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

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

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

- $\mathrm{E}\{g(X)\} \neq g\{\mathrm{E}(X)\}$, 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:
  - $\mathrm{E}\{g(X)\} \approx g(\mu)$
  - $\mathrm{Var}\{g(X)\} \approx g'(\mu)^2\sigma^2$

Proprietà degli stimatori

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

Stimatore di massima verosimiglianza

- 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 $\ell(\theta) = \log L(\theta)$
- Informazione osservata: $J(\theta) = -\dfrac{\partial^2 \ell(\theta)}{\partial \theta^2}$
- Informazione attesa: $I(\theta) = \mathrm{E}\left\{-\dfrac{\partial^2 \ell(\theta)}{\partial \theta^2}\right\}$
- Errore standard: $\mathrm{SE}(\hat{\theta}) \approx I(\theta)^{-1/2}$ oppure $\mathrm{SE}(\hat{\theta}) \approx J(\theta)^{-1/2}$
- Distribuzione limite: $N\{\theta, I(\theta)^{-1}\}$ oppure $N\{\theta, J(\theta)^{-1}\}$

Intervalli di confidenza

Intervalli basati sulla statistica Z

- caso generico: $\hat{\theta} \pm z_{\alpha/2}\widehat{\mathrm{SE}}(\hat{\theta})$
  - $z_{\alpha/2}$ quantile normale standard di posizione $1 - \alpha/2$
  - **R**: `qnorm(1 - alpha / 2)`
- media con varianza nota: $\bar{X} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
- media con varianza ignota e dimensione campionaria grande: $\bar{X} \pm z_{\alpha/2}\dfrac{S}{\sqrt{n}}$
- differenza di due medie con varianze note: $(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}$
- 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 per stimare la media con una data precisione: $n \geq \left(\dfrac{z_{\alpha/2}\,\sigma}{\Delta}\right)^2$

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

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

- dimensione campionaria per stimare una proporzione con una data precisione: $n \geq 0.25\left(\dfrac{z_{\alpha/2}}{\Delta}\right)^2$

Intervalli basati sulla statistica T

- media con varianza ignota: $\bar{X} \pm t_{\alpha/2}\,\dfrac{S}{\sqrt{n}}$
  - $t_{\alpha/2}$ 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: $(\bar{X}-\bar{Y}) \pm t_{\alpha/2} S_p \sqrt{\dfrac{1}{n}+\dfrac{1}{m}}$
  - $t_{\alpha/2}$ quantile distribuzione T di Student con $n+m-2$ gradi di libertà
  - varianza 'pooled' $S_p^2 = \dfrac{1}{n+m-2}\left\{\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{m}(Y_i-\bar{Y})^2\right\}$

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

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

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 $\{\mathrm{H}_0 : \theta = \theta_0\}$: $Z = \dfrac{\hat{\theta}-\theta_0}{\mathrm{SE}(\hat{\theta})}$

- media con varianza nota $\{\mathrm{H}_0 : \mu = \mu_0\}$: $Z = \dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$

- media con varianza ignota e dimensione campionaria grande $\{\mathrm{H}_0 : \mu = \mu_0\}$: $Z = \dfrac{\sqrt{n}(\bar{X}-\mu_0)}{S}$

- proporzione con dimensione campionaria grande $\{\mathrm{H}_0 : p = p_0\}$: $Z = \dfrac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}$

- differenza di due medie con varianze note $\{\mathrm{H}_0 : \mu_X - \mu_Y = D\}$:

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

- differenza di due medie con varianze ignote $\{\mathrm{H}_0 : \mu_X - \mu_Y = D\}$:

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

- differenza di due proporzioni con dimensioni campionarie grandi $\{\mathrm{H}_0 : p_X - p_Y = D\}$:
  - se $D \neq 0$

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

  - se $D = 0$

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

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

Statistiche test T

- media con varianza ignota $\{\mathrm{H}_0 : \mu = \mu_0\}$: $T = \dfrac{\sqrt{n}(\bar{X} - \mu_0)}{S}$
  - T distribuito come T di Student con $n-1$ gradi di libertà sotto H_0
- differenza di due medie con varianze ignote ma uguali $\{\mathrm{H}_0 : \mu_X - \mu_Y = D\}$: $T = \dfrac{\bar{X} - \bar{Y} - D}{S_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$
  - T distribuito come T di Student con $n+m-2$ gradi di libertà sotto H_0
  - $S_p^2 = \dfrac{1}{n+m-2}\left\{\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2\right\}$
- differenza di due medie con varianze ignote non uguali $\{\mathrm{H}_0 : \mu_X - \mu_Y = D\}$:

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

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

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

- 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 $\{\mathrm{H}_0 : \theta = \theta_0\}$:

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

Statistica test χ^2 $\{\mathrm{H}_0 : O_{ij} = E_{ij}$, 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
  - frequenze osservate $O_{ij} = n_{ij}$
  - frequenze attese stimate $\widehat{E}_{ij} = \frac{n_i.n._j}{n}$
  - **R**:
    `tabella <- as.table(matrix(frequenze.osservate, nrow = numero.righe))`
    `margin1 <- margin.table(tabella, margin = 1)` ($n_i.$)
    `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: $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$ $\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2}$
- Residui: $e_i = y_i - \widehat{y}_i$
- Regressione e correlazione: $\widehat{\beta}_1 = r_{xy}\frac{s_y}{s_x}$
- Decomposizione della varianza: $\mathrm{SQ}_{\mathrm{tot}} = \mathrm{SQ}_{\mathrm{reg}} + \mathrm{SQ}_{\mathrm{res}}$
  - somma dei quadrati totale $\mathrm{SQ}_{\mathrm{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)s_y^2$
  - somma dei quadrati spiegata $\mathrm{SQ}_{\mathrm{reg}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - somma dei quadrati residua $\mathrm{SQ}_{\mathrm{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Coefficiente di determinazione $R^2 = \frac{\mathrm{SQ}_{\mathrm{reg}}}{\mathrm{SQ}_{\mathrm{tot}}}$
  - 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 $\{\text{H}_0 : \beta_1 = \beta_1^0\}$: $T = \dfrac{s_x\sqrt{n-1}}{s_e}\left(\widehat{\beta}_1 - \beta_1^0\right)$
  - 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 $1-\alpha/2$