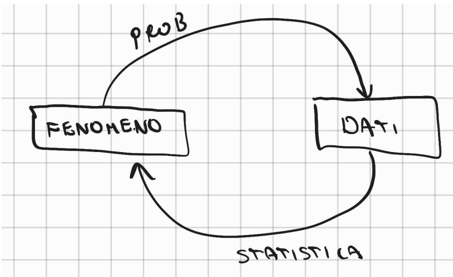
Contents

1 Introduzione

In probabilità quello che facciamo noi è quello di supporre che le nostre distribuzioni siano **note**, in statistica facciamo il contrario, ossia diciamo 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 $v\hat{a}r \to \hat{\boldsymbol{\theta}}$ (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 :

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

Dividiamo la *somma dei nostri valori* per il numero di dati

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

Sommiamo il numero più *piccolo* e il più *grande* e calcoliamo poi la **media** dei due

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

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 e servono per 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 e 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:

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

Derivando rispetto a p possiamo ottenere un'espressione per la stima $\boldsymbol{\hat{p}}$:

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

2.2 MLE di una Poisson

La funzione di *likelihood* è data da:

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

Derivando 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 $X_1, X_2, \dots, X_n$:

$$d(X_1, X_2, \dots, X_n) := \frac{1}{n}\sum_{i=1}^{n} X_i$$

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*

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

(capire sto risultato) Cosi otteniamo che la media della poissoniana è 2.7, la stima desiderata è data da:

$$P(X \leq 2) = \sum_{k=1}^{2} \frac{e^{-2.7} \cdot (2.7)^1}{k!}$$

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

2.3 MLE distribuzione Uniforme

Per la MLE delle uniformi dobbiamo trovare i valori di limite *inferiore* e *superiore* che massimizzano la probabilità di ottenere i dati osservati.

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

La formula per la stima di θ:

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

Lo stimatore di massima verosomiglianza della media della distribuzione è quindi

$$\hat{\theta}_{\frac{\text{MLE}}{2}} = \textit{media} = \frac{\max\{X_1, \dots, X_n\}}{2}$$

L'esempio (molto discutibile) sul libro si trova a pagina *255*

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:

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

La log-likelihood è data da:

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

La risoluzione ci porta alle seguenti formule per le stime:

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

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

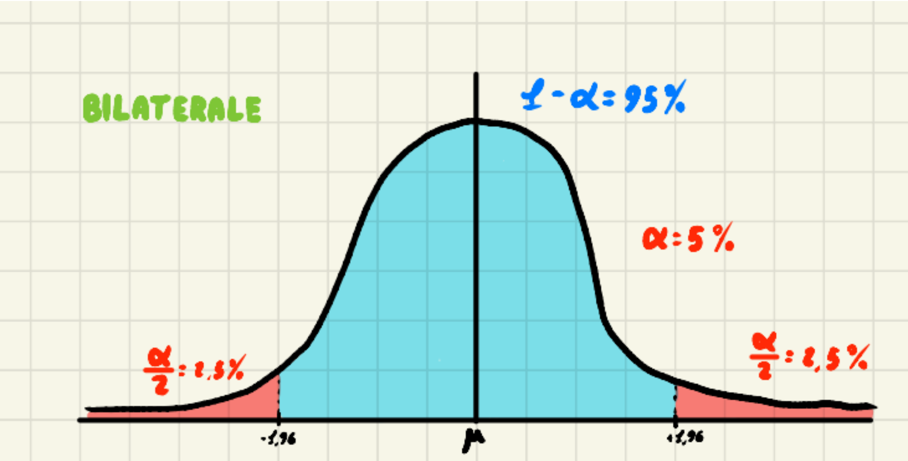
3 Teorema del limite centrale

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

Figure 1: Rappresentazione grafica



4.1 Tabella valori alpha

Figure 2: Valori comuni di α

VALORI COMUNI:

α	$1-\alpha$	$z_{\frac{\alpha}{2}}$
0,1	90%	1,645
0,05	95%	1,96
0,01	99%	2.58
0,005	99,5%	2.81

4.2 Distribuzione normale

4.2.1 μ incognita e varianza σ^2 nota

Definizione: 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:

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

alla fine otteniamo:

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

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 $\boldsymbol{\overline{X} = \overline{x}}$, allora possiamo dire che "con il 95% di confidenza" la media **vera** della distribuzione appartiene al seguente intervallo:

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

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

Esempio (Bilaterale) Messagio inviato da sorgente A a ricevente B con segnale elettrico con il valore di μ e la varianza $\sigma^2 = 4 \to \boldsymbol{\sigma = 2}$
i valori registrati sono i seguenti: { 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5 } $(n = 9)$

Risoluzione Otteniamo $\boldsymbol{\overline{x}}$ (sommando i valori e *dividendo* per la media):

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

Un intervallo di confidenza al 95% per μ è

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

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

Per caso Unilaterale noi in questo caso fino ad ora abbiamo visto degli intervalli di confidenza *bilaterali*, può capitare però di essere interessati solo ad un singolo valore che ci permette di affermare con il *95%* di confidenza che μ gli è superiore Le formule quindi per i due intervalli (sinistro e destro) sono i seguenti:

intervallo destro:

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

intervallo sinistro:

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

Esempio (Unilaterale) determiniamo al 95% di confidenza gli intervalli sinistro e destro del seguente valore:

$$1.645 \frac{\sigma}{\sqrt{n}} = \frac{3.29}{3} \approx 1.097$$

Intervallo destro:

$$(9 - 1.097, \infty) = (7.903, \infty)$$

Intervallo sinistro:

$$(-\infty, 9 + 1.097) = (-\infty, 10.097)$$

4.2.2 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:

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \longrightarrow \boldsymbol{\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}}$$

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

Per caso 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 caso Unilaterale

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

Esempio Usiamo i dati dell'esempio precedente: { 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5 } però senza conoscere σ ($\overline{x} = 9$)
Calcoliamo prima di tutto la *varianza campionaria*:

$$S^2 = \frac{\sum_i x_i^2 - 9\overline{x}^2}{8} = \mathbf{9.5}$$

$$S \approx 3.082$$

Dato che la nostra è una var. aleatoria di tipo t dobbiamo ricorrere alla sua tabella, troviamo che il valore $t_{0.025,8} \approx 2.306 \quad (t_{\alpha/2,n-1})$, quindi un intervallo di confidenza al 95% per μ è dato da:

$$9 \pm 2.306 \cdot \frac{3.092}{3} \Longrightarrow (6.63, \quad 11.37)$$

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 con media μ e varianza σ^2 entrambe *incognite*, dobbiamo prevedere l'elemento X_{n+1}
La sua distribuzione è:

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

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

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

Sostituiamo quindi σ nell'espressione sopra con la varianza campionaria, quindi:

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

Possiamo finalmente ottenere il nostro *intervallo di predizione*, siano $\alpha \in (0, 1/2)$:

$$1 - \alpha = P\left(\overline{X}_n - t_{\alpha/2,n-1}S_n\sqrt{1 + 1/n} < X_{n+1} < \overline{X}_n + t_{\alpha/2,n-1}S_n\sqrt{1 + 1/n}\right)$$

Se i valori che osserviamo per $\overline{X}_n, S_n$ sono $\overline{x}_n, s_n$, possiamo prevedere con un livello di confidenza $1 - \alpha$ che X_{n+1} cadrà nel seguente intervallo:

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

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 cerchiamo le statistiche del campione (X_{n+1}):

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

Dalle tabelle ricaviamo che $t_{0.025,6} \approx 2.447$, calcoliamo ora l'intervallo di predizione:

$$t_{\alpha/2,n-1} S_7 \sqrt{1 + \frac{1}{7}} \approx \mathbf{1919.97}$$

concludiamo col dire che il 95% di confidenza che X_8 cadrà nell'intervallo **[4796, 8637]**

6 Intervalli di confidenza per la varianza

Se $X_1, X_2, \ldots, X_n$ un campione di una distribuzione *normale* con parametri μ e σ^2 **incogniti**

Formula generica: costruiamo degli *intervalli di confidenza* per σ^2:

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

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 :

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

Esempio produzione di rondelle con spessori simili ma diversi: { 0.123, 0.133, 0.124, 0.125, 0.126, 0.128, 0.120, 0.124, 0.130, 0.126 }
Quale è l'intervallo di confidenza al 90% per la dev. standard dello spessore delle rondelle.

Risoluzione :

Un calcolo diretto mostra che $s^2 \approx 1.366 \times 10^{-5}$. Consultando la Tabella A.2 in Appendice, o eseguendo il Programma 5.8.1b si trova che $\chi^2_{0.05,9} \approx 16.917$ e $\chi^2_{0.95,9} \approx 3.334$, quindi

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \approx \frac{9 \times 1.366 \times 10^{-5}}{16.917} \approx 7.26 \times 10^{-6}$$
$$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \approx \frac{9 \times 1.366 \times 10^{-5}}{3.334} \approx 36.87 \times 10^{-6}$$

per cui

$$\sigma^2 \in (7.26 \times 10^{-6}, 36.87 \times 10^{-6})$$

con il 90% di confidenza, o equivalentemente, prendendo le radici quadrate,

$$\sigma \in (2.69 \times 10^{-3}, 6.07 \times 10^{-3})$$

sempre con il 90% di confidenza. □

Gli intervalli di confidenza unilaterali per σ^2 si ottengono in maniera del tutto analoga, e sono presentati nella Tabella 7.1, che riassume tutti i risultati di questa sezione.

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

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

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

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

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

Siano $X_1, X_2, \dots, X_n$ e $Y_1, Y_2, \dots, Y_m$ due campioni normali e differenti, denotiamo con μ_1 e σ_1^2 e con μ_2 e σ_2^2 le nostri variabili
$\boldsymbol{\overline{X} - \overline{Y}}$ 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 (dal fatto che la **somma** di normali indipendenti è ancora una aleatoria normale) 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 $\boldsymbol{\sigma_1^2}$ e $\boldsymbol{\sigma_2^2}$ 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 ulteriormente dedurre 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.

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

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

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

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

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

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

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

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

Definizione 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 e la verifica di questo è immediata:

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

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

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

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

7.2 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

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

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*

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

da cui otteniamo (se calcoliamo la derivata e ne studiamo il segno denotando con $\hat{\lambda}$ il valore di λ che produce il minimo):

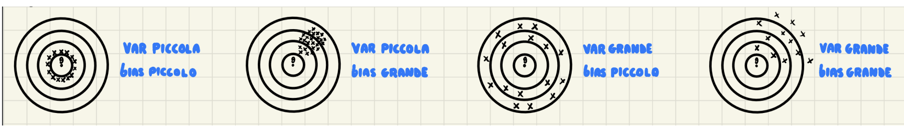
$$\boldsymbol{\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}}$$

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:

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

Bias/Polarizzazione Per uno stimatore *non distorto* l'errore quadratico medio coincide con la varianza, questa cosa si può generalizzare ad uno stimatore qualsiasi cambiando la formula e **sommando** il *quadrato dei bias*
Se $d(X)$ è **distorto**:

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



7.3 Stimatore della media di una distribuzione uniforme

Siano $X_1, X_2, \dots, X_n$ un campione estratto da una popolazione con distribuzione *uniforme* su $(0, \theta)$ dove θ è un parametro incognito.
Dato che $\mathbb{E}\left[\boldsymbol{X_i}\right] = \boldsymbol{\theta/2}$ è uno stimatore naturale per θ ed 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:

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

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 $\boldsymbol{r(d_2,\theta)}$:

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

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

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

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
Se le informazioni **a priori** assumono la forma di distribuzione di probabilità si dice che abbiamo **una distribuzione a priori per $\boldsymbol{\theta}$**

Se i valori che osserviamo sono $X_i = x_i$ e $i = 1, 2, \ldots, n$
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** della distribuzione a posteriori 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 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*:

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

e successivamente la *varianza*:

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

8.2 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* su un intervallo (a, b)
La sua densità a posteriori di θ è data da:

$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n|\theta)}{\int_a^b f(x_1, x_2, \ldots, x_n|\theta')d\theta'}$$

Questa formula è uguale allo stimatore di massima verosomiglianza

Esempio: Abbiamo una trasmissione da una sorgente A e un ricevente B con distribuzione $\mathcal{N}(s, 60)$ (s è il valore del segnale)
Considerando il rumore consideriamo *a priori* $\mathcal{N}(50, 100)$, si determini un intervallo che contenga il valore inviato col 90% di probabilità nel caso in cui il valore ricevuto sia **40**

Risoluzione Sapendo che abbiamo ricevuto un valore di 40 calcoliamo la media e la varianza:

$$\mathbb{E}\left[S|\textit{dati}\right] = \frac{1/60}{1/60 + 1/100} \cdot 40 + \frac{1/100}{1/60 + 1/100} \cdot 50 = \mathbf{43.75}$$

$$Var(S|\textit{dati}) = \frac{1}{1/60 + 1/100} = \mathbf{37.5}$$

Quindi:

$$0.90 = P(43.75 - 1.645\sqrt{37.5} < S < 43.75 + 1.645\sqrt{37.5}|\textit{dati})$$

Con probabilità 0.90 il segnale inviato appartiene all'intervallo **(33.68, 53.82)**

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 questo parametro.
Se F_θ è una distribuzione normale con media θ e varianza 1 due possibili ipotesi nulle sono:

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

2. $H_0 : \theta \leq n \longrightarrow$ *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*.

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

e

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

Esistono **due tipi di errori** che si commettono:

1. **Prima Specie**: Si rifiuta l'ipotesi di H_0 anche se è corretta

2. **Seconda Specie**: si accetta l'ipotesi di H_0 anche se è falsa

Esiste un livello di tolleranza specificato all'interno della regione critica per cui un'ipotesi può essere *ancora accettata*, mentre per rifiutarla occorre che i dati siano molto *improbabili* quando H_0 è soddisfatta. Questa tolleranza è definita dal **livello di significatività**, ovvero viene definito α e imponendo che il test abbia la *proprietà* che quando l'ipotesi H_0 è **vera**, la probabilità di rifiutarla non superi α

In poche parole: un test con livello di significatività α deve avere una probabilità di errore di *prima specie* minore o uguale ad α.

9.2 Verifica di ipotesi sulla media di una popolazione normale

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

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

Dove μ_0 è una costante che abbiamo fissato

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 α 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 Un segnale (μ) viene trasmesso da una sorgente A e ricevuto dal ricevente B, con rumore normale di media *nulla* e varianza *4*, quindi la sua distribuzione sarà: $\boldsymbol{\mathcal{N}(\mu, 4)}$
Il segnale viene inviato per *5* volte e supponiamo che la media campionaria dei *5* segnali ricevuti sia *8.5* ($\overline{X}$), mentre B aveva motivo di suppore che il valore inviato fosse *8*.

Risoluzione:

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

Dato che:

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

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 $\boldsymbol{\alpha < 0.576}$
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**.

9.2.1 Quando la varianza non è nota: il test t

Fino ad ora abbiamo supposto che l'unico parametro incognito fosse la *media*, in questo caso anche 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 $\mu \neq \mu_0$

$$H_0 : \mu = \mu_0$$

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

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:

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

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

Possiamo anche invertire il segno ponendo un - prima del tan:

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

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

Vedere tabella sotto per tutt'e cose

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

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

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

Figure 4: $X_1, X_2, \ldots, X_n$ 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 $P(T_{n-1} > t_{\alpha,n-1}) = \alpha$.

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

Risoluzione: 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:

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

il valore che abbiamo trovato è minore di $t_{0.05,19} \approx 1.729$ l'ipotesi nulla è *accettata* ad un livello di significatività del *5%*
però, un calcolo *p-dei-dati* fornisce il seguente valore:

$$p - dei - dati \approx P(|T_{19}| > 0.778) = 2P(T_{19} > 0.778) \approx 0.446$$

L'ipotesi nulla viene **accettata** a qualsiasi livello di significatività

I dati **non** sono in disaccordo che il consumo medio per abitazione sia di *350* galloni al giorno

9.3 Test unilaterali

Fino ad ora abbiamo verificato l'ipotesi nulla di $\mu = \mu_0$, però cosa succede quando μ può essere **solo** maggiore a μ_0?
Verifichiamo queste 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 critica* ***unilaterale*** o a una coda, quindi il problema di verificare le ipotesi alternative L'ipotesi *unilaterale*

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

contro l'alternativa

$$H_1 : \mu > \mu_0$$

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 α, 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 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.*

9.4.1 Il caso in cui le varianze sono note

Supponiamo che $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ siano 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

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

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:

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

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

9.4.2 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:

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

Quello che facciamo noi è di rifiutare H_0 quando $\overline{X} - \overline{Y}$ è lontano da zero, per capirlo 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 $\boldsymbol{n+m-2}$ 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 v il valore assunto da T

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

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 valori elevati di T, il test di significatività α è:

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

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

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 $\overline{Y} = 7.125$

Risoluzione: Calcoliamo le due S:

$$S_x^2 \approx 0.581 \qquad S_y^2 \approx 0.778$$

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:

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

9.4.3 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	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_x^2/n+S_y^2/m}}$	$\ldots \lvert D_{ts}\rvert > z_{\frac{\alpha}{2}}$	$2P(Z > \lvert t\rvert)$

9.5 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 $i = 1, 2, \ldots, n$
$X_1, X_2, \ldots, X_n \qquad Y_1, Y_2, \ldots, Y_m \qquad$ 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 $i = 1, 2, \ldots, n$ possiamo verificare queste due ipotesi

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

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

9.6 Verifica di ipotesi sulla varianza di una popolazione normale

Sia $X_1, X_2, \ldots, X_n$ 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

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

si rifiuta H_0 negli altri casi

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

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

9.7 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:

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

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

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% che le due varianze siano *uguali* ?
Calcoliamo la funzione di ripartizione delle *distribuzioni F*, quindi:

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

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.**

10 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:

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

10.1 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*

10.1.1 Metodo dei minimi quadrati

Definizione: Il metodo dei minimi quadrati consiste nello scegliere come stimatori di α e β i due valori A e B che **massimizzano** SS 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*.

11 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 (ossia gli stimatori dei minimi quadrati) 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:

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

In sintesi:

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

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

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

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:

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

Ed ha la seguente distribuzione:

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

Esempio Vengono misurati i consumi di un automobile a diverse velocità tra le *45* e le *70* miglia orarie, con i seguenti risultati:

Velocità	45	50	55	60	65	70	75
Miglia con un gallone	24.2	25.0	23.3	22.0	21.5	20.6	19.8

Risoluzione vediamo le due ipotesi:

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

Calcoliamo ora S_{xx}, S_{YY}, S_{xY}:

$$S_{xx} = 700 \qquad S_{YY} \approx 21.757 \qquad S_{xY} = -119$$

Il valore di SS_R viene calcolato in questo modo:

$$SS_R \approx \frac{700X21.757 - 119^2}{700} \approx \mathbf{1.527}$$

Per B troviamo:

$$B = S_{xY}/S_{xx} = -119/700 = 0.17$$

Infine per trovare il valore della statistica:

$$|-0.17|\sqrt{5X700/1.527} \approx \mathbf{8.139}$$

Ricavando dalle tabelle troviamo che $t_{0.005,5} \approx 4.032$ quindi l'ipotesi va **rifiutata** all' *1% di significatività*

12 Inferenza sui parametri della regressione

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

12.1 Inferenza su β

Un'ipotesi molto importante da verificare riguardo il *modello di regressione lineare semplice*

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

è che l'ipotesi β sia **pari a 0**
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.

Esempio in relazione all'esempio di prima si calcoli l'intervallo di confidenza al 95% per il parametro β

Risoluzione: Dato che $t_{0.025,5} \approx 2.571$ l'intervallo cercato è dato da:

$$-0.170 \pm 2.571\sqrt{\frac{1.527}{3500}} \approx \mathbf{-0.170 \pm 0.054}$$

Otteniamo che β sia compreso fra -0.224 e -0.116

12.2 Inferenza su α

Per determinare gli intervalli di confidenza che riguardano α si ottengono in maniera analoga a quanto fatto per β, quindi: 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.

12.3 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 invece la distribuzione di $A + Bx_0$?

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

12.3.1 Intervalli di confidenza

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

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

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

12.4 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	$\sqrt{\frac{n(n-2)S_{xx}}{SS_{\mathrm{R}} \cdot \sum_i x_i^2}}(A - \alpha) \sim t_{n-2}$
$\alpha + \beta x_0$	$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_{\mathrm{R}}}{n-2}}} \sim t_{n-2}$
$Y(x_0)$	$\frac{Y - A - Bx_0}{\sqrt{\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}}\sqrt{\frac{SS_{\mathrm{R}}}{n-2}}} \sim t_{n-2}$

12.5 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:

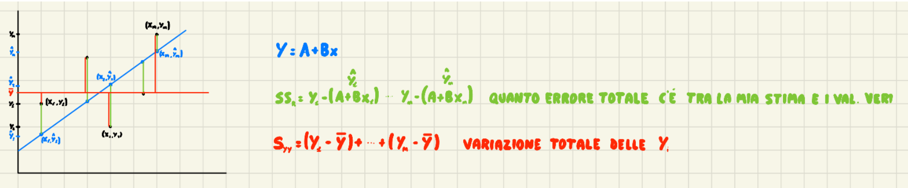
1. Se $R^2 = 1$:

   (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.



12.6 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:

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

12.7 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.

12.8 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:

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

12.9 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 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* i quali possono essere stimati
Costruiamo le loro *combinazioni lineari* con gli stimatori dei parametri e vengano determinati gli stimatori $\hat{\mu}_i$ per le medie vere μ_i
In questa ipotesi possiamo dimostrare ciò:

$$\sum_{i=1}^{N} \frac{(X_i - \hat{\mu}_i)^2}{\sigma^2} \sim \mathcal{X}_{N-k}^2 \Longrightarrow \sum_{i=1}^{N} \frac{(X_i - \mathbb{E}\left[X_i\right])^2}{\sigma^2} \sim \mathcal{X}_N^2$$

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:

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

13.1 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 $X_{ij} \quad i = 1, \ldots, m$ con quello che indica il campione mentre con $j = 1, \ldots, n$ indichiamo la posizione all'interno del campione stesso

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

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \quad \text{contro} \quad H_1 : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_m$$

Dato che ci sono $\boldsymbol{nm}$ *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 $\boldsymbol{SS_W}$ e 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.

13.1.1 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	$SS_{\text{b}} := n\sum_i (X_{i*} - X_{**})^2$	$m - 1$
Entro i campioni	$SS_{\text{W}} := \sum_i \sum_j (X_{ij} - X_{i*})^2$	$nm - m$

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

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:

$$\begin{aligned}\lambda(s) &:= \frac{f(s)}{1-F(s)} \\ &= \frac{F'(s)}{1-F(s)} \\ &= \boldsymbol{-\frac{d}{ds}\log(1-F(s))}\end{aligned} \tag{4}$$

Integrando con i membri tra 0 e t otteniamo alla fine che:

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

Questo sta a significare che 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

$$\tau = \sum_{j=1}^{r} Y_j$$

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 è di 525 ore (TTT). 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 $\hat{\theta}$:

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

Figure 5: Tabella di Φ

Tabella A.1 Funzione di ripartizione della distribuzione normale standard

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Esempio: per trovare un valore se dobbiamo trovare $\mathbf{\Phi(1.77)}$ cerco:

1.7 nelle *righe*

0.07 nelle *colonne*
$= 0.9616$

Figure 6: Tabella di $\chi^2_{\alpha,n}$

Tabella A.2 Valori assunti da $\chi^2_{\alpha,n}$

	α							
n	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.00098	0.00393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672

Dove: $n =$ *gradi di libertà*

se abbiamo

$\boldsymbol{n = 10}$

$\boldsymbol{\alpha = 0.05}$

cerco:

10 nelle *righe*

0.05 nelle *colonne*

Trovo subito che $\chi^2 = 18.307$

Figure 7: Tabella di $T_{\alpha,n}$

Tabella A.3 Valori assunti da $t_{\alpha,n}$

	α				
n	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
70	1.294	1.667	1.994	2.381	2.648
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.326	2.576

Esempio: per trovare un valore se dobbiamo trovare $\boldsymbol{T_{\alpha,n} \to (T_{0.025,10})}$ cerco:

10 nelle *righe*

0.025 nelle *colonne*
= 2.228

Figure 8: Tabella di $F_{0.05,n,m}$

Tabella A.4 Valori assunti da $F_{0.05,n,m}$; n rappresenta i gradi di libertà al numeratore e m quelli al denominatore.

	n						
m	1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01

[1]

[1]Padre nostro, che sei nei cieli, sia santificato il tuo nome, venga il tuo regno, sia fatta la tua volontà, come in cielo così in terra. Dacci oggi il nostro pane quotidiano, e rimetti a noi i nostri debiti come anche noi li rimettiamo ai nostri debitori, e non abbandonarci alla tentazione, ma liberaci dal male.