTT}{r} \qquad \text{numero medio di guasti in 525 ore}$$
$$\hat{\theta} = \frac{525}{15} = 35$$

Per trovare il livello di confidenza (dio merda) utilizziamo:

$$\theta \in \Big(\frac{2TTT}{\mathcal{X}^2_{\frac{\alpha}{2},2r}}, \frac{2TTT}{\mathcal{X}^2_{1-\frac{\alpha}{2},2r}}\Big)$$

$$\boldsymbol{\theta \approx (22.35, 62.54)}$$

14.4 prove simultanee

Dobbiamo analizzare una sequenza una serie di oggetti, ciascuno e tempo di vita esponenziale con media sconosciuta θ. L'esperimento viene concluso dopo un periodo prefissato T. I dati che abbiamo sono il numero r di oggetti guasti entro T e il tempo di vita di ogni oggetto $x_1, x_2, \ldots, x_r$.

MLE della media : numero medio di oggetti che si rompono fino a T:

$$\hat{\theta} = \frac{T}{r}$$

Intervallo di confidenza di $\hat{\theta}$:

$$\theta \in \left(\frac{2T}{\mathcal{X}^2_{1-\frac{\alpha}{2},2r}}, \frac{2T}{\mathcal{X}_{\frac{\alpha}{2}},2r}^{2} \right)$$