Formulario

Analisi dei dati 2022/23

Integrazione per parti

- $\int f'(x)g(x)dx = f(x)g(x) - \int f(x)g'(x)dx$

Valore atteso e varianza

- $\mathrm{E}(aX + bY + c) = a\mathrm{E}(X) + b\mathrm{E}(Y) + c$
- $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$
- $\mathrm{Var}(aX + bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$

Distribuzioni

- Bernoulli $X \sim \mathrm{Ber}(p)$
  - $\Pr(X = x) = p^x(1-p)^{1-x}$, $x = 0, 1$
  - $\mathrm{E}(X) = p \quad \mathrm{Var}(X) = p(1-p)$
  - **R**: `dbinom(size = 1, prob = p)`
    R: `pbinom(q, size = 1, prob = p)`
    R: `qbinom(p, size = 1, prob = p)`
- Binomiale $X \sim \mathrm{Binom}(n, p)$
  - $\Pr(X = x) = \binom{n}{x} p^x(1-p)^{1-x}$, $x = 0, 1, \ldots, n$
  - $\mathrm{E}(X) = np \quad \mathrm{Var}(X) = np(1-p)$
  - **R**: `dbinom(x, size = n, prob = p)`
    R: `pbinom(q, size = n, prob = p)`
    R: `qbinom(p, size = n, prob = p)`
- Poisson $X \sim \mathrm{Poi}(\lambda)$
  - $\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, \ldots$
  - $\mathrm{E}(X) = \lambda \quad \mathrm{Var}(X) = \lambda$
  - **R**: `dpois(x, lambda)`
    R: `ppois(q, lambda)`
    R: `qpois(p, lambda)`
- Normale $X \sim \mathrm{N}(\mu, \sigma^2)$
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$, $x \in \mathbb{R}$
  - $\mathrm{E}(X) = \mu \quad \mathrm{Var}(X) = \sigma^2$
  - **R**: `dnorm(x, mean = mu, sd = sigma)`
    R: `pnorm(q, mean = mu, sd = sigma)`
    R: `qnorm(p, mean = mu, sd = sigma)`

- Uniforme $X \sim \mathrm{U}(a, b)$
  - $f(x) = \dfrac{1}{b-a}$, $x \in [a, b]$
  - $\mathrm{E}(X) = (a+b)/2 \quad \mathrm{Var}(X) = (b-a)^2/12$
  - **R**: `dunif(x, min = a, max = b)`
    R: `punif(q, min = a, max = b)`
    R: `qunif(p, min = a, max = b)`
- Esponenziale $X \sim \mathrm{Exp}(\lambda)$
  - $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$
  - $\mathrm{E}(X) = 1/\lambda \quad \mathrm{Var}(X) = 1/\lambda^2$
  - **R**: `dexp(x, rate = lambda)`
    R: `pexp(q, rate = lambda)`
    R: `qexp(p, rate = lambda)`

Momenti

siccome la stima somme le probabilità di xi al momento k, serve anche vedere quante volte su n, una certa probabilità compare

- Momenti semplici:
  - popolazione $\mu_k = \mathrm{E}(X^k)$
  - campione $M_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$

stima = mk = (xi ^ k) / n = 1/n * (xi ^ k + xi+1 ^ k ... + xn ^k) al momento k-esimo
i=1
e se + noto che x0 = -1 appare 18 volte allora hai:

mk = 1/n * ((-1)^k * 18 + xi ^ k + xi+1 ^ k ... + xn ^k)

- Momenti centrali:
  - popolazione $\mu'_k = \mathrm{E}(X-\mu)^k$
  - campione $M'_k = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$

mk' = 1/n *
Xn
i=1(xi x‾)
k

Principali momenti campionari

- Media campionaria: $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$
- Varianza campionaria: $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2\right)$
- Covarianza campionaria: $S_{xy} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\,\bar{X}\,\bar{Y}\right)$
- Correlazione campionaria: $R_{xy} = \dfrac{S_{xy}}{S_x\,S_y}$

Scarto interquartile

- Scarto interquantile: $IQR = Q_3 - Q_1$
- Valori anomali: osservazioni superiori a $\widehat{Q}_3 + 1.5\widehat{IQR}$ o inferiori a $\widehat{Q}_1 - 1.5\widehat{IQR}$

Teoremi limite

- Legge dei grandi numeri: $\bar{X} \xrightarrow{p} \mu$, per $n \to \infty$
- Teorema del limite centrale: $\bar{X}$ ha distribuzione limite $N(\mu, \sigma^2/n)$

Transformazioni

- $E\{g(X)\} \neq g\{E(X)\}$, l'uguaglianza vale se $g(\cdot)$ è una funzione lineare
- $X \xrightarrow{p} \theta$ allora $g(X) \xrightarrow{p} g(\theta)$, se $g(\cdot)$ è una funzione continua
- $X \xrightarrow{d} N(\mu, \sigma^2)$ allora $g(X) \xrightarrow{d} N(g(\mu), g'(\mu)^2\sigma^2)$, se $g'(\mu)$ esiste e non è nulla, ovvero:
  - $E\{g(X)\} \approx g(\mu)$
  - $\text{Var}\{g(X)\} \approx g'(\mu)^2\sigma^2$

Proprietà degli stimatori

- Distorsione: $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$
- Errore quadratico medio: $\text{MSE}(\hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$
- Se $\text{Bias}(\hat{\theta}) \to 0$ e $\text{Var}(\hat{\theta}) \to 0$, allora $\hat{\theta} \xrightarrow{p} \theta$

109 volte x = 1 -> P(x = 1) = Teta/2
49 volte x = 0 -> P(x = 0) = Teta - 1
103 volte x = -1 -> P(x = -1) = Teta/2
L(teta) = [(Teta/2)^109] + [(Teta - 1) ^ 49] + [(Teta/2)^103]

Stimatore di massima verosimiglianza

Dipende se la variabile di riferimento è discreta o meno:
- se è discreta produttoria di P(X = xi) -> probabilità
- se è continua produttoria di f(xi) -> densità
eventualmente è necessario, con un campione di n = x + y + z elementi, specificare quante volte un valore P(X = x), P(X = y), P(X = y) ; la produttoria sarà:

L(teta) = P(X = x) ^ xvolte + P(X = y) ^ yvolte + P(X = z) ^ zvolte oppure
L(teta) = P(X = x) ^ xvolte + P(X = y) ^ yvolte + P(X = z) ^ zvolte

- Verosimiglianza:
  - caso discreto $L(\theta) \propto \prod_{i=1}^{n} \Pr(X_i = x_i; \theta)$
  - caso continuo $L(\theta) \propto \prod_{i=1}^{n} f(x_i; \theta)$
  - log-verosimiglianza $\ell(\theta) = \log L(\theta)$

Funzione punteggio: l'(Teta) -> deriva per Teta e non per x
Equazione -> l'(Teta) = 0 -> risolvi in Teta
Verificare che coincida con lo stimatore di massima verosomiglianza:
è vero se e solo se l''(Teta) < 0

- Informazione osservata: $J(\theta) = -\dfrac{\partial^2 \ell(\theta)}{\partial \theta^2}$ = - l''(teta) è più semplice da calcolare siccome l''(teta) lo ottieni già dal calcolo della stima di massima verosomiglianza
- Informazione attesa: $I(\theta) = E\left\{-\dfrac{\partial^2 \ell(\theta)}{\partial \theta^2}\right\}$
- Errore standard: $\text{SE}(\hat{\theta}) \approx I(\theta)^{-1/2}$ oppure $\text{SE}(\hat{\theta}) \approx J(\theta)^{-1/2}$
- Distribuzione limite: $N\{\theta, I(\theta)^{-1}\}$ oppure $N\{\theta, J(\theta)^{-1}\}$

Calcolare una stima dell'errore standard significa calcolare l'errore standard approssimato tramite l'informazione osservata/attesa -> più veloce se fai con quella osservata (J(teta)) :
-> SE(teta) circa = 1/sqrt(J(teta))

Intervalli di confidenza

1. trovare uno stimatore Teta^ non distorto di Teta
2. controllare che lo stimatore si distribuisca come una variabile normale
3. calcolare l'errore standard dello stimatore SE(Teta^) -> = SD(Teta^) / sqrt(n) -> corretto ricordati assolutamente
4. calcolare il quantile z/2 : ossia il valore critico della statistica Z che coincide col livello di confidenza a/2:
L'esercizio chiede un'intervallo di livello 1-a = x% allora a = 1 - x% e poi ricavo a/2 e mi ricordo che ricavo z(a/2) con Pr(Z > z(a/2)) = a/2 = 1 - Pr(Z < za) --> Pr(Z < za) = a/2 - 1 -> z(a/2) non è altro che la somma riga e colonna della tabella dove c'è il valore a/2 - 1 teoricamente
5. calcolare gli estremi dell'intervallo Tetaˆ±z/2SE(Tetaˆ)

Intervalli basati sulla statistica Z

- caso generico: $\hat{\theta} \pm z_{\alpha/2}\widehat{\text{SE}}(\hat{\theta})$
  - $z_{\alpha/2}$ quantile normale standard di posizione $1 - \alpha/2$
  - **R**: `qnorm(1 - alpha / 2)` z[a/2] = qnorm(1 - alpha/2)

Diseguaglianza per trovare la dimensione campionaria necessaria per garantire una certa precisione allo stimatore:
z[a/2] * (/n) <= triangolo ottenendo: n >= [(z[a/2]*) /triangolo] ^ 2
dove triangolo è un valore detto margine di errore da garantire tramite una certa dimensione campionaria n

MEDIE N > 30

- media con varianza nota: $\bar{X} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
df = n - 1
- media con varianza ignota e dimensione campionaria grande: $\bar{X} \pm z_{\alpha/2}\dfrac{S}{\sqrt{n}}$
df = n - 1
- differenza di due medie con varianze note: $(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}$
df = ?
- differenza di due medie con varianze ignote e dimensioni campionarie grandi: $(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}$

DIMENSIONE CAMPIONARIA MINIMA PER MARGINE DI ERRORE

- dimensione campionaria per stimare la media con una data precisione: $n \geq \left(\dfrac{z_{\alpha/2}\sigma}{\Delta}\right)^2$

la precisione è il triangolo

PROPORZIONI

QUANDO USARE LA STATISTICA T (qt)-> SE N <= 30 E LA VARIANZA O LA SD E' IGNOTA -> altrimenti per grandi dimensioni di n usi la statistica z (qnorm)

- proporzione con dimensione campionaria grande: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- differenza di due proporzioni con dimensioni campionarie grandi: $\hat{p}_X - \hat{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}$

- dimensione campionaria per stimare una proporzione con una data precisione: $n \geq 0.25\left(\frac{z_{\alpha/2}}{\Delta}\right)^2$ la precisione è il triangolo di confidenza

  0.25 è sempre vero

Intervalli basati sulla statistica T

i df = gradi di libertà = degree freedom -> controllano la forma della distribuzione T e sono di quantità [n - 1]

Pr(T 6 ta/2) = 1 - a/2

- media con varianza ignota: $\bar{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$

  La varianza o^2 viene stimata come S^2 = (1/n 1) Sommatoria[i=1 -> n ; (Xi - Xmedia)^2]

  al crescere di n, l'intervallo basato su T, converge a quello della statistica Z

  – $t_{\alpha/2}$ quantile distribuzione T di Student con $n-1$ gradi di libertà di posizione $1-\alpha/2$
  – **R**: `qt(1 - alpha / 2, df = n - 1)`

- differenza di due medie con varianze ignote ma uguali: $(\bar{X}-\bar{Y}) \pm t_{\alpha/2}S_p\sqrt{\frac{1}{n}+\frac{1}{m}}$
  – $t_{\alpha/2}$ quantile distribuzione T di Student con $n+m-2$ gradi di libertà
  – varianza 'pooled' $S_p^2 = \frac{1}{n+m-2}\left\{\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{m}(Y_i-\bar{Y})^2\right\}$

- differenza di due medie con varianze ignote non uguali: $(\bar{X}-\bar{Y}) \pm t_{\alpha/2}\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}$
  – $t_{\alpha/2}$ quantile distribuzione T di Student con ν gradi di libertà
  – gradi di libertà (formula di Satterthwaite)

$$\nu = \frac{\left(\frac{S_X^2}{n}+\frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)}+\frac{S_Y^4}{m^2(m-1)}}$$

Intervalli basati sullo stimatore di massima verosimiglianza con dimensioni campionarie grandi

- $\hat{\theta} \pm z_{\alpha/2}\, I(\hat{\theta})^{-1/2}$ Il teta è da dividere per n
- $\hat{\theta} \pm z_{\alpha/2}\, J(\hat{\theta})^{-1/2}$ Teta/n +- za/2 * I or J

Verifica delle ipotesi

Statistiche test Z N > 30

errore del primo tipo: quando l'ipotesi nulla H0 è vera, ma la rifiuto (anche se è vera la rifiuto)
errore del secondo tipo: quando l'ipotesi null H0 è falsa, ma non la rifiuto (ossia la accetto anche se è falsa)

- caso generico $\{H_0 : \theta = \theta_0\} : Z = \frac{\hat{\theta}-\theta_0}{\mathrm{SE}(\hat{\theta})}$

MEDIA

- media con varianza nota $\{H_0 : \mu = \mu_0\}$: $Z = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$

LIVELLO DI SIGNIFICATIVITA' = la probabilità di commettere un errore del primo tipo (rifiuto H0 anche se è vera) indicato con alfa a

- media con varianza ignota e dimensione campionaria grande $\{H_0 : \mu = \mu_0\}$: $Z = \frac{\sqrt{n}(\bar{X}-\mu_0)}{S}$

PROPORZIONI N GRANDE > 30

- proporzione con dimensione campionaria grande $\{H_0 : p = p_0\}$: $Z = \frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}$

DIFFERENZA MEDIE N > 30

- differenza di due medie con varianze note $\{H_0 : \mu_X - \mu_Y = D\}$:

$$Z = \frac{\bar{X}-\bar{Y}-D}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}$$

Un test Z per l'ipotesi alternativa bilaterale
H0 : Teta = Teta0 contro HA : Teta != Teta0
non rifiuta l'ipotesi nulla al livello di significativita`
se e solo se l'intervallo di confidenza per basato sulla
statistica Z con livello di confidenza 1 a contiene teta0.
Quindi, possiamo risolvere problemi di verifica
d'ipotesi (con alternativa bilaterale) utilizzando gli
intervalli di confidenza

POTENZA DEL TEST: probabilità di rifiutare H0 quando HA è vera (HA = ipotesi alternativa).
Viene anche definita come la probabilità di NON commettere un errore del secondo tipo.
Di solito dipende da Teta perchè HA contiene diversi valori di Teta

- differenza di due medie con varianze ignote $\{H_0 : \mu_X - \mu_Y = D\}$:

$$Z = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

pooled, quando si usa?
per risultati più accurati
quando si pone la differenza
di due medie o di due proporzioni
dove si considera uguale la varianza
dei 2 campioni. SERVE PER
MAGGIORE PRECISIONE NEL
TEST ->

per proporzione si intende x/totale e y/totale -> casifavorevoli/totale

DIFFERENZA PROPORZIONI N > 30

- differenza di due proporzioni con dimensioni campionarie grandi $\{H_0 : p_X - p_Y = D\}$:

p1hat <- 90 / 200
p2hat <- 145 / 250

  - se $D \neq 0$

$$Z = \frac{\hat{p}_X - \hat{p}_Y - D}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}}$$

Per pooled si intende in comune
pooled <- (90 + 145) / (450)
pooled

CASO POOLED
ossia basta che guardi
all'ipotesi nulla
per capire

  - se $D = 0$

$$Z = \frac{\hat{p}_X - \hat{p}_Y - D}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}, \quad \text{con} \quad \hat{p} = \frac{n\hat{p}_X + m\hat{p}_Y}{n+m}$$

p.pooled

p >= 0.1 non si può rifiutare H0 se p < 0.1 rifiuto H0

il p value dipende dall'ipotesi alternativa fatta

- livello di significatività osservato

se
hai già
il liv. di
significatività
usi
qnorm(1 - a/2)

  - alternativa bilaterale $p = 2\{1 - \Pr(Z \leq |z|)\}$
    R: `p = 2 * (1 - pnorm(abs(z)))`
    R: `p = 2 * pnorm(abs(z), lower.tail = FALSE)` (maggiore precisione numerica)
  - alternativa unilaterale destra $p = 1 - \Pr(Z \leq z)$
    R: `p = 1 - pnorm(z)`
    R: `p = pnorm(z, lower.tail = FALSE)` (maggiore precisione numerica)
  - alternativa unilaterale sinistra $p = \Pr(Z \leq z)$
    R: `p = pnorm(z)`

Statistiche test T   DIMENSIONE <= 30

- media con varianza ignota $\{H_0 : \mu = \mu_0\}$: $T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$
  - T distribuito come T di Student con $n-1$ gradi di libertà sotto H_0

- differenza di due medie con varianze ignote ma uguali $\{H_0 : \mu_X - \mu_Y = D\}$: $T = \frac{\bar{X} - \bar{Y} - D}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$

Sp^2 ossia pooled -> proporzioni o medie uguali

MEDIE
DI
CAMPIONI
CON DIMENSIONE
N <= 30

  - T distribuito come T di Student con $n+m-2$ gradi di libertà sotto H_0
  - $S_p^2 = \frac{1}{n+m-2}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2\right\}$

- differenza di due medie con varianze ignote non uguali $\{H_0 : \mu_X - \mu_Y = D\}$:

se non viene detto nulla riguardo alle
varianze campionarie, dobbiamo per forza
usare Satterwaite per calcolare i gradi di
libertà e il livello di significatività
approssimato p-value

$$T = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

usa questa se hai
le SDx e SDy
detto anche scarto
quadratico medio

sono delle stime
delle varianze (sd^2)

  - T approssimativamente distribuito come T di Student con con ν gradi di libertà sotto H_0
  - gradi di libertà (formula di Satterthwaite)

$$\nu = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

p >= 0.1 non si può rifiutare H0 se p < 0.1 rifiuto H0

LIVELLO DI SIGNIFICATIVITA' OSSERVATO detto anche p-value, evita di fissare un livello di significativita in modo da misurare ` l'ammontare di evidenza a favore delle ipotesi -> quindi usiamo quello per vedere l'evidenza a favore o contro l'ipotesi nulla -> se il p-value contiene (<=) del livello di significatività richiesto, vuoldire posso anche rifiutare eventualmente H0, mentre se il p-value non contiene (>) il livello di significatività richiesto, allora non posso rifiutare H0.

- livello di significatività osservato:
  - alternativa bilaterale $p = 2\{1 - \Pr(T \leq |t|)\}$
    R: `p = 2 * (1 - pt(abs(t), df = gradi.liberta))`
    R: `p = 2 * pt(abs(t), df = gradi.liberta, lower.tail = FALSE)` (maggiore precisione numerica)
  - alternativa unilaterale destra $p = 1 - \Pr(T \leq t)$
    R: `p = 1 - pt(t, df = gradi.liberta)`
    R: `p = pt(t, df = gradi.liberta, lower.tail = FALSE)` (maggiore precisione numerica)
  - alternativa unilaterale sinistra $p = \Pr(T \leq t)$
    R: `p = pt(t, df = gradi.liberta)`

Statistiche test basate sullo stimatore di massima verosimiglianza con dimensioni campionarie grandi $\{H_0 : \theta = \theta_0\}$:

- $Z = I(\theta_0)^{1/2}(\hat{\theta} - \theta_0)$
- $Z = J(\theta_0)^{1/2}(\hat{\theta} - \theta_0)$

Statistica test χ^2 $\{H_0 : O_{ij} = E_{ij}$, per ogni scelta di i e $j\}$:

- $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
  - χ^2 distribuito come variabile casuale χ^2 con $(k-1)(m-1)$ gradi di libertà
- tabelle contingenza
  le osservate sono Nij, ossia la tabella di contingenza
  Nij = elemento in posizione: riga i, colona j della tabella di contingenza
  - frequenze osservate $O_{ij} = n_{ij}$
    in posizione i,j metto il prodotto delle marginali / n
  - frequenze attese stimate $\widehat{E}_{ij} = \frac{n_i.n._j}{n}$
    Ni* = sommatoria di Nij con j = 0, .. , m
    N*j = sommatoria di Nij con i = 0, .. , k
    N = totale = sommatoria di Nij con j = 0, .. , m E con i = 0, .. , k
  - **R**:

v <- c(9, 18, .. , 10)
tab.oss <- as.table(matrix(v, nrow = 2))

`tabella <- as.table(matrix(frequenze.osservate, nrow = numero.righe))`
`margin1 <- margin.table(tabella, margin = 1)` $(n_i.)$ righe <- rowSums(tab.oss)
`margin2 <- margin.table(tabella, margin = 2)` $(n._j)$ colonne <- colSums(tab.oss)
`outer(margin1, margin2) / sum(tabella)` (frequenze attese stimate)
`summary(tabella)` (test χ^2 d'indipendenza)

il p-value indica il livello di significatività osservato, ossia la probabilità che l'ipotesi nulla sia vera -> se il p-value è grande allora posso non posso rifiutare H0

P-value	Significatività	Decisione
p < 0.001	forte	rifiuto netto di H0
0.001 <= p < 0.01	moderata	rifiuto di H0
0.01 <= p < 0.05	modesta	rifiuto debole di H0
0.05 <= p < 0.1	incerta	situazione dubbia
p >= 0.1	nessuna	non si può rifiutare H0

Regressione lineare

- Retta di regressione: $y_i = \beta_0 + \beta_1 x_i + \epsilon$ GUARDA LA TABELLA: ------>>>>
- Stime ai minimi quadrati: $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$ $\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2}$
- Residui: $e_i = y_i - \widehat{y}_i$
- Regressione e correlazione: $\widehat{\beta}_1 = r_{xy} \frac{s_y}{s_x}$
- Decomposizione della varianza: $SQ_{tot} = SQ_{reg} + SQ_{res}$
  - somma dei quadrati totale $SQ_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2$
  - somma dei quadrati spiegata $SQ_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
  - somma dei quadrati residua $SQ_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Coefficiente di determinazione $R^2 = \frac{SQ_{reg}}{SQ_{tot}}$
  - retta di regressione $R^2 = r_{xy}^2$
- Distribuzione limite $\widehat{\beta}_1$: $N\{\beta_1, \mathrm{var}(\beta_1)\}$
  - $\mathrm{var}(\widehat{\beta}_1) = \frac{\sigma^2}{(n-1)s_x^2}$

  - $\widehat{\text{var}}(\widehat{\beta}_1) = \dfrac{s_e^2}{(n-1)s_x^2}$
  - $s_e^2 = \dfrac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$
- Intervallo di confidenza per β_1: $\widehat{\beta}_1 \pm t_{\alpha/2}\dfrac{s_e}{s_x\sqrt{n-1}}$
- Test sul predittore $\{\mathrm{H}_0 : \beta_1 = \beta_1^0\}$: $T = \dfrac{s_x\sqrt{n-1}}{s_e}\left(\widehat{\beta}_1 - \beta_1^0\right)$
  - T distribuito come T di Student con $n-2$ gradi di libertà
- Previsione $\widehat{y}_p = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$
  - varianza stimata $\widehat{\text{Var}}(\widehat{y}_p) = s_e^2\left(1 + \dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{(n-1)s_x^2}\right)$
  - intervallo di previsione $\widehat{y}_p \pm t_{\alpha/2}\widehat{\text{Var}}(\widehat{y}_p)^{1/2}$
  - $t_{\alpha/2}$ quantile distribuzione T di Student con $n-2$ gradi di libertà di posizione $1-\alpha/2$