

Formulario

Analisi dei dati 2023/24

Integrazione per parti

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

Valore atteso e varianza

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

Distribuzioni

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

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

Momenti

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

Principali momenti campionari

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

Scarto interquartile

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

Teoremi limite

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

Trasformazioni

- $E\{g(X)\} \neq g\{E(X)\}$, 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 $\text{Bias}(\hat{\theta}) \rightarrow 0$ e $\text{Var}(\hat{\theta}) \rightarrow 0$, allora $\hat{\theta} \xrightarrow{p} \theta$

Stimatore di massima verosimiglianza

- Verosimiglianza:
 - caso discreto $L(\theta) \propto \prod_{i=1}^n \text{Pr}(X_i = x_i; \theta)$
 - caso continuo $L(\theta) \propto \prod_{i=1}^n f(x_i; \theta)$
 - log-verosimiglianza $\ell(\theta) = \log L(\theta)$
- Informazione osservata: $J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2}$
- Informazione attesa: $I(\theta) = E \left\{ -\frac{\partial^2 \ell(\theta)}{\partial \theta^2} \right\}$
- 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 $N\{\theta, J(\theta)^{-1}\}$

Intervalli di confidenza

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)`
- media con varianza nota: $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- media con varianza ignota e dimensione campionaria grande: $\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$
- differenza di due medie con varianze note: $(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$
- differenza di due medie con varianze ignote e dimensioni campionarie grandi: $(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$
- dimensione campionaria per stimare la media con una data precisione: $n \geq \left(\frac{z_{\alpha/2} \sigma}{\Delta} \right)^2$

- 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: $n \geq 0.25 \left(\frac{z_{\alpha/2}}{\Delta} \right)^2$

Intervalli basati sulla statistica T

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

- $\hat{\theta} \pm z_{\alpha/2} I(\hat{\theta})^{-1/2}$
- $\hat{\theta} \pm z_{\alpha/2} J(\hat{\theta})^{-1/2}$

Verifica delle ipotesi

Statistiche test Z

- caso generico $\{H_0 : \theta = \theta_0\} : Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$
- media con varianza nota $\{H_0 : \mu = \mu_0\} : Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$
- media con varianza ignota e dimensione campionaria grande $\{H_0 : \mu = \mu_0\} : Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$
- 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 di due medie con varianze note $\{H_0 : \mu_X - \mu_Y = D\} :$

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

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

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

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

- se $D \neq 0$

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

- se $D = 0$

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

- livello di significatività osservato

- 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

- 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 $\{H_0 : \mu_X - \mu_Y = D\}$: $T = \frac{\bar{X} - \bar{Y} - D}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$

- 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 $\{H_0 : \mu_X - \mu_Y = D\}$:

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

- T approssimativamente distribuito come T di Student con ν gradi di libertà sotto H_0

- gradi di libertà (formula di Satterthwaite)

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

- 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 $\chi^2 \{H_0 : O_{ij} = E_{ij}, \text{ per ogni scelta di } i \text{ e } j\}$:

- $\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
 - χ^2 distribuito come variabile casuale χ^2 con $(k-1)(m-1)$ gradi di libertà
- tabelle contingenza
 - frequenze osservate $O_{ij} = n_{ij}$
 - frequenze attese stimate $\hat{E}_{ij} = \frac{n_{i.}n_{.j}}{n}$
 - **R:**
`tabella <- as.table(matrix(frequenze.osservate, nrow = numero.righe))`
`margin1 <- margin.table(tabella, margin = 1) (ni.)`
`margin2 <- margin.table(tabella, margin = 2) (n.j)`
`outer(margin1, margin2) / sum(tabella)` (frequenze attese stimate)
`summary(tabella)` (test χ^2 d'indipendenza)

Regressione lineare

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

$$- \widehat{\text{var}}(\hat{\beta}_1) = \frac{s_e^2}{(n-1)s_x^2}$$

$$- s_e^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Intervallo di confidenza per β_1 : $\hat{\beta}_1 \pm t_{\alpha/2} \frac{s_e}{s_x \sqrt{n-1}}$
- Test sul predittore $\{H_0 : \beta_1 = \beta_1^0\}$: $T = \frac{s_x \sqrt{n-1}}{s_e} (\hat{\beta}_1 - \beta_1^0)$
 - T distribuito come T di Student con $n-2$ gradi di libertà
- Previsione $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$
 - varianza stimata $\widehat{\text{Var}}(\hat{y}_p) = s_e^2 \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{(n-1)s_x^2} \right)$
 - intervallo di previsione $\hat{y}_p \pm t_{\alpha/2} \widehat{\text{Var}}(\hat{y}_p)^{1/2}$
 - $t_{\alpha/2}$ quantile distribuzione T di Student con $n-2$ gradi di libertà di posizione $1 - \alpha/2$