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 scimage.pngonosciuti il massimo che possiamo fare è quello di ottenere *una stima* dei parametri *incogniti*.

Codesti signorini 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:

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

Possiamo derivare rispetto a p:

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

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:

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

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

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

2.4 MLE distribuzione Normale

TODO AGGIUNGERE THETA MAX THETA / 2

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 (metodo semplificato per migliorarci la vita che è già una merda) 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 (che lasciamo al libro) ci porta alle formule per le stime:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_1$$

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

TODO TEORIA DEL LIMITE CENTRALE

3 Intervalli di confidenza

3.1 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*

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

Chiedo aiuto alla regia, non so cosa stia sta roba ma comunque:

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \approx 0.95\right)$$

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%

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 $\overline{x}$:

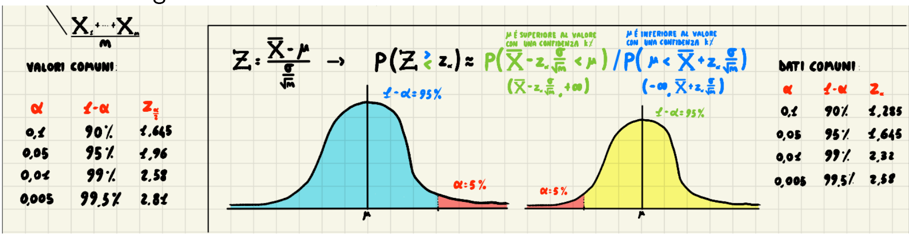
$$\overline{x} = \frac{81}{9} = 9$$

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

Figure 1: TODO CAPIRE CHE SFACCIMM è STA ROBA



3.2 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

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

Per Unilaterale

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

3.3 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:

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

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:

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

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

4 Intervalli di predizione

4.1 μ incognita e varianza σ^2 incognita

Supponiamo che $X_1, X_2, \ldots, X_n, X_{n+1}$ sia un campione normale di media μ e varianza σ^2 entrambe *incognite*

$$\mu = \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Per la riproducibilità

$$X_{n+1} - \overline{X}_n \sim \mathcal{N}(0, \sigma^2 + \frac{\sigma^2}{n}) \longrightarrow \frac{X_{n+1} - \overline{X}_n}{\sigma\sqrt{1 + 1/n}}$$

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

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

Questa grandezza è *indipendente* da $\overline{X}_n$ quindi otteniamo

$$\frac{X_{n+1} - \overline{X}_n}{S_n\sqrt{1 + 1/n}} \sim t_n - 1$$

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:

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

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]

4.2 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 :

$$P\left(0 < \sigma^2 < \frac{(n-1)S^2}{\mathcal{X}^2_{1-\alpha,n-1}}\right) / P\left(\frac{(n-1)S^2}{\mathcal{X}^2_{\alpha,n-1}} < \sigma^2\right) \tag{2}$$

4.3 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
$\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}$ è lo stimatore di massima verosomiglianza $\mu_1 - \mu_2$

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

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

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

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	$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)$	$\left(0, \quad \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, \quad \infty\right)$

Per ottenere uno *stimatore non puntuale*, dobbiamo conoscere la distribuzione di $\overline{X}-\overline{Y}$ poiche:

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

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

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

Per caso Unilaterale

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

5 Intervalli di confidenza

5.1 Intervalli approssimati 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

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

Per caso Unilaterale

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

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	$\left(\hat{p} - z_{\alpha} \sqrt{\hat{p}(1-\hat{p})/n}, \quad \infty\right)$

5.2 Qualità ed efficienza degli stimatori

Sia $X := (X_1, X_2, \ldots, X_n)$ 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:

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

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

5.2.1 Bias e Polarizzazione

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

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

Questo viene detto *bias* di d come stimatore di θ
Se il bias è nullo, si dice che è uno stimatore *corretto* o *non distorto*

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

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

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

sono entrambi *stimatori non distorti di* θ

verifichiamo:

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

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

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

Questo è lo stimatore migliore con Varianza minima

Regola generale

$$d_3(X_1, X_2, \ldots, X_n) := \sum_{i=1}^{n} \lambda_i X_i \text{ è corretto se } \sum_{i=1}^{n} \lambda_i = 1$$

5.2.2 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

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

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

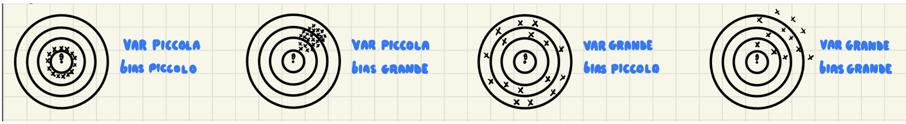
La migliore combinazione lineare delle d_i per l'errore quadratico medio è:

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

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

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

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



5.3 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) $\mathbb{E}[\boldsymbol{X_i}] = \boldsymbol{\theta/2}$ è uno stimatore naturale per θ è dato da

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

$$d_2 = \hat{\theta} = MLE = \max(X_i)$$

Siccome $\mathbb{E}[d_1] = \theta$ otteniamo che:

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

$$P\{X_i = x\} = P^x(1-P)^x \quad x \in \{0,1\}$$

Dove **X** è una *variabile aleatoria* e **x** una *variabile sperimentale*

$$f(x_1 \dots x_n/P) = P^{x_1}(1-P)^{1-x_1} \cdot P^{x_2}(1-P)^{1-x_2} \dots P^{x_n}(1-P)^{1-x_n} =$$

$P^{\sum_i^n x_i}(1-P)^{n-\sum_1^n x_i} \longrightarrow$ Bisogna trovare il **massimo** della funzione

$$\begin{aligned} log(f(x_1 \dots x_n/P)) &= \sum_1^n x_i logP - (n - \sum_i^n x_i) log(1-P) \\ &= \frac{d}{dP}[log(f)] = 0 = \frac{1}{\hat{P}}\sum_i^n x_i - \frac{n - \sum_i x_i}{(1-\hat{P})} \\ &= (1-\hat{P})\sum_i x_i = \hat{P}(n - \sum_i x_i) \\ &= \hat{P} = \frac{\sum_i x_i}{n} \quad \text{MLE} \end{aligned}$$

Esercizio 1 Probabilità che Oneto dia 30L (Lode)
$n = 120$

$\sum_i^{120} x_i = 18$
$\hat{P} = \frac{18}{120} = 0.15 \rightarrow 15\%$

Esercizio 2 N studenti da 30 e lode
$n_1 = 18 \leftarrow$ Oneto
$n_2 = 20 \leftarrow$ Anguita
$n_{1,2} = 10 \leftarrow$ 30L sia con Oneto che con Anguita
$N = ?$ Studenti da **30 e Lode**

$$\hat{P}_1 \approx \frac{n_1 2}{n_2} \qquad \hat{P}_1 \approx \frac{n_1}{N} \qquad \frac{n_{1,2}}{n_2} = \frac{n_1}{N}$$

$\Longrightarrow N = \frac{n_1 \cdot n_2}{n_{1,2}} \rightarrow \frac{18 \cdot 20}{10} = 36$

MLE POISSON

$$\begin{aligned} f(x_1, x_2 \dots x_n / \lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_i x_i}}{x_1! x_2! \dots x_n!} \end{aligned}$$

Formula generica: $\lambda = \frac{\sum_i x_i}{\lambda}$ MLE

Esercizio 3 Stima del numero di incidenti medio in auto n = 10
$x_1 = \{4, 0, 6, 5, 2, 1, 2, 0, 4, 3\}$
$\hat{\lambda} = \frac{\sum_i x_i}{n} = \frac{27}{10} = 2.7$

$$P\{x \leq 2\} = e^{-2.7}(\frac{2.7^0}{0!} + \frac{2.7^1}{1!} + \frac{2.7^2}{2!}) \approx .4936 \rightarrow 49.36\%$$

Probabilità che non ci siano più di **2 incidenti**

MLE UNIFORME

$$f(x_1, x_2 \dots x_n/\theta) = \begin{cases} \frac{1}{\theta^n} & 0 < x_i < \theta \\ 0 & \text{altrimenti} \end{cases}$$

$\hat{\theta} = \max\{x_i\}$
$\frac{\hat{\theta}}{2} = \frac{\max\{x_i\}}{2}$

MLE GAUSSIANA

$$f(x_1, x_2 \dots x_n/\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}$$

$$(\frac{1}{2\pi})^{\frac{n}{2}} \frac{1}{\sigma^n} e^{\frac{-\sum_i (x_i-\mu)^2}{2\sigma}}$$

$log[f] = -\frac{n}{2} log 2\pi - n log\sigma - \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$

$\frac{dlogf}{d\mu} = 0 = \frac{\sum_i (x_i-\mu)^2}{\sigma^2} \longrightarrow \hat{\mu} = \frac{\sum_i x_i}{n}$
$\frac{dlogf}{d\sigma} = 0 = -\frac{n}{\sigma} + \frac{\sum_i (x_i-\mu)^2}{4\sigma^4} \rightarrow \sigma = \sqrt{\frac{\sum_i (x_i-\mu)^2}{n}}$

Esercizio primo
$x_1 = 1.7$
$x_2 = 1.82$
$x_3 = 1.73$
$x_4 = 1.7$
$x_5 = 1.8$

$$\hat{\mu} = \frac{\sum_i x_i}{n} = \frac{1.7 + 1.82 + 1.73 + 1.7 + 1.8}{5} = 1.75$$

$$\hat{\sigma} = \sqrt{\frac{0.05^2 + 0.07^2 + 0.02^2 + 0.05^2 + 0.05^2}{5}} \approx 0.051$$

Intervalli di confidenza normali TODO

Intervalli di confidenza gaussiani σ^2 Nota

$x_1 m x_2 \dots x_n$

$\hat{\mu} \longleftarrow \mu$

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

$$P(-1.96 < \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < +1.96) = 0.95$$

$$\longrightarrow P(-1.96\frac{\sigma}{\sqrt{n}} < \overline{x} - \mu < 1.96\frac{\sigma}{\sqrt{n}})$$

$$P(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

Esempio: Sistema di comunicazione $\sigma^2 = 4 \quad n = 9$

$$x_1 = \{5.85, 12, 15, 7, 9, 7.5, 6, 5, 10.5\}$$

$$\hat{\mu} = \frac{1}{n}\sum_i^n x_i = \frac{1}{9}\sum_i^n x_i = \frac{81}{9} = 9$$

$$P\left(9 - 1.96\frac{\sigma}{\sqrt{m}} < \mu < 9 + 1.96\frac{\sigma}{\sqrt{m}}\right) = 0.95$$

$$p\left(9 - 1.96\frac{2}{3} < \mu < 9 + 1.96\frac{2}{3}\right) = 0.95$$

$\longrightarrow [7.693, 10.31] \rightarrow \mu$ si trova tra 7.693 e 10.31

In generale Prob $= 1-\alpha$

$$(\overline{x} - z_a \frac{\sigma}{\sqrt{n}}, \overline{x} + z_a \frac{\sigma}{\sqrt{n}}) \to \text{Si rileva dalle tavole}$$

5.4 Intervalli di confidenza (Bilaterali)

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

$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\overline{X} \sim (\mu, \frac{\sigma^2}{n})$$

$$\mathcal{Z} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1) \quad Var(\frac{x}{2}) = \frac{1}{\sigma^2} Var(x)$$

Supponiamo che σ sia nota:

$$\Pr\left\{-z_{\frac{\alpha}{2}} < Z < +z_{\frac{\alpha}{2}}\right\} = 1-\alpha$$

$$\Pr\left\{-z_{\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{m}}} < +z_{\frac{\alpha}{2}}\right\} = 1-\alpha$$

$$\Pr\left\{-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{m}} < \bar{x} - \mu < +z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{m}}\right\}$$

$$\Pr\left\{-\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{m}} < -\mu < -\bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{m}}\right\} =$$

$$\Pr\left\{\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{m}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{m}}\right\} = 1-\alpha$$

5.5 Intervalli di confidenza (Unilaterali)

$$\Pr\{z < z_\alpha\} = 1 - \alpha$$
$$\Pr\left\{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{m}}} < z_\alpha\right\} = 1 - \alpha$$
$$\Pr_r\left\{\bar{x} - \mu < z_\alpha \frac{\sigma}{\sqrt{m}}\right\} = 1 - \alpha$$
$$\Pr\left\{-\mu < -\bar{x} + z_\alpha \frac{\sigma}{\sqrt{m}}\right\} = 1 - \alpha$$
$$\Pr\left\{\bar{x} - z_\alpha \frac{\sigma}{\sqrt{m}} < \mu\right\} = 1 - \alpha$$
$$\mu \in \left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{m}}, +\infty\right)$$

5.6 Esempio:

Pesca stagionale dei salmoni (*Fisso intervallo -> trovo* n)
Ad ogni stagione il peso medio dei salmoni è diverso ma $\sigma = 0.3$ Kg
Intervallo di confidenza al 95%, quindi $\alpha = 0.05$

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

$$1.96\frac{\sigma}{\sqrt{n}} \geq 0.1 \quad \sqrt{n} \geq \frac{1.96}{0.1}\sigma$$

$$n \geq (\frac{1.96}{0.1}0.3)^2 = 5.88^2 \approx 34.6 \leftarrow \text{salmoni}$$

5.7 Intervallo di confidenza

con *media* e *varianza* **incognite**

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sigma \quad \text{Non nota}$$

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_i^n \left(x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{n-1} \sum_i \left(x_i^2 + \bar{x}^2 - 2x_i\bar{x}\right)$$

$$= \frac{1}{n-1} \sum_i x_i^2 + \frac{n\bar{x}^2}{n-1} - 2\bar{x}\frac{\bar{x}n}{n-1}$$

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim T_n - 1 \quad \text{(T studenti con n gradi di libertà)}$$

Esempio: Trasimttente (μ) e ricevitore (μ + rumore)

$$95\%(7.69, 10.31) \quad \hat{\mu} = 9, \sigma^2 = 4$$

$X_i\{5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5\}$
$\hat{\mu} = \overline{X} = \frac{1}{9}\sum_i^n X_i = \frac{81}{9} = 9$
$s^2 = \frac{1}{8}\sum_i (X_i^2 - 9.81) \approx 9.5 \quad s = 3.082$

$$\mu \in (9 - 2.306\frac{3.082}{3}, 9 + 2.306\frac{3.082}{3}) = (6.63, 11.37)$$

Si può dimostrare che $T_{\frac{\alpha}{2} \cdot n-1}\mathbb{E}[S] \geq z_\alpha \sigma$

5.8 Integrali Monte Carlo

$$\theta = \mathbb{E}[f(u)] = \int_{-\infty}^{+\infty} f(u)p(u)\,du = \int_{-\infty}^{+\infty} f(u)\,du$$

Esempio :
$\int_0^1 \sqrt{1-x^2}\,dx =? \ \mathbb{E}[\sqrt{1-x^2}] \quad n = 100$
$X_i = \sqrt{1-U_i^2} \quad X = \{X_1, X_2 \dots X_1 00\}$
$\hat{\theta} = \overline{X} \pm t_{\frac{\alpha}{2}}, 99\frac{s}{\sqrt{100}}$ $\rightarrow$ Per vedere se il risultato è corretto *(confidenza)*

5.9 Intervallo di confidenza di Bernoulli

n esperimenti
Binomiale
media np
varianza np(1-p)

$$\hat{P} = \frac{1}{n}\sum_i^n X_i \quad X_i \in \{0,1\}$$

$$X = n\hat{P} \quad P_r\{-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}} \approx 1-\alpha\}$$

Dove z = $\frac{X-np}{\sqrt{np(1-p)}}$

$$\frac{x-nP}{\sqrt{nP(1-P)}} \sim \mathcal{N}(0,1)$$

$$\rho_r \left\{ -z_{\frac{a}{2}} < \frac{x - mp}{\sqrt{mp(1-\hat{p})}} < z_{\frac{a}{2}} \right\} \cong 1 - \alpha$$
$$\rho_r \left\{ \hat{p} - z_{\frac{a}{2}} \sqrt{\frac{p(1-p)}{m}} < \mu < \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-p)}{m}} \right\} \simeq 1 - \alpha \tag{3}$$

6 Intervalli di confidenza

Se σ^2 è nota allora:

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \overline{X} = \frac{1}{n} \sum_i^n X_i$$

$$\mu \in (-\infty, \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}})$$
$$\mu \in (\overline{X} - z_{\frac{\sigma}{\sqrt{n}}}, \overline{X} + z_{\frac{\sigma}{\sqrt{n}}}) \quad p_r(1-\alpha)$$
$$\mu \in (-\infty, \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}})$$
$$\mu \in (\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$$

$s^2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})$
Se σ^2 è ignota allora:

$$\mu \in (\overline{X} - z_{\frac{\alpha}{2}}, n - 1 \frac{s}{\sqrt{n}}) \quad \sigma^2 \to s^2 = z \to t$$

6.1 Intervallo di confidenza nella varianza

$$(n-1)\frac{S^2}{\sigma^2} \sim \mathcal{X}^2 \quad X_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$p_r \left\{ \mathcal{X}^2_{1-\frac{\alpha}{2},n-1} \leq (n-1)\frac{s^2}{\sigma^2} \leq \mathcal{X}_{\frac{\alpha}{2},n-1} \right\}$$

$$p_r \left\{ \frac{s^2(n-1)}{\mathcal{X}_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\mathcal{X}^2_{1-\frac{\alpha}{2},n-1}} \right\} = 1-\alpha$$

$$\sigma^2 \in \left(\frac{s^2(n-1)}{\mathcal{X}^2_{\frac{\alpha}{2},n-1}}, \frac{s^2(n-1)}{\mathcal{X}_{1-\frac{\alpha}{2},n-1}} \right) \quad p_r = 1-\alpha$$

Esempio: Laminatoio n = 4 $X_i = \{0.123, 0.124, 0.126, 0.12\}$spessore in **mm**

Svolgimento

$$\frac{1}{4}\sum_i^4 X_i = \frac{0.493}{4} = 0.12325$$

$$\frac{1}{4-1}\sum_i^4 (X_i - 0.12325)^2 = 1.875 \cdot 10^{-5}$$

$$\sigma^2 \in \left(\frac{s^2(n-1)}{9.348}, \frac{s^2(n-1)}{0.216} \right)$$

Dove **9.348** e **0.216** sono ricavati dalle tabelle
Facciamo la radice:
$\sigma \in (0.0014, 0.0093) \to 95\%$

6.2 Intervallo di confidenza

della differenza di due medie:

N campioni

$$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

$$\overline{X} = \frac{1}{n}\sum_i^n X_i$$

M campioni

$$Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$\overline{Y} = \frac{1}{m}\sum_i^m Y_i$$

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

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

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

Se σ_1^2, σ_2^2 non sono note:

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

$(n-1)\frac{S_1^2}{\sigma_1^2} \sim \mathcal{X}_{n-1}^2$

$s_2^2 = \frac{1}{m-1}\sum_i^m (X_i - \overline{Y})$

$(n-1)\frac{s_2^2}{\sigma_2^2} \sim \mathcal{X}_{n-1}^2$

Possiamo andare avanti solo se $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$

$$(n-1)\frac{s_1^2}{\sigma^2} + (n-1)\frac{s_2^2}{\sigma^2} \sim \mathcal{X}_{n+m-2}^2$$

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}} \longrightarrow \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p(\frac{1}{n} + \frac{1}{m})}}$$

$$\sim \mathcal{N}(0,1) \qquad\qquad \sim T_{n+m-2}$$

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

Se σ sono ignote ma uguali

$\mu_1 - \mu_2 \in (\overline{X} - \overline{Y} - T_{\frac{\alpha}{2},n+m-2}\sqrt{s^2(\frac{1}{n} + \frac{1}{m})})$
$\overline{X} - \overline{Y} + T_{\frac{\alpha}{2},n+m-2}\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}$

6.3 Intervallo di previsione

$X_1, \dots X_n, X_{n+1} \sim \mathcal{N}(\mu, \sigma^2)$
$\overline{X}_n = \frac{1}{n}\sum_i^n X_i \quad \overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
$\overline{X}_n - X_{n+1} \sim \mathcal{N}(0, \sigma^2 + \frac{\sigma^2}{n}) \to (\mu - \mu, \sigma^2 + \frac{\sigma^2}{n})$
$\sigma^2(1+\frac{1}{n})\ \frac{X_n - X_n + 1}{\sigma\sqrt{1+\frac{1}{n}}} \sim \mathcal{N}(0,1) \qquad s_n^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2$

$$X_{n+1} \in (\overline{X}_n - T_{\frac{\alpha}{2},n-1} s_n\sqrt{1+\frac{1}{n}}, \overline{X}_n + T_{\frac{\alpha}{2},n-1} s_n\sqrt{1+\frac{1}{n}}) \longrightarrow P_r(1-\alpha)$$

Esempio smartwatch contapassi $\boldsymbol{n = 7}$

$$\begin{array}{lll} LUN & 6922 & X_1 \\ MAR & 5333 & X_2 \\ MER & 7420 & X_3 \end{array} \qquad \begin{array}{lll} GIO & 7432 & X_4 \\ VEN & 6252 & X_5 \\ SAB & 7005 & X_6 \end{array}$$

$$DOM \quad 6752 \quad X_7$$

$$\overline{X}_n = \frac{1}{n}\sum_i^m X_i = \frac{47016}{7} \approx 6717$$

$$1 - \alpha = 95\% \quad \alpha = 5\%$$

$$t_{0.0025,6} = 2.997$$

$$S_n = \sqrt{S_n^2} = 7.333.8$$

$$x_{n+1} \in (6717 - 2.447 \cdot 733397\sqrt{1+\frac{1}{7}}, 6717 + 2.447 \cdot 73397\sqrt{1+\frac{1}{7}})$$

$X_{n+1} \in (9796, 8637)\mu \in (6037, 7396)$

6.4 Qualità di uno stimatore

$X = X_1 \dots X_n$ $\theta \leftarrow$ parametro $d(x) \leftarrow$ stimatore di θ

$$\boldsymbol{(d(x)-\theta)^2} \qquad\qquad \mathbb{E}[\boldsymbol{(d(x)-\theta)^2}]$$

Errore Quadratico (*misura della qualità*) Errore Quadratico Medio (*M.S.E*)
Rischio $r(d,\theta) = \mathbb{E}[(d - \theta)^2]$ Lo stimatore "ottimo" sarà quello con il rischio minimo
-> d con r minimo θ

Esempio $d^*(x) = 4$ se$\theta = 4 \Rightarrow d^* =$ stimatore ottimo(per tutti gli altri valori non va

6.5 Proprietà di uno stimatore

Def: $b_\theta(d) = \mathbb{E}[d] - \theta \rightarrow$ bias o polarizzazione Uno stimatore non è **polarizzato** se $b_\theta(d) = 0$

Esempio : $X_1 \dots X_n$ θmedia
$d_1(X_1 \dots X_n) = X_1$
$d_2(X_1 \dots X_n) = \frac{X_1+X_2}{2}$
$d_3(X_1 \dots X_n) = \frac{X_1+X_2+\dots X_n}{n}$ Tutti questi sono **unbiased**

6.6 Stimatore unbaieseo

$r(d,\theta) = \mathbb{E}[(d(x)-\theta)^2] = \mathbb{E}[(d(x)-\mathbb{E}[d(x)])^2] = Var(d)$
tra gli stimatori non polarizzati di ottimo è quello con la varianza minima

6.7 Valutazione di uno stimatore

$$X = X_1 \dots X_n \quad \theta = ?$$

Dove θ è un *parametro* e $d(x)$ è uno *stimatore* di θ

$r(d,\theta)$(mse) rischio $\qquad b_\theta(d) = \mathbb{E}[d] - \theta$
se $b_\theta(d) = 0 \Rightarrow r(d,\theta) = Var(d)$

se $b_\theta(d) \neq 0$? $r(d,\theta)$ = ?

$$\begin{aligned} r(d,\theta) = \mathbb{E}[(d(x)-\theta)^2] &= \mathbb{E}[(d(x)-\mathbb{E}[d]+\mathbb{E}[d]-\theta)^2] \\ &= \mathbb{E}[(d-\mathbb{E}[d])^2 + (\mathbb{E}[d]-\theta)^2 - 2(d-\mathbb{E}[d])(\mathbb{E}[d]-\theta)] \\ &= \mathbb{E}[(d-\mathbb{E}[d])^2] + \mathbb{E}[(\mathbb{E}[d]-\theta)^2] - 2(\mathbb{E}[d]-\theta)\cdot\mathbb{E}[(d-\mathbb{E}[d])] \end{aligned}$$

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

6.8 Esempio:

Stimatore della media di una *distribuzione uniforme*

$\mathbb{E}[X_i] = \theta/2 \qquad d_1 = 2\frac{1}{n}\sum_i^n X_i$
$X_1, X_2 \dots X_n \qquad d_2 = \max X_i$

$d_1 : \mathbb{E}[d_1] = \frac{2}{n}\sum_i \mathbb{E}[X_i] = \frac{2}{n} n \frac{\theta}{2} = \theta$

$r(d_1,\theta) = Var(d_1) = \frac{4}{n^2} n Var(X_i) = \frac{4}{n}\frac{\theta^2}{12} = \boldsymbol{\frac{\theta^2}{3n}} \Leftarrow$ *Unbiased*

$$F_2(x) = P_r\{d_2(x) \le x\} = P_r\{\max X_1 \le x\}$$
$$= P_r\{X_1 \le, \forall i \in 1\} = \prod_{i=1}^{n} P_r\{X_i \le x\} = (\frac{x}{\theta})^n$$
$$f_2(x) = \frac{d}{dx} F_2(x) = n\frac{x^{n-1}}{\theta^n} \quad x \le \theta$$

$$\mathbb{E}[d_2] = \int_0^\theta x f_x(x)\,dx = \int_0^\theta \frac{n}{\theta^n} x^n\,dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\bigg|_0^\theta\right] = \frac{n}{n+1}\theta$$

$$\mathbb{E}[d_2^2] = \frac{n}{\theta^n}\int_0^\theta x^2 f(x)\,dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx = \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\bigg|_0^\theta\right] = \frac{n}{n+2}\theta^2$$

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

$r(d_2,\theta) = Var(d_2) + (\mathbb{E}[d_2] - \theta)^2 = \boldsymbol{\frac{2\cdot\theta^2}{(n+1)(n+2)}}$

$n \ge 4 \quad r(d_2,\theta) < r(d_1,\theta) \qquad d_3 = \frac{n+1}{n} d_2$

In sintesi

$$r(d_1, \theta) = \frac{\theta^2}{3n} \Leftarrow \text{Unbiased}$$
$$r(d_2, \theta) = \frac{2\theta^2}{(n+1)(n+2)} \Leftarrow \text{Biased}$$
$$r(d_3, \theta) = \frac{\theta^2}{n^2+2n} \Leftarrow \text{Unbiased}$$
$$r(d_4, \theta) = \frac{\theta^2}{(n+1)^2} \Leftarrow \text{Biased}$$

7 Test di ipotesi

Ipotesi: Affermazione rispetto a uno o più parametri di una distribuzione Ipotesi da confutare: H_0 (ipotesi nulla)

Esempio :

$$\begin{aligned} X_1 \dots X_n \sim \mathcal{N}(\mu, \sigma^2) \\ H_0 : \mu = 0 \\ H_a : \mu \neq 0 \end{aligned} \tag{4}$$

Diamo per scontato che l'ipotesi sia **vera**
Dobbiamo cercare di *confutarla*

Definizione Regione critica tale che:

$(X_1 \dots X_n) \in C \rightarrow H_0$è rifiutata
$(X_1 \dots X_n) \notin C \rightarrow H_0$è accettata
$\alpha =$ Livello di **significatività** del test ($\alpha = 10\%, 5\% \dots$)

Procedimento

- Fisso alpha
- Suppongo che α sia vera
- calcolo stima di μ
- verifico che non sia "*troppo distante*"

$X_1 \dots X_n \sim \mathcal{N}(\mu, \sigma^2)$
$H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$
$\overline{X} = \frac{1}{n} \sum_i X_i$

Regione critica $\quad \{(X_1 \dots X_n) : |\overline{X} - \mu_0| > c\}$
$P_{r_{\mu_0}} \left\{ |\overline{X} - \mu_0| > c \right\} = \alpha$
$P_{r_{\mu_0}} \left\{ \frac{|\overline{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} > \frac{c}{\frac{\sigma}{\sqrt{n}}} \right\} = \alpha$
$P_{r_{\mu_0}} \{|z| > z_\alpha\} = \alpha$

Esempio (5 transimissioni)
$\boldsymbol{n = 5}$ $\qquad H_0 : \mu = 8$
$\overline{X} = 9.5$ $\qquad \alpha = 5\%$

Ipotizzando che H_0 sia vera:

$$\frac{|\overline{X} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|9.5 - 8|}{\frac{2}{\sqrt{5}}} \approx 1.68$$

Se:
$\alpha = P_r$(rifiuto H_0 / H_0 vera)
$\alpha \uparrow$ più *"facile"* rifiutare l'ipotesi
$\alpha \downarrow$ più *"difficile"* rifiutare l'ipotesi

7.1 Metolodogia alternativa

$$Ts = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \rightarrow \text{Statistica di test}$$

$P-$value = **probabilità** di ottenere un valore più "anomalo" di quello osservato

Esempio: $X_i \sim \mathcal{N}(\mu, 4)$

$n = 5$ $\qquad H_0 : \mu = 8$

$\overline{X} = 8.5$ $\qquad H_a : \mu \neq 8$

$$\frac{|\overline{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} = \frac{|8.5 - 8|}{\frac{2}{\sqrt{5}}} = \frac{\sqrt{5}}{2} 0.5 \approx 0.559$$

$$P\{|z| > 0.559\} = 2P\{z > 0.559\} \approx 2 \cdot 0.288 = 0.579 \rightarrow \text{P-value}$$

Se $\boldsymbol{\overline{X} = 11.5}$:

$$\frac{|\overline{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} = \frac{|11.5 - 8|}{\frac{2}{\sqrt{5}}} \approx 3.913$$

$$P\{|z| > 3.913\} = 2P\{z > 3.913\} \leq 0.00005 \rightarrow \underline{\text{Rifiuto ipotesi } H_0}$$

7.2 Test di Hp unilaterale

$$H_0 : \mu = \mu_0 (\mu \leq \mu_0) \qquad\qquad H_a : \mu > \mu_0$$

$$C = \{(X_1 \dots n) \cdot \overline{X} - \mu_0 > c\}$$

$$P_{r_{\mu_0}}\{\overline{X} - \mu_0 > c\} = P_r\{\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{c}{\frac{\sigma}{\sqrt{n}}}\} = P_{r_{\mu_0}}\{z > z_a\} = \alpha$$

Statistica test $\boldsymbol{\frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}} \leq \boldsymbol{z_\alpha}$ accetto

7.3 Test di ipotesi

H_0	H_a	TS	Livello α	P - Value
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{1}}$	Rifiuto H_0 se $TS > \frac{z\alpha}{2}$	$2P(z \geq \|TS\|)$

Altre ipotesi :

H_0	H_a	TS	Livello α	P - Value
$\mu < \mu_0$	$\mu > \mu_0$	$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{2}}$	$H_0 \quad z_\alpha > TS$	$P(z \geq TS)$
$\mu \geq \mu_0$	$\mu < \mu_0$	//	$H_0 \quad z_\alpha < -TS$	$P(z \leq TS)$

7.4 Uguaglianza media di due popolazioni

$X_1 \dots X_n \sim \mathcal{N}(\mu_1, \sigma_2^2)$ $\qquad$ $Y_i \dots Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$

$\overline{X} = \frac{1}{n} \sum_i^n X_i$ $\qquad$ $\overline{Y} = \frac{1}{m} \sum_i^m Y_i$

$S_x^2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2$ $\qquad$ $S_y^2 = \frac{1}{m-1} \sum_i (Y_i - \overline{Y})^2$

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

H_0	H_a	TS		
$\mu_1 = \mu_2$	$\mu \neq \mu_2$	$\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_{1/n}^2 + \sigma_{2/m}^2}}$	Livello α	P - Value
$\mu_1 = \mu_2$	$\mu \neq \mu_2$	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_p^2(\frac{1}{n}+\frac{1}{m})}}$	rif. $\lvert TS \rvert > z_{\frac{\alpha}{2}}$	$2P(z \geq \lvert TS \rvert)$
$\mu_1 = \mu_2$	$\mu \neq \mu_2$	$S_i \in$ T-student		

4) T-test per coppie di dati Se X_1 e X_2 <u>**NON**</u> sono indipendenti

$W_i = X_i - Y_i$ $\qquad$ $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$

$Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$

ES Manutenzione (n guasti) tagliand

$H_0 : \mu_a - \mu_b \geq 0 \quad \overline{W} = \frac{1}{5}(-7.5 + 2.5 - 2.5 - 3.5 - 1.5) = -2.5$

$S_W^2 = \frac{1}{4}(W_i - \overline{W})^2 = 13$

$Ts = \frac{\frac{\overline{W}}{S_W}}{\sqrt{n}} = \frac{\frac{-2.5}{\sqrt{13}}}{\sqrt{5}} = 1.55$

$$P_r\{T_{n-1} \leq Ts\} = \{T_4 \leq Ts\}$$

5) Test sulla varianza

$$H_0 : \sigma^2 = \sigma_0^2 \quad H_a : \sigma^2 \neq \sigma_0^2$$

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

$Pr\{\mathcal{X}^2_{1-\frac{\alpha}{2},n-1} \leq \frac{(n-1)s^2}{\sigma_0^2} \leq \mathcal{X}^2_{\frac{\alpha}{2},n-1}\} = 1 - \alpha$

Uguaglianza di varianza :
$X_1 \dots X_n \qquad Y_1 \dots Y_n \quad H_0 : \sigma_x^2 = \sigma_y^2 \qquad H_a : \sigma_x^2 \neq \sigma_y^2$

$S_x^2 - S_y^2 \quad Ts = \frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}} = \frac{S_x^2}{S_y^2}$

$\frac{S_x^2}{S_y^2} \sim F_{n-1,m-1} \qquad Pr\{F_{1-\frac{\alpha}{2},n-1,m-1} \leq \frac{S_x^2}{S_y^2} \leq -F_{1-\frac{\alpha}{2},n-1,m-1}\}$
Non rifiuto se soddisfa la disuguaglianza

Test parametro Bernoulli (Var discrete.)
$H_0 : p \leq p_0 \quad H_a i p > p_0$

$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ *n campioni (Bernoulli)*

Binomiale $\sim$ Gaussiana (quando n è grande)
X n eventi favorevoli
$\mathbb{E}[X] = np \qquad Var(X) = np(1-p) \quad \mathcal{N}(np, np(1-p))$

Esempio Difetti di fabbricazione:
n = 300 $H_o ip \leq p_0 \quad p_0 = 2\%$
X = 10 n difetti

$$\frac{X - np}{\sqrt{nP_0(1-p_0)}} = \frac{10 - 300 \cdot 0.02}{\sqrt{300 \cdot 0.02 \cdot 0.98}} = 1.65$$

$$Pr\{z > 1.65\} = 0.0495$$

7.5 Modelli previsionali

7.5.1 Modelli di regressione previsionale

$Y_i = \alpha + \beta x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma^2)$

Problema $\{x_i, y_i\}_{i=1}^n \quad \alpha, \beta =?$

Sum of square -> SS
$SS = \sum_i^n (y_i - \alpha + \beta x_i)^2$ *Dove B e A -> var aleatoria*

$$\begin{cases} \dfrac{dSS}{dA} = -2\displaystyle\sum_i^n (y_i - A - Bx_i) = 0 \\ \dfrac{dSS}{dB} = -2\displaystyle\sum_i^n (y_i - A - Bx_i)^2 x_i = 0 \end{cases}$$

$$\begin{cases} \sum_i^n y_i = nA + B\sum_i x_i \\ \sum_i^n x_i y_i = n\sum_i^n + B\sum_i^n x_i \end{cases}$$

$$A = \frac{1}{n}\sum_i y_i - B\frac{1}{n}\sum x_i = \overline{y} - \beta\overline{x}$$

7.5.2 Regressione lineare

$y = \alpha + \beta x \quad e \sim (0, 1) \quad y_i = A + \beta x$

$\mathbb{E}[B] = \beta \qquad \mathbb{E}[A] = \alpha$
$Var[B] = \frac{\sigma^2}{\sum_i x_i^2 - n\overline{x}} \qquad Var[A] = \frac{\sigma^2 \sum x_i^2}{n(\sum x_i^2 - n\overline{x}^2)}$

$SS_R = \sum_i (y_i - (A + Bx_i))^2$ (Somma dei quadrati dei residui)

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

$\mathbb{E}[\frac{SS_R}{\sigma^2}] = n - 2 \qquad \mathbb{E}[\frac{SS_R}{n-2}] = \sigma^2$

MLE :

$$f_{y_1 \dots y_n}(y_1 \dots y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\sum(y-(\alpha+\beta x_i 0))^2/2\sigma^2}$$

MSE = MLE

Notazione :

$$
\begin{aligned}
S_{xy} &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \ldots = \sum_i x_i y_i - n\bar{x}\bar{y} \\
S_{xx} &= \sum_i (x_i - \bar{x})^2 = \ldots = \sum_i x_i^2 - n\bar{x} \\
S_{yy} &= \sum_i (y_i - \bar{y})^2 = \cdots = \sum_i y_i^2 - n\bar{y}
\end{aligned}
\tag{5}
$$

S_{xy} (Dispersione di x e y) S_{xy} (Dispersione di x) S_{xy} (Dispersione di y)

$A = \overline{y} - B\overline{x}$ $B = \frac{S_{xy}}{S_{xx}}$ $SS_R = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}$

Inferenza su β $\frac{B-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim \mathcal{N}(0,1)$ $\frac{SS_R}{\sigma^2} \sim \mathcal{X}_{n-2}^2$

$$
\frac{\frac{B-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{-\sqrt{\frac{SS_R}{j^2(n-2)}}} \sim t_{2-2}
$$

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

$\beta \in B \pm \sqrt{\frac{SS_R}{(n-2)S_{xx}}} \quad t_{\frac{\alpha}{2},n-2} \rightarrow$ Livello di confidenza

Inferenza su α $\quad \frac{A-\alpha}{\sqrt{\frac{\sigma^2 \sum x_i^2}{nS_{xx}}}} \sim \mathcal{N}(0,1) \qquad \frac{SS_R}{\sigma^2} \sim \mathcal{X}^2_{n-2}$
coeffieciente della retta:

$\alpha \in A \pm \frac{SS_R \sum x_i^2}{\sqrt{n(n-2)S_{xx}}} \sim t_{\frac{\alpha}{2},n-2} \rightarrow$ Livello di confidenza

Interferenza su $\alpha + \beta x_0 \quad :$

$\mathbb{E}[A + Bx_0] = \mathbb{E}[A] + x_0\mathbb{E}[B] = \alpha + \beta x_0$
$Var(A + Bx_0) = \ldots = \sigma^2[\frac{1}{n} + \frac{(\overline{x}-x_0)^2}{S_{xx}}]$

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

$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{(\frac{1}{n} + \frac{(x_0-x)^2}{S_{xx}}(\frac{SS_R}{n-2}))}} \sim t_{n-2}$$

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

Piccolo se i punti sono vicini alla media

7.5.3 Regressione Lineare (e non)

$\{x_i, y_i\}_{i=1}^2 \leftarrow$ punti stocastici

Inferenza $\alpha + \beta x_0 = \mathbb{E}[y] \rightarrow$ non so niente del valore della y in quel punto

Inferenza $y_0 = y(x_0)\theta$

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

$\alpha + \beta \mathbf{x_0} \rightarrow$ Il punto x_0 che sta sulla retta $\alpha + \beta x_0$

Inferenza $y_0 = y(x_0) \rightarrow$ predittivo

$$y \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2)$$
$$A + Bx_0 \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}))$$
$$y_0 - (A + Bx_0) \sim \mathcal{N}(\sigma, \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}))$$
$$y_0 = y(x_0) = A + Bx_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}})\frac{SS_R}{n-2}}$$

Coefficiente di determinazione

Definizione: La verifica dei miei valori

Formula generica: $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \rightarrow$ dispersione di y
La dispersione è data da due fattori:

- Retta (regressione)

- Rumore

$SS_R = \sum_{i=1}^{n}(y_i - (A + Bx_i))^2 \rightarrow$ Dipende dalla porzione non spiegata della retta
Utilizzo coefficienti di determinazione:

$$R^2 = \frac{S_{yy} - SS_R}{S_{yy}} = 1 - \frac{SS_R}{S_{yy}} \quad 0 \leq R^2 \leq 1$$

Se $\mathbf{R^2 = 1}$ la dispersione è data solo dalla retta *(regressione)*
Se $\mathbf{R^2 = 0}$ la dispersione è data solo dal *rumore*

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

Coefficiente di correlazione

Formula generica:

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

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \ldots = 1 - \frac{SS_R}{S_{yy}} \rightarrow \text{Dimostrazione matematica di } R^2$$

Analisi dei residui $y - (A + Bx) \rightarrow$ verifico tutti gli errori residui
Per la non linearità

Trasformazione lineare

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

Dove ce e $-dt$ sono *parametri*

$\log(W(t)) = \log(c) - dt \rightarrow$ Prob. soluzione al non lineare $y = \alpha + \beta x$

Rimedio al caso eteroschedastico :
$y_i = \alpha + \beta x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma_i^2) \rightarrow$ errore in crescita x
$Var(e_i) = \dfrac{\sigma^2}{W_i} \sum W_i(y - (A + Bx_0))^2$

- Regressione lineare multipla
  - $\mathbf{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_k x_k + e}$
  - $min \sum_i (y_i - (B_0 + B_1 x_{i1} + B_2 x_{i2} + \ldots + B_k x_{ik}))^2$
- Regressione (lineare) polinomiale
  - $y = \beta_0 = \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + e$
  - $\{\underline{x_i}, y_i\}_{i=1}^n$

$$\frac{\mathrm{d}}{\mathrm{d}B_0} = 0 = \sum_i (y_i - 1 - B_1 x_{i1} + B_2 x_{i2} + \ldots + B_k x_{ik})$$

$$\frac{\mathrm{d}}{\mathrm{d}B_1} = 0 = \sum_i x_{i1}(y_i - B_0 - B_1 x_{i1} + B_2 x_{i2} + \ldots + B_k x_{ik})$$

$x^t x \underline{\beta} = x^t \underline{y} \Longrightarrow \underline{\beta} = (x^x x)^{-1} x^t \underline{y}$

AN.O.VA (analysis of variance)
Analisi delle varianze / estensione del test di ipotesi sulle medie

Esempio voti medi degli anni scolastici

Anno.	**Voti medi.**
2020-2021 lockdown	μ_a
2021-2022 lockdown parziale	μ_b
2022-2023 presenza	μ_c

$H_0 : \mu_a = \mu_b = \mu_c$

1) stimatore di σ^2:

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

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

$\mathbb{E}[\frac{SS_w}{\sigma^2}] = n \cdot m - m$

$\mathbb{E}[\frac{SS_w}{nm-m}] = \sigma^2$ stimatore 1

2) stimatore di σ^2 supponendo $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_m = \mu$

$$n \sum_{i=1}^{m} \frac{(x_{i*} - \mu)^2}{\sigma^2} \sim \mathcal{X}_m^2 \qquad x_{**} = \frac{\sum_{i=1}^{m} \sum_{i=1}^{n} x_{ij}}{m \cdot n}$$

$SS_b = n \sum_{i=1}^{m} (x_{i*} - x_{**})^2 \sim \mathcal{X}_{m-1}^2$
$\mathbb{E}[\frac{SS_b}{m-1}] = \sigma^2 \rightarrow$ Stimatore 2

Verifico stimatori :

$$Ts = \frac{SS_b/m-1}{SS_W/nm-m} \rightarrow \text{intorno a 1 va bene}$$

F Distribution: $F_{m-1}, mn - m, \alpha$

ANOVA :
Se i gruppi sono uguali : n camp = $n \cdot m$
Se sono diversi : n camp = $\sum_i n_i$

Life testing (Misura di affidabilità)
$x \geq 0|$ *tempo di vita* $\qquad \lambda(t) = \frac{f(t)}{1-F(t)}$
$f(t) =$ Densità di popolazione
$\lambda(t) =$ Intensità di rottura (failure rate)

$$\begin{aligned} P(x \in (t, t + \triangle t) | x > t) &= \frac{P(x \in (t, t + \triangle t), x > t)}{P(x > t)} \\ &= \frac{P(x \in (t, t + \triangle t))}{P(x > t)} \\ &\approx \frac{F(t) \triangle t}{1 - F(t)} \end{aligned}$$

Intensità di rottura

Definizione: Densità condizionale di probabilità che un oggetto funzionante almeno fino a t si guasti "subito dopo"

Formula generica:

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

Se la distribuzione è esponenziale:

$$\lambda(t) = \frac{\lambda e^{-\lambda t}}{1 - 1 + e^{-\lambda t}} = \lambda \rightarrow \text{dove } \lambda \text{ è una costante}$$

Proprietà $\lambda(t) \Rightarrow F(t)$

$$\lambda(s) = \frac{f(s)}{1 - F(s)} = \frac{F'(s)}{1 - F(s)} = \frac{d}{dS}[-\log(1 - F(s))]$$

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

Esempio Tasso di mortalità di un fumatore ($\lambda_{\mathbf{s}}$) e di un <u>non</u> fumatore ($\lambda_{\mathbf{n}}$)

$$\lambda_s(t) = 2\lambda_n(t)$$

$$= P(\text{Non fumatore di età } \mathbf{A} \text{ vive fino a } \mathbf{B})$$
$$= P(\text{Non fumatore vive fino a } \mathbf{B} \mid \text{è vissuto fino } \mathbf{A})$$
$$= \frac{P(\text{Non fumatore viva fino a } \mathbf{B})}{P(\text{Non fumatore viva fino a } \mathbf{A})}$$
$$= \frac{1 - F_N(B)}{1 - F_N(A)}$$
$$= \frac{e^{-\int_0^B \lambda(t)\,dt}}{e^{-\int_0^A \lambda(t)\,dt}}$$

Quindi:
$P(\text{Non fumatore di eta } \mathbf{A} \text{ vive fino a } \mathbf{B}) = e^{-\int_A^B \lambda(t)\,dt}$

Per i non fumatori invece:
$P(\text{Fumatore di eta } \mathbf{A} \text{ vive fino a } \mathbf{B}) = e^{-\int_A^B \lambda(t)\,dt} = Ps$

Dove $Ps = (Pn)^2 \rightarrow$ quindi la probabilità di soppravivenza del fumatore è uguale alla probabilità di soppravivenza del non fumatore *al quadrato*

Probabilità che un non fumatore arrivi ai 60 anni sapendo che è arrivato ai 50:

$$\lambda_N(t) = \frac{1}{20} \qquad 50 \leq t \leq 60$$

$P_N = e^{-\int_{50}^{60} \frac{1}{20}\,dt} = e^{-\frac{1}{20}(60-50)} = e^{-\frac{1}{2}} \approx 0.607 \approx 61\%$
$P_{\leq}(e^{-\frac{1}{2}})^2 = e^{-1} \approx 0.368 \approx 37\%$

Stima di affidabilità N oggetti che si possono guastare *indipendenti* tra di loro
Tempi di vita: $\lambda e^{-\lambda t}$ $\quad \lambda = \frac{1}{\theta} \Rightarrow \frac{1}{\theta} e^{-\frac{t}{\theta}}$

Dati a disposizione: $x_1 \leq x_2 \leq x_3 \leq x_4 = r \quad i_1 = 2, i_2 = 3, i_3 = n, i_4 = 1$

Studio la variabile aleatoria X_i, i_j indica quale *oggetto si è guastato* per j-esimo all'istante x_j

(n-r) non si sono guastati $\Rightarrow$ per questi $X_i > x_r$

$$f x_1, x_2 \dots x_r(x_1, x_2, \dots, x_r) = \prod_j^r \frac{1}{\theta} e^{-\frac{x_j}{\theta}} = \frac{1}{\theta^r} e^{-\frac{\sum_j x_j}{\theta}}$$

Ora per i non guasti:

$$P\left(X_j > x_j \text{ con } j \notin \{i_1, \dots, i_r\}\right) = \prod_{r+1}^n (1 - F_{X_j}(x_r)) = \left[1 - (1 - e^{\frac{-x_r}{\theta}})\right]$$

$$\log L = -r \log\theta - \frac{1}{\theta}\left[\sum_i^r x_i + (n-r)x_r\right]$$

$$\frac{d\log}{d\theta} = -\frac{r}{\theta} + \frac{1}{\theta^2}\left[\sum_i^r x_i + (n-r)x_r\right] = 0$$

$$-\theta r + \left[\sum_i^r x_i + (n-r)x_r\right] = 0$$

$$\hat{\theta} = \frac{\sum_i^r x_i + (n-r)x_r}{r} = \frac{t}{r} = \frac{TTT}{r} \rightarrow \text{Total Time Test}$$