Università di Pisa

Dipartimento di Informatica
Corso di Laurea Triennale in Informatica

Formulario di Statistica

Professore:
Prof. Francesco Grotto

Autore:
Filippo Ghirardini

Anno Accademico 2023/2024

Contents

Formulario di Statistica

Realizzato da: Ghirardini Filippo

A.A. 2023-2024

---

1 Formulario

1.1 Statistica descrittiva

1.1.1 Indici statistici

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(Media campionaria)}$$

$$var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{(Varianza campionaria)}$$

$$\sigma(x) = \sqrt{var(x)} \qquad \text{(Deviazione standard)}$$

$$b = \frac{1}{\sigma^3}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \qquad \text{(Sample skewness)}$$

Proposizione 1.1.1.

$$\frac{\#\{x_i : |x_i - \bar{x}| > d\}}{n-1} \leq \frac{var(x)}{d^2} \tag{1}$$

1.1.2 Quantili

$$F_e(t) = \frac{\#\{i | x_i \leq t\}}{n} \qquad \text{(Funzione di ripartizione empirica)}$$

1.1.3 Dati multivariati

$$cov(x,y) = \sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad \text{(Covarianza campionaria)}$$

$$r(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{(Coefficiente di correlazione)}$$

1.2 Probabilità e indipendenza

1.2.1 Spazi di probabilità

$$\mathbb{P}\Big(\bigcup_{n=1}^{+\infty} A_n\Big) = \sum_{n=1}^{+\infty}\mathbb{P}(A_n) \qquad (\sigma\text{-addittività})$$

Proposizione 1.2.1. *Operazioni su spazi di probabilità:*

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\emptyset) = 0$

- $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$

Proposizione 1.2.2. *Vale:*

$$\mathbb{P}(A) = \lim_{n \to \infty} \mathbb{P}(A_n) \tag{2}$$

1.2.2 Probabilità discreta

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} \qquad \text{(Probabilità)}$$

Proposizione 1.2.3. *Calcolo combinatorio:*

- *Sequenze ordinate con possibile ripetizione di k numeri da* 1 *a n:* n^k
- *Numero di modi in cui si può ordinare* $\{1, \ldots, n\}$*:* $n!$
- *Numero di sequenze ordinate senza ripetizione di k numeri di* $\{1, \ldots, n\}$*:* $\frac{n!}{(n-k)!}$
- *Numero di sottoinsiemi di* $\{1, \ldots, n\}$ *formati da k elementi:* $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Definizione 1.2.1 (Funzione di massa)**.** *Se* $\Omega = \{x_1, x_2, \ldots\} \subset \mathbb{R}$ *è un sottoinsieme numerabile:*

$$\Omega \ni x_i \mapsto p(x_i) = \mathbb{P}(\{x_i\}) \in [0, 1] \tag{3}$$

e valgono:

$$\mathbb{P}(A) = \sum_{i : x_i \in A} p(x_i) \qquad \forall A \subseteq \mathbb{R} \tag{4}$$

$$p(x_i) \geq 0 \tag{5}$$

$$\sum_{i=1,2,\ldots} p(x_i) = 1 \tag{6}$$

1.2.3 Probabilità condizionata

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \text{(Probabilità condizionata)}$$

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \ldots \cdot \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1}) \qquad \text{(Condizionamento ripetuto)}$$

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i) \qquad \text{(Formula di fattorizzazione)}$$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \qquad \text{(Formula di Bayes)}$$

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{n} \mathbb{P}(A|B_j)\mathbb{P}(B_j)} \qquad \text{(Formula di Bayes - Sistema di alternative)}$$

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}(A_{i_k}) \qquad \text{(Eventi indipendenti)}$$

1.2.4 Entropia di Shannon

$$H^{(n)}(p_1,\dots,p_n) = -\sum_{i=1}^{n} p_i \log(p_i) \qquad \text{(Entropia)}$$

Proposizione 1.2.4. *L'entropia ha queste proprietà:*

- *È una funzione simmetrica*
- $H^{(n)}(1,0,\dots,0) = 0$
- *È coerente tra n diversi:* $H^{(n)}(p_1 = 0, p_2,\dots,p_n) = H^{(n-1)}(p_2,\dots,p_n)$
- $H^{(n)}(p_1,\dots,p_n) \le H^{(n)}(\frac{1}{n},\dots,\frac{1}{n})$
- *Data una probabilità su* $n \times m$ *oggetti,* $\Omega = \{x_{11},\dots,x_{ij},\dots,x_{nm}\}$, $\mathbb{P}(\{x_{ij}\}) = q_{ij}$, *considerando gli eventi* $A_i = \{x_{i,1},\dots,x_{i,m}\}$, $\mathbb{P}(A_i) = p_i$*:*

$$H^{(nm)}(q_{11},\dots,q_{ij},\dots,q_{nm}) = H^{(n)}(p_1,\dots,p_n) + \sum_{i=1}^{n} p_i H^{(m)}\left(\frac{q_{i1}}{p_i},\dots,\frac{q_{im}}{p_i}\right) \tag{7}$$

Teorema 1.2.1 (Teorema di Shannon)**.** Una funzione continua che soddisfa queste proprietà deve avere la forma:

$$cH^{(n)} \qquad c > 0 \tag{8}$$

1.2.5 Densità di probabilità

$$\mathbb{P}(A) = \int_A f(x)dx \qquad A \subseteq \Omega \qquad \text{(Probabilità di una densità)}$$

$$\mathbb{P}(A \cup B) = \int_{A\cup B} f(x)dx = \int_A f(x)dx + \int_B f(x)dx = \mathbb{P}(A) + \mathbb{P}(B) \qquad A \cap B = \emptyset$$

(Somma di probabilità)

$$\mathbb{P}(\{t\}) = \int_{\{t\}} f(x)dx = 0 \qquad \text{(Probabilità di un punto)}$$

$$f(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{altrove} \end{cases} \qquad \text{(Densità uniforme)}$$

$$\mathcal{X} : \mathbb{R} \to \{0,1\} \qquad \mathcal{X}_S(x) = \begin{cases} 1 & x \in S \\ 0 & x \notin S \end{cases} \qquad \text{(Funzione indicatrice)}$$

1.3 Variabili aleatorie

1.3.1 Legge di una variabile aleatoria

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \sum_{x_i \in A} p_X(x_i) \qquad \text{(Variabile aleatoria discreta)}$$

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \int_A f(x)dx \qquad \text{(Variabile aleatoria continua)}$$

$$\mathbb{P}(X \in A) = \mathbb{P}(a \le X \le b) = \int_a^b f(x)dx \qquad \text{(Variabile aleatoria continua con segmento)}$$

1.3.2 Funzione di ripartizione e quantili

$$F_X : \mathbb{R} \to [0,1] \qquad F_X(x) = \mathbb{P}\{X \leq x\} \qquad \text{(Cumulative Distribution Function)}$$
$$\mathbb{P}\{a < X \leq b\} = F(b) - F(a) \tag{9}$$
$$F_X(t) = \sum_{x_i \leq t} p(x_i) \qquad \text{(c.d.f. discreta)}$$
$$\mathbb{P}\{X = x\} = F(x) - F_{-}(x) \tag{10}$$
$$f(x) = \frac{dF(x)}{dx} \qquad \text{(c.d.f. continua)}$$

Proposizione 1.3.1. *Proprietà della c.d.f.:*

- *F non è decrescente* $(x < y \Rightarrow F(x) \leq F(y))$
- $\lim_{x\to-\infty} F(x) = 0 \qquad \lim_{x\to+\infty} F(x) = 1$
- *F è continua a destra* $(\forall x \in \mathbb{R} \quad F(x_n) \to F(x)$ *per ogni successione* $x_n \to x \quad x_n \geq x)$

$$r_\beta = inf\{r \in \mathbb{R} : F(r) \geq \beta\} \qquad \beta \in (0,1) \qquad (\beta\text{-quantile})$$
$$F^{\leftarrow} : (0,1) \to \mathbb{R} \qquad F^{\leftarrow}(t) = inf\{r \in \mathbb{R} : F(r) \geq t\} \qquad \text{(Inversa generalizzata)}$$

Proposizione 1.3.2. *Proprietà dell'inversa generalizzata:*

- *Se F è strettamente crescente* $F^{\leftarrow} = F^{-1}$
- $F^{\leftarrow}$ *è sempre non decrescente*
- $F^{\leftarrow}(F(t)) \leq t \quad \forall t \in \mathbb{R}$
- $F(F^{\leftarrow}(t)) \geq t \quad \forall t \in \mathbb{R}$
- $F^{\leftarrow}(t) \leq s \Leftrightarrow F(s) \geq t$

1.3.3 Variabili aleatorie notevoli discrete

$$\mathbb{P}(X = h) = \binom{n}{h} p^h (1-p)^{n-h} \qquad \text{(Binomiale } B(n,p))$$
$$\mathbb{P}(X = h) = (1-p)^{h-1} p \qquad h \in \mathbb{N}_0 \qquad \text{(Geometriche } G(p))$$
$$\mathbb{P}(X = k) = \frac{\binom{h}{k}\binom{n-h}{r-k}}{\binom{n}{r}} \qquad k = 0, \ldots, h \qquad \text{(Ipergeometriche } I(n,h,r))$$
$$\mathbb{P}(X = h) = e^{-\lambda} \frac{\lambda^h}{h!} \qquad h \in \mathbb{N} \qquad \text{(Poisson } P(\lambda))$$
$$\sum_{k=0}^{h} \binom{h}{k}\binom{n-h}{r-k} = \binom{n}{r} \qquad \text{(Identità di Vandermonde)}$$

1.3.4 Variabili aleatorie notevoli con densità

$$f(t) = \begin{cases} \frac{1}{b-a} & a < t < b \\ 0 & \text{altrove} \end{cases} \qquad F(t) = \begin{cases} 0 & t \le a \\ \frac{t}{b-a} & 0 < t \le b \\ 1 & t > b \end{cases} \qquad \text{(Uniformi su intervalli)}$$

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t > 0 \\ 0 & t \le 0 \end{cases} \qquad F(t) = \begin{cases} 1 - e^{-\lambda t} & t > 0 \\ 0 & t \le 0 \end{cases} \qquad \text{(Esponenziali)}$$

$$f(t) = \begin{cases} \alpha x_m^{\alpha} t^{-1-\alpha} & t > x_m \\ 0 & t \le x_m \end{cases} \qquad F(t) = \begin{cases} 1 & t < x_m \\ 1 - (\frac{x_m}{t})^{\alpha} & t \ge x_m \end{cases} \qquad \text{(Pareto)}$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \qquad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} \qquad \text{(Gaussiane standard } N(0,1))$$

$$f_Y(t) = \frac{1}{\sigma} f_X\left(\frac{t-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-m)^2}{2\sigma^2}} \qquad \text{(Gaussiane non standard } N(m,\sigma^2))$$

$$F_Y(t) = \mathbb{P}\{Y \le t\} = \mathbb{P}\{\sigma X + m \le t\} = \mathbb{P}\left(X \le \frac{t-m}{\sigma}\right) = \Phi\left(\frac{t-m}{\sigma}\right) \quad \text{(Gaussiane non standard 2)}$$

Proposizione 1.3.3. *Proprietà delle Gaussiane standard:*

$$\mathbb{P}\{-t \le Z \le t\} = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1 \tag{11}$$

$$\Phi(0) = \mathbb{P}\{X \ge 0\} = \mathbb{P}\{X \le 0\} = \frac{1}{2} \tag{12}$$

Definizione 1.3.1 (Variabili Gaussiane Non Standard -)**.** *Data X una v.a. $N(0,1)$ e Y una v.a. del tipo $Y = \sigma X + m$.*

1.3.5 Trasformazioni di variabili con densità

Proposizione 1.3.4. *Sia $h : A \to B$ biunivoca, differenziabile e con inversa differenziabile:*

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)) \cdot \left| \frac{dh^{-1}(y)}{dy} \right| & y \in B \\ 0 & y \notin B \end{cases} \qquad \text{(Cambio di variabile)}$$

h crescente:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(h(X) \le y) = \mathbb{P}(X \le h^{-1}(y)) = F_X(h^{-1}(y)) \tag{13}$$

$$f_Y(y) = f_X(h^{-1}(y)) \cdot \frac{dh^{-1}(y)}{dy} \tag{14}$$

h decrescente:

$$\mathbb{P}(h(X) \le y) = \mathbb{P}(X \le h^{-1}(y)) = 1 - F_X(h^{-1}(y)) \tag{15}$$

1.3.6 Valore atteso, varianza e momenti

$$\mathbb{E}[X] = \sum_i x_i p_X(x_i) \qquad \mathbb{E}[X] = \int_{-\infty}^{+\infty} t f_X(t) dt \qquad \text{(Valore atteso)}$$

Proposizione 1.3.5. *Sia X discreta, la variabile $g(x)$ ammette valore atteso se $\sum_i |g(x_i)| p(x_i) < +\infty$. In quel caso vale:*

$$\mathbb{E}[g(X)] = \sum_i g(x_i) p(x_i) \tag{16}$$

Sia X con densità, la variabile $g(x)$ ammette valore atteso se $\int_{-\infty}^{+\infty} |g(x)| f(x) dx < +\infty$. In quel caso vale:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx \tag{17}$$

Proposizione 1.3.6. *Se X ha valore atteso, valgono:*

- $\forall a, b \in \mathbb{R} \quad \mathbb{E}[aX + b] = a\mathbb{E}[X] + b \quad \mathbb{E}[b] = b$
- $|\mathbb{E}[X]| \leq E[|X|]$
- $\mathbb{P}(X \geq 0) = 1 \Longrightarrow \mathbb{E}[X] \geq 0$

$$\mathbb{E}[X^n] \qquad \mathbb{E}[|X|^n] < +\infty \qquad \text{(Momento)}$$

$$\mathbb{E}[|X|^m]^{\frac{1}{m}} \leq \mathbb{E}[|X|^n]^{\frac{1}{n}} \qquad 1 \leq m \leq n \qquad \text{(Disuguaglianza di Jensen)}$$

$$a\mathbb{P}\{X \geq a\} \leq \mathbb{E}[X] \qquad a > 0 \qquad \text{(Disuguaglianza di Markov)}$$

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \qquad \text{(Varianza)}$$

$$\sigma(X) = \sqrt{Var(X)} \qquad \text{(Scarto quadratico medio)}$$

$$\mathbb{P}\{|X - \mathbb{E}[X]| > d\} \leq \frac{Var(X)}{d^2} \qquad d > 0 \qquad \text{(Disuguaglianza di Chebyshev)}$$

1.3.7 Momenti di variabili aleatorie notevoli

$$\mathbb{E}[X^k] = p \qquad Var(X) = p - p^2 = p(1-p) \qquad k \geq 1 \qquad \text{(Bernoulli)}$$

$$\mathbb{E}[X] = np \qquad Var(X) = np(1-p) \qquad \text{(Binomiali)}$$

$$\mathbb{E}[X] = \sum_{h=0}^{+\infty} he^{-\lambda}\frac{\lambda^h}{h!} = \lambda \qquad \text{(Poisson)}$$

$$\mathbb{E}[X^2] = \lambda + \lambda^2 \qquad (18)$$

$$Var(X) = \lambda \qquad (19)$$

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a}dx = \frac{a^2 + ab + b^2}{3} \qquad Var(X) = \frac{(b-a)^2}{12} \qquad \text{(Uniformi su intervalli finiti)}$$

$$\mathbb{E}[X^n] = \frac{n!}{\lambda^n} \qquad Var(X) = \frac{1}{\lambda^2} \qquad \text{(Esponenziali)}$$

$$\mathbb{E}[X^{2h+1}] = 0 \qquad \text{(Gaussiana standard)}$$

$$\mathbb{E}[X^{2h+2}] = (2h+1)\mathbb{E}[X^{2h}] \qquad (20)$$

$$Var(X) = 1 \qquad (21)$$

$$\mathbb{E}[Y] = \mathbb{E}[\sigma X + m] \qquad Var(Y) = Var(\sigma X + m) = \sigma^2 Var(X) \qquad \text{(Gaussiana non standard)}$$

1.4 Distribuzioni multivariate

Proposizione 1.4.1 (Distribuzione di probabilità di variabile doppia discreta)**.**

$$p(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j) \qquad (22)$$

$$\mathbb{P}_{(X,Y)}(A) = \mathbb{P}\{(X,Y) \in A\} = \sum_{(x_i,y_j)\in A} p(x_i, y_j) \qquad (23)$$

$$p_X(x_i) = \sum_{j=1}^{\infty} p(x_i, y_j) \qquad (24)$$

$$p_Y(y_j) = \sum_{i=1}^{\infty} p(x_i, y_j) \qquad (25)$$

Proposizione 1.4.2 (Distribuzione di probabilità di variabile doppia con densità)**.**

$$\mathbb{P}_{(X,Y)}(A) = \mathbb{P}\{(X,Y) \in A\} = \int\int_A f(x,y)dxdy \qquad (26)$$

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy \tag{27}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx \tag{28}$$

1.4.1 Indipendenza di variabili aleatorie

Definizione 1.4.1 (Variabili aleatorie indipendenti)**.**

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot \ldots \cdot \mathbb{P}(X_n \in A_n) \tag{29}$$

$$p(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j) \qquad \forall (x_i, y_j) \tag{30}$$

$$f(x,y) = f_X(x) \cdot f_Y(y) \qquad \forall (x,y) \tag{31}$$

1.4.2 Funzioni di variabili indipendenti

Proposizione 1.4.3 (Somma di Binomiali)**.**

$$X \to B(n,p) \quad Y \to B(m,p) \Longrightarrow Z = X + Y \to B(n+m,p) \tag{32}$$

Proposizione 1.4.4 (Funzione di massa di somma di variabili discrete)**.**

$$Z = X + Y \Longrightarrow p_Z(n) = \sum_{h=0}^{n} p_X(h) \cdot p_Y(n-h) \tag{33}$$

Proposizione 1.4.5 (Funzione di massa di somma di variabili con densità o formula della convoluzione)**.**

$$Z = X + Y \Longrightarrow p_Z(n) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x)dx = \int_{-\infty}^{+\infty} f_Y(y) f_X(z-y)dy \tag{34}$$

1.4.3 Covarianza e correlazione

Proposizione 1.4.6 (Valore atteso di somma di variabili)**.** *Dati X e Y con valore atteso:*

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $X \geq Y \Longrightarrow \mathbb{E}[X] \geq \mathbb{E}[Y]$

Se sono anche indipendenti:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y] \tag{35}$$

$$\mathbb{E}[h(X)k(Y)] = \mathbb{E}[h(X)] \cdot \mathbb{E}[k(Y)] \tag{36}$$

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]} \cdot \sqrt{\mathbb{E}[Y^2]} \qquad \text{(Disuguaglianza di Schwartz)}$$

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \qquad \text{(Covarianza)}$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \qquad \text{(Coefficiente di correlazione)}$$

1.5 Variabili indipendenti e teoremi limite

$$\forall \epsilon > 0 \qquad \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0 \quad \text{(Convergenza in probabilità)}$$

$$\lim_{n\to\infty} \mathbb{E}[X_n] = c \in \mathbb{R} \wedge \lim_{n\to\infty} Var(X_n) = 0 \Longrightarrow (X_n)_{n\geq 1} \longrightarrow c$$

(Convergenza in probabilità ad una costante)

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}(|S_n^2 - \sigma^2| > \epsilon) = 0 \tag{37}$$

$$\lim_{n\to\infty} F_n(t) = F(t) \quad \forall t \in \mathbb{R} \Longrightarrow (X_n)_{n\geq 1} \longrightarrow X$$

(Convergenza in distribuzione)

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-1\frac{x^2}{2}} dx = \Phi(b) - \phi(a) \quad \text{(Teorema centrale del limite)}$$

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \qquad n \geq 50$$

(Approssimazione a variabile Gaussiana)

(38)

1.5.1 Variabili Chi-Quadro e di Student

Definizione 1.5.1 (Gamma di Eulero)**.**

$$\Gamma(r) = \int_0^{+\infty} x^{r-1} e^{-x} dx = (r-1)\Gamma(r-1) \qquad r > 0 \tag{39}$$

Definizione 1.5.2 (Densità Gamma - $\Gamma(r, \lambda)$)**.**

$$f(x) = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x} & x > 0 \\ 0 & z \leq 0 \end{cases} \tag{40}$$

$$\mathbb{E}[X^\beta] = \frac{\Gamma(r+\beta)}{\Gamma(r)\lambda^\beta} \tag{41}$$

$$Var(X) = \frac{r}{\lambda^2} \tag{42}$$

Proposizione 1.5.1 (Somma di densità gamma)**.**

$$X \to \Gamma(r, \lambda) \quad Y \to \Gamma(s, \lambda) \Longrightarrow (X + Y) \to \Gamma(r + s, \lambda) \tag{43}$$

Definizione 1.5.3 (Densità Chi-quadro)**.**

$$X_1, \ldots, X_n \to N(0,1) \Longrightarrow (X_1^2, \ldots, X_n^") \to \Gamma(\frac{n}{2}, \frac{1}{2}) = \mathcal{X}^2(n) \tag{44}$$

$$F_{X^2}(t) = \mathbb{P}(X^2 \leq t) = \begin{cases} 0 & t < 0 \\ \mathbb{P}(-\sqrt{t} \leq X \leq \sqrt{t}) = 2F_X(\sqrt{t}) - 1 & t \geq 1 \end{cases} \tag{45}$$

Proposizione 1.5.2 (Approssimazioni Chi-quadro)**.** *Data C_n variabile con densità $\mathcal{X}^2(n)$:*

- $\lim_{n\to\infty} \frac{C_2}{n} \approx 1$
- $\lim_{n\to\infty} \frac{C_n - n}{\sqrt{2n}} \approx N(0,1)$

Definizione 1.5.4 (Densità Student)**.**

$$T_n = \sqrt{n}\frac{X}{\sqrt{C_n}} \qquad X \to N(0,1) \quad C_n \to \mathcal{X}^2(n) \tag{46}$$

$$f_{T_n}(t) = \frac{\Gamma(\frac{n}{2} + \frac{1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n}{2} - \frac{1}{2}} \tag{47}$$

Proposizione 1.5.3.

$$(T_n)_{n\geq 1} \longrightarrow N(0,1) \tag{48}$$

Proposizione 1.5.4. *Date $X_1, \ldots, X_n$ v.a. $N(m, \sigma^2)$:*

- $\bar{X}_n$ *e* S_n^2 *sono indipendenti*
- $\bar{X} \to N(m, \frac{\sigma^2}{n})$
- $\frac{n-1}{\sigma^2} S_n^2 \to \mathcal{X}^2(n-1)$
- $T = \sqrt{n}\frac{(\bar{X}_n - m)}{S} \to T(n-1)$

1.6 Campioni statistici e stimatori

1.6.1 Stima parametrica

Definizione 1.6.1 (Stimatore corretto)**.** *Dato un parametro θ:*

$$\mathbb{E}_\theta[g(X_1, \ldots, X_n)] = \theta \tag{49}$$

Proposizione 1.6.1.

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X_i] \tag{50}$$

$$Var(\bar{X}) = \frac{Var(X_i)}{n} \tag{51}$$

$$\mathbb{E}[S_n^2] = Var(X_i) \tag{52}$$

Definizione 1.6.2 (Stimatore consistente)**.**

$$\lim_{n\to\infty} \mathbb{P}_\theta\{|g_n(X-1, \ldots, X_n) - \theta| > \epsilon\} = 0 \tag{53}$$

Definizione 1.6.3 (Stimatore efficiente)**.**

$$Var_\theta(g(X_1, \ldots, X_n)) \leq Var_\theta(h(X_1, \ldots, X_n)) \tag{54}$$

1.6.2 Massima verosimiglianza e metodo dei momenti

Definizione 1.6.4 (Funzione di verosimiglianza)**.**

$$L(\theta, x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\theta(x_i) \tag{55}$$

$$L(\theta, x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) \tag{56}$$

Definizione 1.6.5 (Stima di massima verosimiglianza)**.**

$$L(\hat{\theta}; x_1, \ldots, x_n) = \max_{\theta\in\Theta} L(\theta; x_1, \ldots, x_n) \tag{57}$$

Definizione 1.6.6 (Metodo dei momenti)**.**

$$\mathbb{E}_{\tilde{\theta}}[X^k] = \frac{1}{n}\sum_{i=1}^{n} x_i^k \qquad \forall (x_1, \ldots, x_n) \tag{58}$$

1.7 Intervalli di fiducia

Definizione 1.7.1 (Intervallo di fiducia per la media)**.**

$$\left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \tag{59}$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{60}$$

$$\left[\bar{X}_n \pm \frac{S_n}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \tag{61}$$

Definizione 1.7.2 (Intervallo di fiducia per la media - Bernoulli)**.**

$$\left[\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} q_{1-\frac{\alpha}{2}}\right] \tag{62}$$

Definizione 1.7.3 (Intervallo di fiducia per la varianza - Gaussiani)**.**

$$\left(0, \frac{(n-1)S_n^2}{\mathcal{X}^2_{\alpha,n-1}}\right] \qquad \left[\frac{(n-1)S_n^2}{\mathcal{X}^2_{1-\alpha,n-1}}, +\infty\right) \tag{63}$$

1.8 Test statistici

Definizione 1.8.1 (Regione critica di livello α)**.**

$$C = \left\{\sqrt{n}\frac{|\bar{X} - m_0|}{\sigma} > q_{1-\frac{\alpha}{2}}\right\} = \left\{|\bar{X}_n - m_0| > \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right\} \tag{64}$$

Definizione 1.8.2 (p-value)**.**

$$\bar{\alpha} = \mathbb{P}_{m_0}\left(\sqrt{n}\frac{\bar{X}_n - m_0|}{\sigma} > \frac{\sqrt{n}}{\sigma}|\bar{X}_n - m_0|\right) = 2\left[1 - \Phi\left(\frac{\sqrt{n}}{\sigma}|\bar{X}_n - m_0|\right)\right] \tag{65}$$

Definizione 1.8.3 (Curva operativa)**.**

$$\beta(m) = \Phi\left(\sqrt{n}\frac{|m_0 - m|}{\sigma} + q_{1-\alpha}\right) \tag{66}$$

1.8.1 Campione di Bernoulli

$$C = \left\{\frac{\sqrt{n}|\bar{X} - m_0|}{\sqrt{p_0(1-p_0)}} > q_{1-\frac{\alpha}{2}}\right\} \tag{67}$$

$$\bar{\alpha} = 2\left[1 - \Phi\left(\frac{\sqrt{n}|\bar{X}_n - p_0|}{\sqrt{p_o(1-p_0)}}\right)\right] \tag{68}$$

1.8.2 Campione Gaussiano

$$C = \left\{\frac{n-1}{\sigma_0^2}S_n^2 > \mathcal{X}^2_{2-\alpha,n-1}\right\} \tag{69}$$

$$\bar{\alpha} = 1 - G_{n-1}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{\sigma_0^2}\right) \tag{70}$$

$$\beta(\sigma) = G_{n-1}\left(\frac{\sigma_0^2}{\sigma^2}\mathcal{X}^2_{1-\alpha,n-1}\right) \tag{71}$$

2 Domande di teoria

Considerando il lancio di una moneta equilibrata e la variabile X che prende valore 1 se esce testa e 0 se croce, e la variabile Y che invece prende valore 0 se esce testa e 1 se croce, le variabili sono di Bernoulli $p = \frac{2}{2}$ e dunque la loro somma $X+Y$ è di tipo binomiale con $n = 2$ e $p = \frac{1}{2}$.
FALSO: la variabile $X + Y$ prende sempre valore 1.

Una variabile aleatoria X con varianza $Var(X) = 0$ è tale che esiste un numero $c \in \mathbb{R}$ tale che $\mathbb{P}(X = c) = 1$.
VERO: in effetti $c = \mathbb{E}[X]$ che è finito perché X ammette momento secondo essendo la varianza finita (anzi, nulla), e questo perché dalla disuguaglianza di Chebychev

$$\mathbb{P}(|X - \mathbb{E}[X] > \delta) = \frac{1}{\delta^2} Var(X) = 0$$

da cui mandando $\theta \to 0$ si ha $\mathbb{P}(X \neq \mathbb{E}[X]) = 1$.

Considerando una successione $X_1 = X_2 = X_3 = \ldots$di variabili uguali tra di loro e con momento secondo finito, per la Legge dei Grandi Numeri la quantità $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$ converge in probabilità a $\mathbb{E}[X]$.
FALSO: non è verificata l'ipotesi di indipendenza, e in questo caso $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n) = X_1$.

Una variabile Gaussiana $N(m, \sigma^2)$ ha tutti i momenti finiti.
VERO: la condizione di esistenza del momento n-esimo è:

$$\int_{-\infty}^{\infty} |t|^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(t-m)^2}{2\sigma^2}} dt < \infty$$

che è verificata per ogni n poiché l'esponenziale negativo decresce più velocemente di ogni potenza, e nello specifico:

$$|t|^n e^{\frac{-(t-m)^2}{2\sigma^2}} \leq e^{\frac{-c(t-m)^2}{2\sigma^2}}$$

per $c > 0$ abbastanza piccola.

Dato un campione statistico $X_1, \ldots, X_n$ di v.a. indipendenti con momento secondo finito, la funzione $Var(\bar{X}_n)$ è decrescente in n.
VERO: infatti per indipendenza $Var(\bar{X}_n) = \frac{Var(X_1)}{n}$

Se il p-value di un test di ipotesi è 0.001, allora l'ipotesi nulla è plausibile per ogni ragionevole livello.
FALSO: si rifiuta l'ipotesi nulla per ogni livello $\alpha > 0.001$ quindi per ogni ragionevole livello.

In uno spazio di probabilità, due eventi disgiunti sono indipendenti.
FALSO: se A, B sono disgiunti e indipendenti,

$$\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$$

è soddisfatta se e solo se uno dei due eventi è esso stesso vuoto, per cui l'affermazione non è verificata in generale.

Una variabile aleatoria che assume infiniti valori è una variabile con densità.
FALSO: ad esempio una variabile di Poisson può assumere qualsiasi valore naturale $(0, 1, 2, \ldots)$ ma non è una variabile con densità.

I momenti $\mathbb{E}_n$ di una variabile aleatoria X di Bernoulli di parametro p sono tutti uguali.
VERO: infatti vale sempre $X_n = X$, perché se $X = 0$ allora $X_n = 0^n = 0$, e similmente se $X = 1$ allora $X_n = 1^n = 1$, di conseguenza:

$$\mathbb{E}[X^n] = 1^n\mathbb{P}(X^n = 1) + 0^n\mathbb{P}(X^n = 0) = 1\mathbb{P}(X = 1) + 0\mathbb{P}(X = 0) = \mathbb{E}[X]$$

Lo stimatore $\frac{1}{n-1}\sum_i /X_i - \bar{X})^2$ ha denominatore $n-1$ per far sì che esso sia consistente.
FALSO: il denominatore è scelto in modo che lo stimatore sia corretto, e quest'ultimo sarebbe consistente anche con denominatore n per la Legge dei Grandi numeri e il fatto che $\frac{n-1}{n} \to 1$ per $n \to \infty$.

Due eventi A e B di probabilità positiva sono indipendenti se e solo se $P(A|B) = P(B)$.
FALSO: A e B sono indipendenti se e solo se $P(A|B) = P(A)$.

Per una variabile aleatoria X con funzione di ripartizione $F(x) = P(X \leq x)$, abbiamo $P(a < X \leq b) = \int_a^b F(x)dx$.
FALSO: $P(a < X \leq b) = F(b) - F(a)$.

Un intervallo di fiducia ha come estremi due variabili aleatorie e può non contenere il parametro stimando.
VERO: un intervallo di fiducia è dato da una coppia di variabili aleatorie $I = [a(\omega), b(\omega)]$ date da opportune funzioni del campione aleatorio, e contiene il parametro θ con una certa probabilità, determinata dal livello di fiducia.

Imporre il livello di un test statistico fissa un limite alla probabilità di commettere errori di prima specie, ma non dà vincoli su quelli di seconda.
VERO: la prima parte dell'enunciato è la definizione di livello, mentre va ricordato che è la potenza del test a quantificare la probabilità di commettere errori di seconda specie.

Considerando per fissare le idee lo Z-test, all'aumentare della numerosità del campione la regione di accettazione si riduce, ossia è più facile che l'ipotesi nulla sia rifiutata.
VERO: l'ampiezza della regione di accettazione è $\frac{2\sigma q_{1-\frac{\alpha}{2}}}{n}$, dunque è decrescente nella numerosità n.

Due variabili aleatorie indipendenti (di varianza finita) sono sempre scorrelate.
VERO: come visto a lezione l'indipendenza di X, Y implica che $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, che è il caso con valori attesi nulli, da cui quello generale segue con passaggi elementari.

In un test statistico, il p-value è la probabilità che l'ipotesi nulla sia vera.
FALSO: per definizione il p-value è la probabilità di esiti di coda della statistica su cui il test si basa, sotto l'ipotesi nulla. NON si può determinare "la probabilità che l'ipotesi nulla sia vera" nell'approccio frequentista alla statistica.

Una variabile aleatoria è discreta se e solo se assume una quantità finita di esiti.
FALSO: è falsa la parte "solo se", perché esistono variabili discrete con infiniti esiti di probabilità positiva, ad esempio le variabili di Poisson.

Una variabile aleatoria X con legge Gaussiana $N(0, \sigma^2)$ prende valori nell'intervallo $[-\sigma, \sigma]$ con probabilità di circa il 95%.
FALSO: La probabilità cercata è $\Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0.6826$.

In un test statistico, fissare un livello α è equivalente a decidere se il p-value è abbastanza alto da ritenere accettabile H_0).
VERO: il p-value è per definizione il livello α sotto cui, fissati gli esiti dell'esperimento, si accetta H_0).

Se un campione statistico $X_1, \ldots, X_n$ ha legge con momento primo finito, allora la media campionaria è uno stimatore corretto del valore atteso.
VERO: infatti per linearità del valore atteso, e usando che le variabili del campione hanno lo stesso valore atteso,

$$\mathbb{E}[\frac{1}{n}(X_1 + \ldots + X_n)] = \frac{1}{n}(\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]) = \mathbb{E}[X_1]$$

Dato un campione aleatorio $X_1, \ldots, X_n$, il valore atteso delle X_i coincide con la media campionaria del campione $X_1, \ldots, X_n$.
FALSO: Ad esempio, per l'esperimento lancio di una moneta equilibrata, e $X = 1$ se esce testa, $X = 0$ se esce croce, il valore atteso di X è $\frac{1}{2}$, ma la media campionaria del campione (testa, testa) è 1.

Se due eventi A, B sono indipendenti, lo sono anche i loro complementari A^c , B^c .
VERO: se A, B sono eventi indipendenti, dalla legge di De Morgan e la definizione di indipendenza si deduce che

$$P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A) - P(B) + P(A \cap B) =$$
$$= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P(A^c)P(B^c)$$

Dato un campione aleatorio $X_1, \ldots, X_n$ la cui legge dipende da un solo parametro $\theta \in \Theta \subseteq \mathbb{R}$, un intervallo di fiducia per il parametro θ è dato da una qualsiasi coppia di numeri $a, b \in \mathbb{R}$ in modo che $I = (a, b)$ contenga θ.
FALSO: un intervallo di fiducia è dato da una coppia di variabili aleatorie, $I = [a(\omega), b(\omega)]$, e contiene il parametro θ con una certa probabilità, determinata dal livello di fiducia.

Una variabile aleatoria X ammette funzione di densità $f : \mathbb{R} \to \mathbb{R}$ se la probabilità che X assuma il valore $x \in \mathbb{R}$ è dato da $f(x)$.
FALSO: se $f(x)$ è una funzione tale che

$$\mathbb{P}(X = x) = f(x)$$

allora la condizione $\mathbb{P}(\Omega) = 1$ impone che f sia non nulla su al più numerabili valori, e X è dunque una variabile discreta.

Un p-value dell'ordine di 10^{-5} è da considerarsi evidenza statistica contro l'ipotesi nulla.
VERO: se il livello $\alpha > \alpha = 10^{-5}$, l'ipotesi nulla viene rifiutata dal test di regione critica di livello α, dunque un p-value così piccolo fa sì che l'ipotesi venga rifiutata per ogni livello α ragionevole.

Per effettuare il test di Student sulla media di una popolazione è necessario conoscere la deviazione standard della popolazione.
FALSO: Il test di Student usa la deviazione standard campionaria.

Se due variabili aleatorie X, Y hanno la stessa distribuzione di probabilità, allora non sono indipendenti.
FALSO: le singole componenti di una scelta uniforme su $\{(0,0),(0,1),(1,0),(1,1)\}$ sono $Bernoulli\left(\frac{1}{2}\right)$ indipendenti.

Due variabili aleatorie X, Y soddisfano una relazione lineare $Y = aX + b$ se e solo se la loro correlazione vale $\rho(X,Y) = \pm 1$.
VERO: come visto nel corso vale

$$min_{a,b}\mathbb{E}[(Y - aX - n)^2] = Var(Y)(1 - \rho(X,Y)^2)$$

per cui quando il minimo a sinistra è nullo deve valere $\rho(X,Y) = 1$.

Data una generica variabile aleatoria X, la sua funzione di ripartizione è data dall'integrale $F_X(t) = \int_0^t f(x)dx$ di una certa funzione f .
FALSO: ciò è vero, per definizione, solo per le variabili con densità, e per cui la densità f sia zero per $x < 0$; ad esempio non esiste una funzione f che soddisfa la relazione proposta se X è una variabile di Bernoulli.

Scambiare l'ipotesi nulla e quella alternativa può cambiare l'esito di un test statistico.
VERO: le condizioni su livello e potenza non sono simmetriche nelle due ipotesi. Ad esempio se si è misurata l'altezza di 100 persone ottenendo media campionaria $175.5cm$, supponendo deviazione standard $5cm$, consideriamo due test con ipotesi nulle $H_0)m \leq 175eH_0')m > 175$. In entrambi i casi $Z = \sqrt{N}\frac{(\bar{X}-m)}{\sigma} = 1$, e i quantili da confrontare sono $q_{0.95} = 1.64$ e$q0.05 = -1.64$. In particolare, in entrambi i casi siamo nella regione critica, quindi scambiare un ipotesi nulla che sta venendo rigettata con la corrispondente alternativa non conduce ad accettazione.

In generale, più basso è il livello imposto a un test statistico, inferiore è la potenza.
VERO: il livello è la probabilità sotto H_0 della regione critica, ed imporlo più basso significa ridurre la regione critica; a sua volta la potenza è la probabilità sotto l'ipotesi alternativa della regione critica, ma se quest'ultima si è ridotta la potenza non può aumentare.

In un test statistico una forte evidenza a favore dell'ipotesi nulla può produrre un p-value superiore a 1.
FALSO: il p-value è una probabilità e non può mai assumere valori superiori a 1.

Una variabile aleatoria $X : \Omega \to \mathbb{R}$ costante, ovvero $X(\omega) = c$ per ogni $\omega \in \Omega$, è indipendente da qualsiasi altra variabile aleatoria $Y : \Omega \to \mathbb{R}$.
VERO: l'evento $\{X \in A\}$ è sempre uno tra $\emptyset$ e Ω, ma questi ultimi sono indipendenti da ogni evento $\{Y \in B\}$.

Il β-quantile di una variabile aleatoria con densità esponenziale di parametro $\lambda > 0$ è $\frac{-\log(1-\beta)}{\lambda}$.
VERO: $F(x) = 1 - e^{-\lambda x} = \beta$ se e solo se $x = \frac{-\log(1-\beta)}{\lambda}$.

Due variabili aleatorie scorrelate sono anche indipendenti.
FALSO: Ad esempio: X uniforme su $[-1, 1]$ e X^2 sono scorrelate ma non indipendenti.

La media campionaria $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$ di un campione di variabili aleatorie con momento secondo finito è uno stimatore corretto e consistente del parametro $m = \mathbb{E}[X_i]$.
VERO: la correttezza discende dalla linearità del valore atteso, mentre la consistenza dalla Legge Debole dei Grandi Numeri.

Un intervallo di fiducia ha livello 95% se esso contiene almeno il 95% dei dati del campione.
FALSO: non ci sono legami tra intervallo di fiducia e numero di dati in esso contenuti.

Se X è una v.a. discreta con funzione di massa p_X, la funzione di ripartizione F_X di X è $F_X(x) = \int_{-\infty}^{x} p_X(y)dy$.
FALSO: La funzione di ripartizione è $F_X(x) = \sum_{x_i \leq x} p_X(x_i)$.