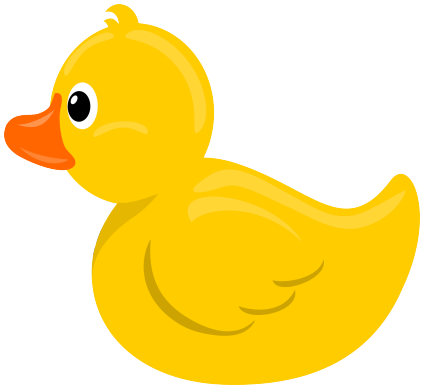
Probabilità e Statistica per l'Informatica

Ver 1.1

Falbo Andrea

A.A 2022/2023

LilQuacky



Indice

FORMULARIO DI PROBABILITÀ E STATISTICA PER L'INFORMATICA

Prima parte

Statistica descrittiva

Insieme di dati $x_1, \ldots, x_N$, riordinamento $x_{(1)} \leq \ldots \leq x_{(N)}$.

- Media campionaria: $\overline{x} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i$

  $\rightsquigarrow$ $\overline{x} = \dfrac{\sum_{j=1}^{M} z_j f_j}{\sum_{j=1}^{M} f_j}$ (valori assunti $z_1, \ldots, z_M$ con frequenze $f_1, \ldots, f_M$)

- Mediana campionaria: $m = \begin{cases} x_{(\frac{N+1}{2})} & \text{se } N \text{ è dispari} \\ \dfrac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} & \text{se } N \text{ è pari} \end{cases}$

- $100p$-esimo percentile campionario: $\begin{cases} x_{(i)} \ (i \text{ intero succ. a } Np) & \text{se } Np \text{ non è intero} \\ \dfrac{x_{(Np)} + x_{(Np+1)}}{2} & \text{se } Np \text{ è intero} \end{cases}$

  $\rightsquigarrow$ Quartili: q_1 $(p = \frac{1}{4})$, $q_2 = m$ $(p = \frac{1}{2})$, q_3 $(p = \frac{3}{4})$

- Varianza campionaria: $s^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(x_i - \overline{x})^2 = \dfrac{1}{N-1}\left\{ \displaystyle\sum_{i=1}^{N} x_i^2 - N\overline{x}^2 \right\}$

- Deviazione standard campionaria: $s = \sqrt{s^2} = \sqrt{\dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(x_i - \overline{x})^2}$

- Scarto interquartile: $\Delta = \mathrm{IQR} = q_3 - q_1$

- Coeff. di correlazione lineare campionario: $r = \dfrac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{(N-1)s_x s_y} = \dfrac{\sum_{i=1}^{N} x_i y_i \ - \ N\overline{xy}}{(N-1)s_x s_y}$

ULTIMA MODIFICA: 17 maggio 2022

Spazi di probabilità

Assiomi della probabilità:

- $\mathrm{P}(\Omega) = 1$
- se A e B sono disgiunti ($A \cap B = \emptyset$): $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$
- se $(A_i)_{i=1}^{\infty}$ sono disgiunti ($A_i \cap A_j = \emptyset$ per $i \neq j$): $\mathrm{P}\big(\bigcup_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$

Proprietà della probabilità:

$$\mathrm{P}(\emptyset) = 0$$

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$$

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

$$\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$$

$$\text{se } A \subseteq B\text{:} \quad \mathrm{P}(A) \leq \mathrm{P}(B)$$

Calcolo combinatorio

Dato un insieme di n elementi:

- le disposizioni con ripetizione di k elementi sono n^k
- le disposizioni semplici di k elementi sono $n!/(n-k)!$
- le combinazioni di k elementi sono $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Probabilità condizionale e indipendenza di eventi

Siano $A, B, (B_i)$ eventi in uno spazio di probabilità.

- Regola del prodotto: $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A)$

  $\mathrm{P}(B_1 \cap \cdots \cap B_n) = \mathrm{P}(B_1)\,\mathrm{P}(B_2|B_1) \cdots \mathrm{P}(B_n|B_1 \cap \ldots \cap B_{n-1})$

- Formula di disintegrazione: $\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c)$

  $\{B_1, \ldots, B_n\}$ partizione di Ω : $\mathrm{P}(A) = \mathrm{P}(A \cap B_1) + \ldots + \mathrm{P}(A \cap B_n)$

- Formula delle probabilità totali: $\mathrm{P}(A) = \mathrm{P}(A|B)\,\mathrm{P}(B) + \mathrm{P}(A|B^c)\,\mathrm{P}(B^c)$

  $\{B_1, \ldots, B_n\}$ partizione di Ω : $\mathrm{P}(A) = \mathrm{P}(A|B_1)\,\mathrm{P}(B_1) + \ldots + \mathrm{P}(A|B_n)\,\mathrm{P}(B_n)$

- Formula di Bayes: $$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

Variabili aleatorie discrete

Variabile aleatoria $X : \Omega \to \mathbb{R}$ discreta: quantità finita o numerabile di valori assunti

$$X(\Omega) = \{x_i\} = \{x_1, x_2, x_3, \dots\}$$

- Densità discreta: $\mathrm{p}_X(x_i) = \mathrm{P}(X = x_i)$ $\qquad (\mathrm{p}_X(x) = 0 \text{ se } x \notin \{x_i\})$
- Distribuzione: $\mathrm{P}(X \in A) = \sum_{x_i \in A} \mathrm{p}_X(x_i)$
- Valore medio: $\mathrm{E}[X] = \sum_{x_i} x_i \cdot \mathrm{p}_X(x_i)$

$$\mathrm{E}[X + c] = \mathrm{E}[X] + c \qquad \mathrm{E}[cX] = c\,\mathrm{E}[X] \qquad \mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

$$\text{se } X = c \text{ (costante) allora } \mathrm{E}[X] = c \qquad \text{se } X \geq 0 \text{ allora } \mathrm{E}[X] \geq 0$$

- Varianza: $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ $\qquad$ con $\qquad \mathrm{E}[X^2] = \sum_{x_i} x_i^2 \cdot \mathrm{p}_X(x_i)$

$$\mathrm{Var}[X + c] = \mathrm{Var}[X] \qquad \mathrm{Var}[cX] = c^2\,\mathrm{Var}[X]$$

$$\text{se } X \text{ e } Y \text{ sono indipendenti:} \quad \mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

$$X = c \text{ (costante)} \iff \mathrm{Var}[X] = 0$$

- Deviazione standard: $\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$

Distribuzioni notevoli discrete

Distribuzione	$X(\Omega)$	$\mathrm{p}_X(k)$ per $k \in X(\Omega)$	$\mathrm{E}[X]$	$\mathrm{Var}[X]$
Bernoulli $\mathrm{Be}(p)$ $p \in [0,1]$	$\{0,1\}$	$\begin{cases} p & \text{se } k = 1 \\ 1-p & \text{se } k = 0 \end{cases}$	p	$p(1-p)$
Binomiale $\mathrm{Bin}(n,p)$ $n \in \{1,2,\dots\}$ $p \in [0,1]$	$\{0,1,\dots,n\}$	$\binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Poisson $\mathrm{Pois}(\lambda)$ $\lambda \in (0,\infty)$	$\mathbb{N}_0 = \{0,1,\dots\}$	$\mathrm{e}^{-\lambda}\dfrac{\lambda^k}{k!}$	λ	λ
Geometrica $\mathrm{Geo}(p)$ $p \in (0,1]$	$\mathbb{N} = \{1,2,\dots\}$	$p(1-p)^{k-1}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$

Variabili aleatorie assolutamente continue

Variabile aleatoria $X : \Omega \to \mathbb{R}$ assolutamente continua con densità $f_X(x)$:

- Distribuzione: $\mathrm{P}(X \in A) = \displaystyle\int_A f_X(x)\,\mathrm{d}x$
- Valori assunti: $X(\Omega) = \{x \in \mathbb{R} :\ f_X(x) > 0\}$
- Valore medio: $\mathrm{E}[X] = \displaystyle\int_{-\infty}^{+\infty} x \cdot f_X(x)\,\mathrm{d}x$

$$\mathrm{E}[X + c] = \mathrm{E}[X] + c \qquad \mathrm{E}[cX] = c\,\mathrm{E}[X] \qquad \mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

$$\text{se } X = c \text{ (costante) allora } \mathrm{E}[X] = c \qquad \text{se } X \geq 0 \text{ allora } \mathrm{E}[X] \geq 0$$

- Varianza: $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ con $\mathrm{E}[X^2] = \displaystyle\int_{-\infty}^{+\infty} x^2 \cdot f_X(x)\,\mathrm{d}x$

$$\mathrm{Var}[X + c] = \mathrm{Var}[X] \qquad \mathrm{Var}[cX] = c^2\,\mathrm{Var}[X]$$

$$\text{se } X \text{ e } Y \text{ sono indipendenti:} \quad \mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

$$X = c \text{ (costante)} \iff \mathrm{Var}[X] = 0$$

- Deviazione standard: $\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$

Distribuzioni notevoli assolutamente continue

Distribuzione	$X(\Omega)$	$f_X(x)$ per $x \in X(\Omega)$	$F_X(x)$ per $x \in X(\Omega)$	$\mathrm{E}[X]$	$\mathrm{Var}[X]$
Uniforme continua $\mathrm{U}(a, b)$ $a, b \in \mathbb{R}$ con $a < b$	$[a, b]$	$\dfrac{1}{b-a}$	$\dfrac{x-a}{b-a}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Esponenziale $\mathrm{Exp}(\lambda)$ $\lambda \in (0, \infty)$	$[0, \infty)$	$\lambda\,\mathrm{e}^{-\lambda x}$	$1 - \mathrm{e}^{-\lambda x}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Normale $\mathrm{N}(\mu, \sigma^2)$ $\mu \in \mathbb{R}$ $\sigma \in (0, \infty)$	$(-\infty, +\infty)$	$\dfrac{\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$	$\Phi(x)$	μ	σ^2

Funzione di ripartizione

Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria.

- Funzione di ripartizione: $F_X(x) = \mathrm{P}(X \leq x)$
- Probabilità di intervalli: $\mathrm{P}(X \in (a, b]) = F_X(b) - F_X(a)$
- F_X continua dappertutto e derivabile a tratti $\rightsquigarrow$ $\begin{cases} X \text{ v.a. assolutamente continua} \\ f_X(x) = (F_X)'(x) \end{cases}$
- F_X costante a tratti $\rightsquigarrow$ $\begin{cases} X \text{ v.a. discreta} \\ \text{valori assunti } \{x_i\} = \text{punti di discontinuità di } F_X \\ \mathrm{p}_X(x_i) = F_X(x_i) - F_X(x_i^-) \quad \left(\text{con } F_X(x^-) := \lim\limits_{t \to x^-} F_X(t)\right) \end{cases}$

Vettori aleatori

- Covarianza: $\mathrm{Cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$

$$\mathrm{E}[XY] = \begin{cases} \displaystyle\sum_{x_i} \sum_{y_j} x_i \cdot y_j \cdot \mathrm{p}_{(X,Y)}(x_i, y_j) & \text{se } (X, Y) \text{ è un vettore discreto con densità discreta congiunta } \mathrm{p}_{(X,Y)}(x, y) \\ \displaystyle\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \cdot y \cdot f_{(X,Y)}(x, y)\,\mathrm{d}x\,\mathrm{d}y & \text{se } (X, Y) \text{ è un vettore assolut. cont. con densità congiunta } f_{(X,Y)}(x, y) \end{cases}$$

- Varianza della somma: $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X, Y]$
- Indipendenza di v.a. X e Y: $\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A) \cdot \mathrm{P}(Y \in B) \quad \forall A, B \subseteq \mathbb{R}$

  Se (X, Y) è discreto: $\mathrm{p}_{(X,Y)}(x_i, y_j) = \mathrm{p}_X(x_i) \cdot \mathrm{p}_Y(y_j) \quad \forall x_i,\, y_j$

  Se (X, Y) è assolutamente continuo: $f_{(X,Y)}(x, y) = f_X(x) \cdot f_Y(y) \quad \forall x,\, y$

- X e Y indipendenti $\rightsquigarrow$ $\mathrm{Cov}[X, Y] = 0$ $\rightsquigarrow$ $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$
- Indice di correlazione lineare: $\rho[X, Y] = \dfrac{\mathrm{Cov}[X, Y]}{\mathrm{SD}[X]\,\mathrm{SD}[Y]}$

Seconda parte

Teorema del limite centrale

$X_1, ..., X_n, ...$ v.a. i.i.d. con media μ e varianza σ^2 e sia $\bar{X}_n = \dfrac{X_1 + ... + X_n}{n}$

$$\mathrm{P}\Big(\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} \leq t\Big) \to \Phi(t) \text{ se } n \to \infty$$

dove $\Phi(t) = \mathrm{P}(Z \leq t)$, $Z \sim \mathcal{N}(0,1)$

Stima puntuale

• $X_1, ..., X_n$ campione casuale estratto da una popolazione con media incognita.
Stimatore non distorto della media

$$\bar{X}_n = \frac{X_1 + ... + X_n}{n}$$

• $X_1, ..., X_n$ campione casuale estratto da una popolazione con media e varianza incognite .
Stimatore non distorto della varianza

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

• $X_1, ..., X_n$ campione casuale estratto da una popolazione con media nota pari a μ e varianza incognita .
Stimatore non distorto della varianza

$$\bar{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \Big(X_i - \mu\Big)^2$$

Distribuzioni utili per le statistiche campionarie

- $Z \sim \mathcal{N}(0,1)$ e $\alpha \in (0,1)$, si pone $z_\alpha \in \mathbb{R}$ quel valore tale che $\mathbb{P}(Z > z_\alpha) = \alpha$. N.B: $z_\alpha = -z_{1-\alpha}$.
- $Z_1, ..., Z_n$ i.i.d. normali standard

  $$Y = Z_1^2 + ... + Z_n^2, \qquad Y \sim \chi^2(n)$$

  Y ha una distribuzione chi quadrato con n gradi di libertà: $Y \geq 0$
  Per $\alpha \in (0,1)$ si pone $\chi^2_{n,\alpha} \in \mathbb{R}$ quel valore tale che $\mathbb{P}(Y > \chi^2_{n,\alpha}) = \alpha$.

  $$\mathbb{E}[Y] = n, \ \mathrm{var}[Y] = 2n$$

- Siano $Z \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(n)$ indipendenti

  $$T = \frac{Z}{\sqrt{Y/n}}, \qquad T \sim t(n)$$

  T ha una distribuzione t di Student con n gradi di libertà. T simmetrica rispetto a 0.
  Per $\alpha \in (0,1)$ si pone $t_{n,\alpha} \in \mathbb{R}$ quel valore tale che $\mathbb{P}(T > t_{n,\alpha}) = \alpha$. N.B: $t_{n,\alpha} = -t_{n,1-\alpha}$

Stima per intervalli

Daremo formule per intervalli di confidenza, (estremi inferiori o superiori) al livello di $100(1-\alpha)\%$, e daremo la realizzazione dell'intervallo sui dati campionari $x_1, ..., x_n$.

campione numeroso $\rightsquigarrow$ $n \geq 30$

- campione estratto da una popolazione normale con media incognita e varianza nota pari a σ^2 (vale anche per campioni numerosi non necessariamente normali): stima intervallare della media

$$\text{Intervallo di confidenza } \left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$\text{Estremo inferiore } \bar{x}_n - z_\alpha\frac{\sigma}{\sqrt{n}}, \quad \text{intervallo destro } \left(\bar{x}_n - z_\alpha\frac{\sigma}{\sqrt{n}}, +\infty\right)$$

$$\text{Estremo superiore } \bar{x}_n + z_\alpha\frac{\sigma}{\sqrt{n}}, \quad \text{intervallo sinistro } \left(-\infty, \bar{x}_n + z_\alpha\frac{\sigma}{\sqrt{n}}\right)$$

- campione estratto da una popolazione normale con media e varianza incognite (vale anche per campioni numerosi non necessariamente normali): stima intervallare della media

$$\text{Intervallo di confidenza } \left(\bar{x}_n - t_{n-1\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right)$$

$$\text{Estremo inferiore } \bar{x}_n - t_{n-1,\alpha}\frac{s_n}{\sqrt{n}}, \quad \text{intervallo destro } \left(\bar{x}_n - t_{n-1,\alpha}\frac{s_n}{\sqrt{n}}, +\infty\right)$$

$$\text{Estremo superiore } \bar{x}_n + t_{n-1,\alpha}\frac{s_n}{\sqrt{n}}, \quad \text{intervallo sinistro } \left(-\infty, \bar{x}_n + t_{n-1,\alpha}\frac{s_n}{\sqrt{n}}\right)$$

- campione numeroso estratto da una popolazione Bernoulliana con media e varianza incognite (vale anche per campioni numerosi non necessariamente normali): stima intervallare della proporzione-frequenza: ok se $n\bar{x}_n > 5$, $n(1-\bar{x}_n) > 5$.

$$\text{Intervallo di confidenza } \left(\bar{x}_n - z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \bar{x}_n + z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}\right) \text{ N.B.: } \bar{x}_n(1-\bar{x}_n) \leq \frac{1}{4}.$$

$$\text{Estremo inferiore } \bar{x}_n - z_\alpha\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \quad \text{intervallo destro } \left(\bar{x}_n - z_\alpha\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, +\infty\right)$$

$$\text{Estremo superiore } \bar{x}_n + z_\alpha\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \quad \text{intervallo sinistro } \left(-\infty, \bar{x}_n + z_\alpha\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}\right)$$

- campione estratto da una popolazione normale con media e varianza incognite: stima intervallare della varianza

$$\text{Intervallo di confidenza } \left(\frac{(n-1)s_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s_n^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

$$\text{Estremo inferiore } \frac{(n-1)s_n^2}{\chi^2_{n-1,\alpha}}, \quad \text{intervallo destro } \left(\frac{(n-1)s_n^2}{\chi^2_{n-1,\alpha}}, +\infty\right)$$

$$\text{Estremo superiore } \frac{(n-1)s_n^2}{\chi^2_{n-1,1-\alpha}}, \quad \text{intervallo sinistro } \left[0, \frac{(n-1)s_n^2}{\chi^2_{n-1,1-\alpha}}\right)$$

- campione estratto da una popolazione normale con media nota e varianza incognite: stima intervallare della varianza

$$\text{Intervallo di confidenza } \left(\frac{n\bar{s}_n^2}{\chi^2_{n,\alpha/2}}, \frac{n\bar{s}_n^2}{\chi^2_{n,1-\alpha/2}}\right)$$

$$\text{Estremo inferiore } \frac{n\bar{s}_n^2}{\chi^2_{n,\alpha}}, \quad \text{intervallo destro } \left(\frac{n\bar{s}_n^2}{\chi^2_{n,\alpha}}, +\infty\right)$$

$$\text{Estremo superiore } \frac{n\bar{s}_n^2}{\chi^2_{n,1-\alpha}}, \quad \text{intervallo sinistro } \left[0, \frac{n\bar{s}_n^2}{\chi^2_{n,1-\alpha}}\right)$$

Test di Ipotesi

α= livello di significatività

- Test z sulla media di una popolazione normale con varianza nota pari a σ^2 (vale anche per campioni numerosi estratti da popolazioni non necessariamente normali)

H_0	H_1	Statistica	Regione critica
$\mu = \mu_0$	$\mu \neq \mu_0$	$Z = \frac{\bar{X}_n - \mu_0}{\sigma}\sqrt{n}$	$\left\lvert\frac{\bar{x}_n - \mu_0}{\sigma}\sqrt{n}\right\rvert > z_{\alpha/2}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$Z = \frac{\bar{X}_n - \mu_0}{\sigma}\sqrt{n}$	$\frac{\bar{x}_n - \mu_0}{\sigma}\sqrt{n} > z_\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	$Z = \frac{\bar{X}_n - \mu_0}{\sigma}\sqrt{n}$	$\frac{\bar{x}_n - \mu_0}{\sigma}\sqrt{n} < -z_\alpha$

- Test t sulla media di una popolazione normale con varianza incognita (vale anche per campioni numerosi estratti da popolazioni non necessariamente normali)

H_0	H_1	Statistica	Regione critica
$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$	$\left\lvert\frac{\bar{x}_n - \mu_0}{s_n}\sqrt{n}\right\rvert > t_{n-1,\alpha/2}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$T = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$	$\frac{\bar{x}_n - \mu_0}{s_n}\sqrt{n} > t_{n-1,\alpha}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$T = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$	$\frac{\bar{x}_n - \mu_0}{s_n}\sqrt{n} < -t_{n-1,\alpha}$

- Test z approssimato sulla proporzione con $n \geq 30$, $np_0 \geq 5$, $n(1-p_0) \geq 5$.

H_0	H_1	Statistica	Regione critica
$p = p_0$	$p \neq p_0$	$Z = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$\left\lvert\frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}\right\rvert > z_{\alpha/2}$
$p \leq p_0$	$p > p_0$	$Z = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$\frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} > z_\alpha$
$p \geq p_0$	$p < p_0$	$Z = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$\frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} < -z_\alpha$

- Test t sulla differenza delle media di due campioni normali **accoppiati** $X_1, ..., X_n$ di media μ_X e $Y_1, ..., Y_n$ di media μ_Y

  Test t sul campione delle differenze $D_1 = X_1 - Y_1, ..., D_n = X_n - Y_n$, denotiamo con $\bar{D}_n = \frac{1}{n}(D_1 + \cdots + D_n)$ la media campionaria e con S_d^2 la varianza campionaria del campione $D_1, \ldots, D_n$.

H_0	H_1	Statistica	Regione critica
$\mu_X = \mu_Y + \mu_0 \rightsquigarrow \mu_d = \mu_0$	$\mu_X \neq \mu_Y + \mu_0 \rightsquigarrow \mu_d \neq \mu_0$	$T = \frac{\bar{D}_n - \mu_0}{S_d}\sqrt{n}$	$\left\lvert \frac{\bar{d}_n - \mu_0}{s_d}\sqrt{n} \right\rvert > t_{n-1,\alpha/2}$
$\mu_X \leq \mu_Y + \mu_0 \rightsquigarrow \mu_d \leq \mu_0$	$\mu_X > \mu_Y + \mu_0 \rightsquigarrow \mu_d > \mu_0$	$T = \frac{\bar{D}_n - \mu_0}{S_d}\sqrt{n}$	$\frac{d_n - \mu_0}{s_d}\sqrt{n} > t_{n-1,\alpha}$
$\mu_X \geq \mu_Y + \mu_0 \rightsquigarrow \mu_d \geq \mu_0$	$\mu_X < \mu_Y + \mu_0 \rightsquigarrow \mu_d < \mu_0$	$T = \frac{\bar{D}_n - \mu_0}{S_d}\sqrt{n}$	$\frac{d_n - \mu_0}{s_d}\sqrt{n} < -t_{n-1,\alpha}$

- Test t sulla differenza delle media di due campioni normali indipendenti $X_1, ..., X_{n_x}$ di media μ_x e $Y_1, ..., Y_{n_y}$ di media μ_y

  $\bar{X}$ e S_x^2 media e varianza campionarie di $X_1, ..., X_{n_x}$.

  $\bar{Y}$ e S_y^2 media e varianza campionarie di $Y_1, ..., Y_{n_y}$.

  varianza campionaria combinata: $S_p^2 = \dfrac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$

H_0	H_1	Statistica	Regione critica
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$	$\left\lvert \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \right\rvert > t_{n_x+n_y-2,\alpha/2}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$T = \frac{X - Y}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$	$\frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} > t_{n_x+n_y-2,\alpha}$
$\mu_x \geq \mu_y$	$\mu_x < \mu_y$	$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$	$\frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} < -t_{n_x+n_y-2,\alpha}$

- Test χ^2 di buon adattamento

  H_0: la popolazione ha una certa distribuzione assegnata.

  si vuole decidere se accettare o rifiutare H_0

  $C_1, ..., C_k$ classi

  $N_1, ..., N_k$ frequenze assolute delle classi; $n_1, ..., n_k$ frequenze assolute **osservate**; $f_1, \ldots, f_k$ frequenze assolute **attese**. Come le calcolo?

  i) se si vuole testare il buon adattamento a una distribuzione discreta, le classi coincidono con uno o più valori assunti dalla distribuzione incognita; sia $C_1 = \{1\}, ..., C_k = \{k\}$, si assegna $\pi = \Big(\pi(1), ..., \pi(k)\Big)$, densità discreta, e

  $$f_1 = n\pi(1), ..., f_k = n\pi(k)$$

  ii) se si vuole testare il buon adattamento a una distribuzione continua, si avrà $C_1 = [a_0, a_1), C_2 = [a_1, a_2), ..., C_k = [a_{k-1}, a_k)$ si assegna F, funzione di ripartizione, e

  $$f_1 = n\Big(F(a_1) - F(a_0)\Big), ..., f_k = n\Big(F(a_k) - F(a_{k-1})\Big)$$

Statistica:

$$Q = \sum_{j=1}^{k} \frac{(N_j - f_j)^2}{f_j}$$

Regione critica:

se non ci sono parametri da stimare: $q = \sum_{j=1}^{k} \frac{(n_j - f_j)^2}{f_j} > \chi^2_{k-1,\alpha}$

se nella distribuzione verso cui si cerca buon adattamento ci sono r parametri da stimare si ha $q = \sum_{j=1}^{k} \frac{(n_j - f_j)^2}{f_j} > \chi^2_{k-1-r,\alpha}$

Regola empirica di applicabilità: $f_1, \ldots, f_k$ tutte ≥ 1 e almeno l'80% di esse ≥ 5.

Regressione lineare

Modello di regressione lineare semplice: $Y = \alpha + \beta x + e$ ossia per $i = 1, ..., n$

$$Y_i = \alpha + \beta x_i + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

Notazioni

$S_{xx} = \sum_{i=1^n} (x_i - \bar{x})^2 = \sum_{i=1^n} x_i^2 - n\bar{x}^2$

$S_{YY} = \sum_{i=1^n} (Y_i - \bar{Y})^2 = \sum_{i=1^n} Y_i^2 - n\bar{Y}^2, \qquad S_{yy} = \sum_{i=1^n} (y_i - \bar{y})^2 = \sum_{i=1^n} y_i^2 - n\bar{y}^2$

$S_{xY} = \sum_{i=1^n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1^n} x_i Y_i - n\bar{x}\bar{Y}, \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$

• stima di α e β

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

• retta di regressione di y su x : : $y = \hat{\alpha} + \hat{\beta}x$

• Somma dei quadrati residui, con A e B stimatori di α e β:

$$SS_R = \sum_{i=1^n} (Y_i - A - Bx_i)^2 = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}, \qquad \frac{SS_R}{\sigma^2} \sim \chi^2(n-2)$$

• Stima di σ^2: $\hat{\sigma}^2 = \frac{SS_R}{n-2}$,

con SS_R calcolato sulle osservazioni $y_1, ..., y_n$, ossia $\hat{\sigma}^2 = \frac{\frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}}{n-2}$

• Coefficiente di determinazione: $R^2 = 1 - \frac{SS_R}{S_{yy}}$,

con SS_R calcolato sulle osservazioni $y_1, ..., y_n$.
N.B.: $R^2 = r_{x,y}^2$, $r_{x,y}$ coefficiente di correlazione campionaria.

• Intervallo di confidenza per β di livello $100(1-\gamma)\%$:

$$\left(\hat{\beta} - \sqrt{\frac{SS_R}{S_{xx}(n-2)}} t_{n-2,\gamma/2}, \hat{\beta} + \sqrt{\frac{SS_R}{S_{xx}(n-2)}} t_{n-2,\gamma/2}\right)$$

• Verifica dell'ipotesi $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$: si rifiuta H_0 a livello γ se

$$\left| \sqrt{\frac{S_{xx}(n-2)}{SS_R}} \hat{\beta} \right| > t_{n-2,\gamma/2}$$

• Verifica dell'ipotesi $H_0 : \beta = \beta \geq 1$ vs $H_1 : \beta < 1$ (regressione verso la media): si rifiuta H_0 a livello γ se

$$\sqrt{\frac{S_{xx}(n-2)}{SS_R}} (\hat{\beta} - 1) < -t_{n-2,\gamma}$$

Tavola della distribuzione normale

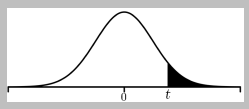
La tabella seguente riporta i valori di $\Phi(z) := \int_{-\infty}^{z} \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\,\mathrm{d}x$, la funzione di ripartizione della distribuzione normale standard $N(0,1)$, per $0 \le z \le 3.5$.

I valori di $\Phi(z)$ per $z < 0$ possono essere ricavati grazie alla formula

$$\Phi(z) \;=\; 1 - \Phi(-z)\,.$$

z	**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

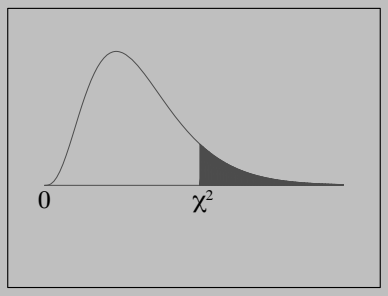
Critical Values for Student's t-Distribution.



df	Upper Tail Probability: $\Pr(T > t)$									
	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850
21	0.859	1.323	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.819
22	0.858	1.321	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.792
23	0.858	1.319	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.768
24	0.857	1.318	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.745
25	0.856	1.316	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.725
26	0.856	1.315	1.706	1.822	1.967	2.056	2.162	2.479	2.779	3.707
27	0.855	1.314	1.703	1.819	1.963	2.052	2.158	2.473	2.771	3.690
28	0.855	1.313	1.701	1.817	1.960	2.048	2.154	2.467	2.763	3.674
29	0.854	1.311	1.699	1.814	1.957	2.045	2.150	2.462	2.756	3.659
30	0.854	1.310	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.646
31	0.853	1.309	1.696	1.810	1.952	2.040	2.144	2.453	2.744	3.633
32	0.853	1.309	1.694	1.808	1.950	2.037	2.141	2.449	2.738	3.622
33	0.853	1.308	1.692	1.806	1.948	2.035	2.138	2.445	2.733	3.611
34	0.852	1.307	1.691	1.805	1.946	2.032	2.136	2.441	2.728	3.601
35	0.852	1.306	1.690	1.803	1.944	2.030	2.133	2.438	2.724	3.591
36	0.852	1.306	1.688	1.802	1.942	2.028	2.131	2.434	2.719	3.582
37	0.851	1.305	1.687	1.800	1.940	2.026	2.129	2.431	2.715	3.574
38	0.851	1.304	1.686	1.799	1.939	2.024	2.127	2.429	2.712	3.566
39	0.851	1.304	1.685	1.798	1.937	2.023	2.125	2.426	2.708	3.558
40	0.851	1.303	1.684	1.796	1.936	2.021	2.123	2.423	2.704	3.551
41	0.850	1.303	1.683	1.795	1.934	2.020	2.121	2.421	2.701	3.544
42	0.850	1.302	1.682	1.794	1.933	2.018	2.120	2.418	2.698	3.538
43	0.850	1.302	1.681	1.793	1.932	2.017	2.118	2.416	2.695	3.532
44	0.850	1.301	1.680	1.792	1.931	2.015	2.116	2.414	2.692	3.526
45	0.850	1.301	1.679	1.791	1.929	2.014	2.115	2.412	2.690	3.520
46	0.850	1.300	1.679	1.790	1.928	2.013	2.114	2.410	2.687	3.515
47	0.849	1.300	1.678	1.789	1.927	2.012	2.112	2.408	2.685	3.510
48	0.849	1.299	1.677	1.789	1.926	2.011	2.111	2.407	2.682	3.505
49	0.849	1.299	1.677	1.788	1.925	2.010	2.110	2.405	2.680	3.500
50	0.849	1.299	1.676	1.787	1.924	2.009	2.109	2.403	2.678	3.496
60	0.848	1.296	1.671	1.781	1.917	2.000	2.099	2.390	2.660	3.460
70	0.847	1.294	1.667	1.776	1.912	1.994	2.093	2.381	2.648	3.435
80	0.846	1.292	1.664	1.773	1.908	1.990	2.088	2.374	2.639	3.416
90	0.846	1.291	1.662	1.771	1.905	1.987	2.084	2.368	2.632	3.402
100	0.845	1.290	1.660	1.769	1.902	1.984	2.081	2.364	2.626	3.390
120	0.845	1.289	1.658	1.766	1.899	1.980	2.076	2.358	2.617	3.373
140	0.844	1.288	1.656	1.763	1.896	1.977	2.073	2.353	2.611	3.361
180	0.844	1.286	1.653	1.761	1.893	1.973	2.069	2.347	2.603	3.345
200	0.843	1.286	1.653	1.760	1.892	1.972	2.067	2.345	2.601	3.340
500	0.842	1.283	1.648	1.754	1.885	1.965	2.059	2.334	2.586	3.310
1000	0.842	1.282	1.646	1.752	1.883	1.962	2.056	2.330	2.581	3.300
∞	0.842	1.282	1.645	1.751	1.881	1.960	2.054	2.326	2.576	3.291
	60%	80%	90%	92%	94%	95%	96%	98%	99%	99.9%
	Confidence Level									

Note: $t(\infty)_{\alpha/2} = Z_{\alpha/2}$ in our notation.

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Capitolo 1

Statistica Descrittiva

1.1 Introduzione

Statistica Descrittiva, Statistica Inferenziale e Probabilità

La statistica è l'arte di imparare dai dati. Possiamo suddividerla in:

- **Statistica Descrittiva**: *Descrive* e riassume i dati
- **Statistica Inferenziale**: *Trae* conclusioni dai dati

Per poter trarre conclusioni dai dati, bisogna tenere conto del ruolo che gioca il caso. Definiamo quindi la **probabilità** come la descrizione matematica di eventi *casuali*.

1.2 Descrivere i dati

1.2.1 Frequenze, Istogrammi, Classi

Frequenza Assoluta e Relativa

Misuriamo una certa variabile in un campione, ottenendo un insieme di dati. Se i dati non sono distinti, ovvero *abbiamo ripetizioni*, possiamo riassumerli in una **tabella delle frequenze**. Possiamo definire dunque:

- **Frequenza Assoluta** f_i := numero di volte in cui i compare nel campione di dati.
- **Frequenza Relativa** $p_i := {}^{f_i}/_{N}$ = numero di volte in cui compare i rispetto al totale.

Esempio: Marta intervista i suoi N = 20 compagni di classe e chiede la squadra di calcio preferita, ottenendo le risposte:
Juve, Milan, Inter, Atalanta, Juve, Milan, Nessuna, Nessuna, Inter, Milan, Juve, Nessuna, Atalanta, Juve, Nessuna, Milan, Inter, Milan, Nessuna, Nessuna.

Squadre di Calcio		
Valori	Frequenze assolute	Frequenze relative
Juve	4	$4/20 = 0.20$
Milan	5	$5/20 = 0.25$
Inter	3	$3/20 = 0.15$
Nessuna	6	$6/20 = 0.30$
Atalanta	2	$2/20 = 0.10$

Istogramma

É utile rappresentare le frequenze mediante un grafico a barre detto **istogramma**.

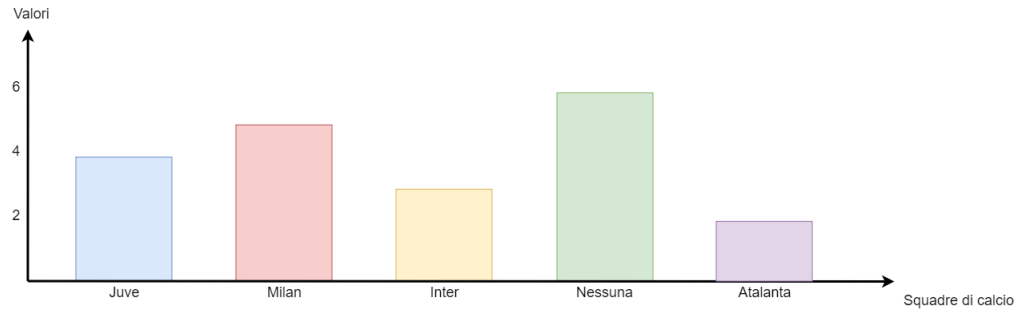


Figura 1.1: Istogramma sulle frequenze assolute delle squadre di calcio

Classi

Può capitare di avere insiemi di dati che assumono un numero elevato di valori distinti. In tal caso conviene *suddividere* i valori assunti in intervalli detti **classi**.

Esempio: Vengono misurati i livelli di colesterolo nel sangue di un insieme di N = 40 individui:

213 174 193 796 220 183 194 200 192 200 200 199 178 183 188 193 187
181 193 205 196 211 202 213 216 206 195 191 171 194 184 191 221 212
221 204 204 191 183 227

Molti valori sono distinti e dunque hanno $f_i = 1$. Scegliamo le classi:

$[170, 180)$ $[180, 190)$ $[190, 200)$ $[200, 210)$ $[210, 220)$ $[220, 230)$

Livelli di colesterolo		
Valori	Frequenze assolute	Frequenze relative
$[170, 180)$	3	$3/40 = 7.5$
$[180, 190)$	7	$7/40 = 17.5$
$[190, 200)$	13	$13/40 = 32.5$
$[200, 210)$	8	$8/40 = 20$
$[210, 220)$	5	$5/40 = 12.5$
$[220, 230)$	4	$4/40 = 10$

1.2.2 Dati Bivariati

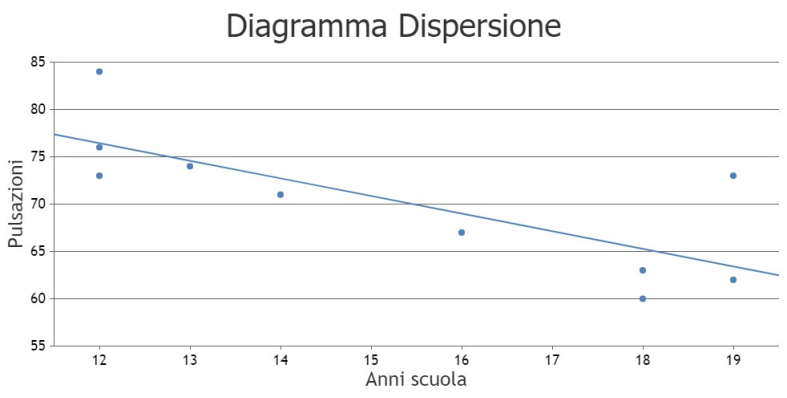
I **dati bivariati** ci permettono di mostrare *due variabili* per ogni singolo elemento dell'insieme. Per un elemento i, indichiamo con x_i la prima variabile e con y_i la seconda.

Diagramma Di Dispersione

Un **diagramma di dispersione** rappresenta i punti (x_i, y_i), in modo da evidenziare una possibile *correlazione* tra i due valori.

Esempio: Rileviamo il numero di anni di scuola (prima variabile) e le pulsazioni a riposo (seconda variabile) in un campione di N=10 individui. I dati (xi , yi) sono

(12,73) (16,67) (13,74) (18,63) (19,73) (12,84) (18,60) (19,62) (12,76) (14,71)



Possiamo evidenziare una correlazione negativa tra i due valori.

1.3 Riassumere i dati

Una **statistica** campionaria riassume l'insieme di dati mediante una quantità numerica.

1.3.1 Indici di Posizione

Un **indice di posizione** *sintetizza la posizione* di una distribuzione sostituendo l'insieme dei dati con un unico valore tale da fornire una rappresentazione globale.

Media Campionaria

Per definire il valore medio dell'insieme dei dati, definiamo la **media campionaria** come

$$\overline{x} := \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

In generale, sapendo le frequenze assolute dei valori, possiamo scrivere la media campionaria come

$$\overline{x} = \frac{z_1 \cdot f_1 + z_2 \cdot f_2 + \cdots + z_M \cdot f_M}{N}$$

con $N = f_1 + f_2 + \cdots + f_M$, dove z_i sono i valori e f_i sono le frequenze assolute associate.

Osservazione: La media è lineare: $\overline{y} = a \cdot \overline{x} + b$

Mediana Campionaria

Un'altra misura del centro dell'insieme dei dati è la **mediana campionaria**, ovvero quel valore che, ordinati i dati, si trova in *posizione centrale*. In base al numero di dati, si calcola in due modi:

- **N dispari**: la mediana è il dato di posto $\frac{N+1}{2}$:

$$m := x_{\left(\frac{N+1}{2}\right)}$$

- **N pari**: la mediana è la media aritmetica tra il dato di posto $\frac{N}{2}$ e quello di posto $\frac{N}{2} + 1$:

$$m := \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2}$$

Percentile Campionario

Fissiamo un numero $k \in [0, 100]$. Definiamo il **k-esimo percentile campionario** il valore t per cui:

- almeno il $k\%$ dei dati è $\leq$ t
- almeno il $(100 - k)\%$ dei dati è $\geq$ t

Quartili

I casi più importanti sono:

- **Primo Quartile** q_1: $p = \frac{1}{4}$, $k = 100p$
- **Secondo Quartile** q_2: $p = \frac{1}{2}$, $k = 100p$
- **Terzo Quartile** q_3: $p = \frac{3}{4}$, $k = 100p$

Come per la mediana, si calcola in due modi:

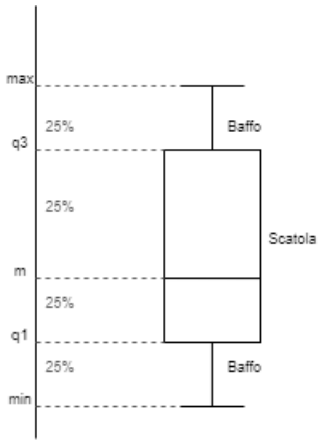
- Se $N \cdot p$ non è intero, $t = x_{(i)}$ è il dato di posizione i definito come l'intero successivo a $N \cdot p$
- Se $N \cdot p$ è intero, $t = \frac{x_{(N \cdot p)} + x_{(N \cdot p+1)}}{2}$

Esempio: Ai 1000 abitanti di un piccolo comune viene chiesto di esprimere un giudizio su un nuovo servizio comunale, usando una scala da 0 a 4. 251 persone hanno votato 0, 260 persone hanno votato 1, 80 persone hanno votato 2, 154 persone hyanno votato 3 mentre 255 persone hanno votato 4
Vogliamo calcolare i 3 indici di posizione: Media, Mediana, Quartili.

- **Media**: $\overline{x} = \dfrac{0 \cdot 251 + 1 \cdot 260 + 2 \cdot 80 + 3 \cdot 154 + 4 \cdot 255}{1000} = 1,902 \simeq 1,9$
- **Mediana**: $m = \dfrac{x_{(500)} + x_{(501)}}{2} = \dfrac{1+1}{2} = 1$
- **Primo quartile**: $1000 \cdot \dfrac{1}{4} = 250$, $q_1 = \dfrac{x_{(250)} + x_{(251)}}{2} = \dfrac{0+0}{2} = 0$
- **Secondo quartile**: mediana $= 1$
- **Terzo quartile**: $1000 \cdot \dfrac{3}{4} = 750$, $q_3 = \dfrac{x_{(750)} + x_{(751)}}{2} = \dfrac{4+4}{2} = 4$

Box Plot

Una rappresentazione grafica di mediana e quartili viene fornita dal **box plot**:



1.3.2 Indici di Dispersione

Un **indice di dispersione** descrive la *variabilità di distribuzione* quantitativa dei dati, ovvero quanto i valori presenti distano da un valore centrale.

Scarti

Gli **scarti** sono la distanza di ogni singolo elemento dal valore medio. La somma di tutti gli scarti è nulla.

$$w := (x_i - \overline{x})$$

Varianza Campionaria

Considerando gli scarti elevati al quadrato e facendone una sorta di media, otteniamo la **varianza campionaria** come:

$$s^2 := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

Deviazione Standard

Per ottenere una statistica omogenea ai dati, si definisce **deviazione standard** come:

$$s := \sqrt{s^2} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2}$$

Scarto Interquartile

Un altro indicatore di dispersione rispetto alla mediana m è lo **scarto interquartile**. Per costruzione, questo intervallo contiene almeno il 50% dei dati. Il suo valore è:

$$IQR = \Delta := q_3 - q_1$$

Osservazione: La varianza e la deviazione standard non sono lineari.

1.3.3 Correlazione

Consideriamo un insieme di dati bivariati. Vogliamo quantificare la correlazione tra le due variabili x e y, ossia la *tendenza* per cui a valori di x grandi corrispondo valori di y grandi o piccoli. Definiamo quindi il **coefficiente di correlazione lineare**:

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{(N-1)\cdot s_x \cdot s_y} \quad \text{oppure} \quad r = \frac{\sum_{i=1}^{N} x_i \cdot y_i - N \cdot \overline{x} \cdot \overline{y}}{(N-1)\cdot s_x \cdot s_y}$$

Il coefficiente di correlazione lineare può assumere valori $\in [-1, 1]$. Nello specifico diremo che:

- $|r| \gtrsim 0,7$ avremo correlazione *significativa*
- $|r| \lesssim 0,3$ avremo correlazione *debole*.

Esercizio Completo: Prendiamo un campione di N = 10 elementi e assegnamo x = numero anni di scuola e y = pulsazioni. Otteniamo i seguenti valori

(12, 73) (16, 67) (13, 74) (18, 63) (19, 73) (12, 84) (18, 60) (19, 62) (12, 76) (14, 71)

Calcoliamo $\overline{x}, m, s^2, s, r$.

- $\overline{x} = \frac{12+16+13+18+19+19+12+18+19+12+14}{10} = 15,3$
- $s_x^2 = \frac{(12+16+13+18+19+19+12+18+19+12+14)^2 - 10\cdot(15,3)^2}{9} \simeq 9,12$
- $s_x = \sqrt{s_x^2} \simeq 3,02$
- $\overline{y} = \frac{73+67+74+63+73+84+60+62+76+71}{10} = 70,3$
- $s_y^2 = \frac{(73+67+74+63+73+84+60+62+76+71)^2 - 10\cdot(70,3)^2}{9} \simeq 54,23$
- $s_y = \sqrt{s_y^2} \simeq 7,36$
- $\sum_{i=1}^{N} x_i \cdot y_i = 12\cdot 73 + 16\cdot 67 + 13\cdot 74 + \cdots + 14\cdot 71 = 10603$
- $r = \frac{10603 - 10\cdot 15,3\cdot 70,3}{9\cdot 3,02\cdot 7,36} \simeq -0,76$

Il valore di r conferma che x e y mostrano una significativa correlazione negativa

Capitolo 2

Spazi di Probabilità

2.1 Introduzione

Il **calcolo delle probabilità** è una teoria matematica che permette di descrivere gli *esperimenti aleatori*, ovvero fenomeni il cui esito non è prevedibile con certezza a priori.

2.2 Assiomi della Probabilità

Spazio Campionario, Evento, Probabilità

La descrizione matematica di esperimento aleatorio si descrive in 3 passi:

1. **Spazio Campionario**: Insieme Ω che contiene tutti i possibili esiti dell'esperimento.
2. **Eventi**: Affermazioni sull'esito dell'esperimento aleatorio. $A \subseteq \Omega$.
3. **Probabilità**: Funzione P che associa a ogni evento $A \subseteq \Omega$ un valore $\in [0, 1]$ che soddisfa opportune proprietà. Ci sono almeno due interpretazioni su che cos'è P(A):
   - **Soggettivista**: P(A) = prezzo equo di una scommessa che paga 1 se si verifica A, altrimenti 0.
   - **Frequentista**: P(A) = frazione asintotica di volte in cui si verifica A ripetendo l'esperimento.

In ogni caso la probabilità deve soddisfare due proprietà:

- $P(\Omega) = 1$
- Se A e B sono eventi disgiunti, cioè $A \cap B = \emptyset$, allora:

$$P(A \cup B) = P(A) + P(B)$$

Esempio: Se prendiamo in considerazione un dado a sei facce, avremo:

- Spazio Campionario: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Evento: A = esce un numero pari = {2,4,6}
- Probabilità di A= 0.5 = 50%

La coppia (Ω, P) è detta **spazio di probabilità**.

Indichiamo con $|A|$ la cardinalità di un insieme A. La **probabilità uniforme** su un insieme finito Ω si definisce come:

$$P(A) := \frac{|A|}{|\Omega|}$$

Abbiamo dunque che per ogni $w \in \Omega$, $P(w) = \frac{1}{|\Omega|}$

Proprietà di base

Fissiamo uno spazio di probabilità (Ω, P). Valgono le seguenti proprietà:

1. **Insieme Vuoto**: $P(\emptyset) = 0$
2. **Regola del complementare**: $P(A^C) = 1 - P(A)$
3. **Regola dell'addizione di probabilità**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. **Monotonia**: se $A \subseteq B$ allora $P(A) \leq P(B)$

2.3 Calcolo Combinatorio

In uno spazio di probabilità uniforme, calcolare una probabilità significa contare gli elementi di un insieme. Contare è un problema non banale per insiemi grandi. Le tecniche di conteggio formano il **Calcolo combinatorio**.

Principio fondamentale: Consideriamo un esperimento costituito da due parti:

1. prima parte: n esiti possibili
2. seconda parte: m esiti possibili

Allora l'esperimento totale può avere $n \cdot m$ esiti possibili.

Esempio: Se lancio due dadi a 6 facce ho $\Omega = 6^2 = 36$ esiti possibili, se lancio tre dadi a 6 facce ho $\Omega = 6^3 = 216$ esiti possibili ecc.

2.3.1 Disposizioni con Ripetizione

Le **disposizioni con ripetizioni** sono sequenze ordinate di k elementi, *anche ripetuti*, scelti tra n possibili. Sono in numero

$$n^k$$

Esempio: Estraggo casualmente k persone: qual è la probabilità che siano nate tutte e k in primavera?
$|\Omega| = 365^k$
A = "Tutti nati in primavera " = "20 marzo - 20 giugno" = 92 giorni
$|A| = 92^k$
$\mathrm{P(A)} = (\frac{92}{365})^k \simeq \frac{1}{4^k}$

2.3.2 Disposizioni Semplici

Le **disposizioni semplici** sono sequenze ordinate di k elementi **distinti** scelti tra n possibili. Sono in numero

$$\frac{n!}{(n-k)!}$$

Osservazione: Se n=k si parla di permutazioni, in questo caso sono in numero n!

Esempio: Estraggo casualmente k persone: qual è la probabilità che almeno due abbiano lo stesso compleanno?
$|\Omega| = 365^k$
A = "almeno due persone hanno lo stesso compleanno"
A^C = "tutti hanno compleanni distinti"
$P(A^C) = 365 \cdot 364 \cdots (365 - k + 1)$
$P(A) = 1 - P(A^C) = 1 - \frac{|A^C|}{|\Omega|} = 1 - \frac{365 \cdot 364 \cdots (365 - k + 1)}{365^k} = 1 - \frac{365}{365} \cdot \frac{364}{365} \cdots \frac{(365 - k + 1)}{365}$

- con K=10 persone avremo il 12% di trovare due persone con lo stesso compleanno.
- con K=23 persone avremo il 50% di trovare due persone con lo stesso compleanno.
- con K=50 persone avremo il 97% di trovare due persone con lo stesso compleanno.

2.3.3 Combinazioni

Fino ad ora abbiamo considerato l'ordine importante. Ad esempio, nel lancio di due dadi, (2,5) $\neq$ (5,2). Le combinazioni si possono ottenere dalle disposizioni semplici dimenticando l'ordine degli elementi.
Le **combinazioni** sono collezioni *non ordinate* di k elementi distinti scelti tra n possibili. Sono in numero

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Esempio: In una mano a Poker un giocatore riceve 5 carte su un mazzo di 52. Le possibili combinazioni sono quindi

$$\binom{52}{5} = \frac{52!}{5! \cdot 47!} = 2.598.960$$

2.4 Probabilità Condizionata

Consideriamo uno spazio di probabilità (Ω, P). Consideriamo un evento $A \subseteq \Omega$ con probabilità P(A). Supponiamo di ricevere informazioni su un evento B che si è verificato.
La **probabilità condizionata** è l'aggiornamento della probabilità di A dopo un *informazione aggiuntiva.* la probabilità condizionata di A dato B si scrive come

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

Esempio: Qual è la probabilità che la somma di due dadi a 6 facce valga 4, sapendo che il primo dado vale 2?
$|\Omega| = 6^2 = 36$
A = "la somma vale 4" = (1,3) (2,2) (3,1), $|A| = 3$
B = "il primo vale 2" = (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), $|B| = 6$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{|A \cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{|A \cap B|}{|B|} = \frac{1}{6} \simeq 16,7\%$$

Proprietà

- **Regola del prodotto**: $P(A \cap B) = P(B) \cdot P(A|B)$
- **Formula di disintegrazione**: $P(A) = P(A \cap B) + P(A \cap B^C)$
- **Formula delle probabilità totali**: $P(A) = P(A|B) \cdot P(B) + P(A|B^C) \cdot P(B^C)$
- **Secondo elemento fissato**: $P(*|B)$ è una probabilità: $P(A^C|B) = 1 - P(A|B)$
- **Formula di Bayes**: $P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$

Esempio: Per rilevare la presenza di un virus viene effettuato un test con le seguenti caratteristiche:

- Sensibilità: se il virus è presente, il test dà esito positivo al 99%
- Specificità: se il virus è assente, il test dà esito negativo al 99.7%
- Prevalenza: è noto che 4 persone su 1000 hanno il virus.

Estraendo un cittadino a caso, qual è la probabilità che l'esito sia positivo? E qual è la probabilità che abbia effettivamente il virus se l'esito è positivo?
Introduciamo gli eventi e le relative probabilità:

- A: "il test dà esito positivo"
- B: "l'individuo ha il virus"
- $P(A|B) = 0.99$
- $P(A^C|B^C) = 0.997$
- $P(B) = 0.004$

Dai seguenti valori ricaviamo che:

- $P(A|B^C) = 1 - P(A^C|B^C) = 1 - 0.997 = 0.003$
- $P(B^C) = 1 - P(B) = 1 - 0.004 = 0.996$

Ora possiamo usare la formula delle probabilità totali per trovare P(A), ovvero la probabilità che il test sia positivo:

$$P(A) = P(A|B) \cdot P(B) + P(A|B^C) \cdot P(B^C) \simeq 0,004 + 0,003 = 0.007$$

Adesso vogliamo sapere $P(B|A)$, ovvero la probabilità che abbia effettivamente il virus se risulta positivo al test, e per farlo usiamo la Formula di Bayes:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} = \frac{0.99 \cdot 0.004}{0.007} \simeq 57\%$$

2.5 Indipendenza di eventi

Due eventi sono **indipendenti** se al verificarsi di uno, la probabilità che si verifichi l'altro non cambia:

$$P(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Esempio: Consideriamo
A = "il primo dado vale 2" e C = "somma = 7" e D = "somma = 4".
A = $\{(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\}$ $|A| = 6$ $P(A) = 1/6$
C = $\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$ $|C| = 6$ $P(C) = 1/6$
D = $\{(1,3),(2,2),(3,1)\}$ $|D| = 3$ $P(D) = 1/12$

$A \cap C = \{(2,5)\}$ $|A \cap C| = 1$ $P(A \cap C) = 1/36$
Dato che $P(A \cap C) = P(A) \cdot P(C)$ A e C sono indipendenti

$A \cap D = \{(2,2)\}$ $|A \cap D| = 1$ $P(A \cap C) = 1/36$
Dato che $P(A \cap D) \neq P(A) \cdot P(D)$ A e C sono dipendenti

Osservazione: Se A e B sono indipendenti, lo sono anche A e B^c, A^c e B, A^c e B^c
Osservazione: Eventi indipendenti $\neq$ eventi disgiunti! Due eventi indipendenti non possono essere disgiunti.

Estensioni

Tre eventi A,B,C si dicono indipendenti se valgono *tutte* le seguenti regole:

- $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$
- $P(A \cap B) = P(A) \cdot P(B)$
- $P(A \cap C) = P(A) \cdot P(C)$
- $P(B \cap C) = P(B) \cdot P(C)$

Capitolo 3

Variabili aleatorie

3.0.1 Introduzione

Consideriamo uno spazio di probabilità (Ω, P). Spesso non siamo interessati a tutti i dettagli dell'esito dell'esperimento, ma solo a una *quantità determinata dall'esito* dell'esperimento. Tale quantità è detta **variabile aleatoria**. Una variabile aleatoria può essere descritta come una **funzione** $\Omega \to \mathbb{R}$

Osservazione: Evento e variabile aleatoria sono diverse ma in relazione. Ogni variabile aleatoria mi permette di determinare più eventi. Sia X una variabile aleatoria e x un suo possibile valore: $\{X = x\}$ è un evento.

3.1 Variabili Aleatorie Discrete

Una variabile aleatoria X si dice **discreta** se i valori che può assumere sono un insieme finito o un insieme infinito numerabile.
Ad ogni variabile aleatoria discreta X possiamo associare un indice chiamato **Densità Discreta**:

$$P_X(x_i) := P(X = x_i)$$

Proprietà

- Se x non è un valore assunto da X, si pone $P_X(x) := 0$
- P_X è una funzione da $\mathbb{R}$ a $[0, 1]$
- $P_X(x_i) \geq 0$

- $\sum_{i \geq 1} P_X(x_i) = 1$

Osservazione: La densità discreta ci risulta utile quando ci interessa sapere la densità di un sottoinsieme:

$$P(X \in B) = \sum_{x_i \in B} P_X(x_i)$$

Esempio: Una pasticceria prepara 3 torte al giorno. Sappiamo che:

- il 20% dei giorni nessun cliente ordina una torta
- il 30% dei giorni un cliente ordina una torta
- il 35% dei giorni due clienti ordinano una torta
- il restante dei giorni tre o più clienti ordinano una torta

Sia X il numero di torte invendute.

1. Qual è la densità discreta di X?
2. Qual è la probabilità q che il numero di torte invendute sia pari?

Sappiamo che:

- $P_X(3) = P(X = 3) = 20\%$
- $P_X(2) = P(X = 2) = 30\%$
- $P_X(1) = P(X = 1) = 35\%$

Ricaviamo che

$$P_X(0) = P(X = 0) = 1 - P(X = 3) - P(X = 2) - P(X = 1) = 15\%$$

La probabilità che il numero di torte invendute sia pari è

$$P_X(0) + P_X(2) = 45\%$$

3.1.1 Valore Medio di Variabile Aleatoria Discreta

Sia X una variabile aleatoria discreta. Si definisce **valore medio di X**:

$$E[X] := \sum_{i=1}^{N} x_i \cdot p_X(x_i)$$

Proprietà Valore Medio

Per ogni variabile aleatoria X,Y e per ogni costante c $\in \mathbb{R}$ avremo:

- *Traslazione*: E[X + c] = E[X]
- *Moltiplicazione*: E[c · X] = c · E[X]
- *Somma*: E[X + Y] = E[X] + E[Y]
- *Formula di trasferimento*: $E[f(x)] = \sum_{i=1}^{N} f(x_i) \cdot P_X(x_i)$

Osservazioni

- Il valore medio è un operatore *lineare*: $E[Z] = X + c$.
- *Momento Secondo*: $E[X^2] = \sum_{i=1}^{N} (x_i)^2 \cdot P_X(x_i)$
- Il valore medio E[X] non è necessariamente un valore x_i assunto da X
- Il valore medio E[X] è quel valore tale per cui se ripetiamo un esperimento più volte, la media di tale esperimenti si avvicinerà a E[X]. Questo fenomeno è detto **Legge dei Grandi Numeri**:

$$\frac{\sum_{i=1} x_i}{N} \simeq E[X]$$

3.1.2 Varianza e Deviazione Standard

Sia X una variabile aleatoria discreta e $\mu = E[X]$. Definiamo la **varianza** come:

$$Var[X] := E[(X - \mu)^2]$$

A livello di calcolo, risulta più semplice utlizzare la formula:

$$Var[X] = E[X^2] - E[X]^2$$

Definiano la *deviazione standard* come:

$$Sd[X] := \sqrt{Var[X]}$$

Propietà della Varianza

Per ogni variabile aleatoria X e per ogni costante c $\in \mathbb{R}$ avremo:

· Var[X + c] = Var[X]

· Var[c · X] = c^2 Var[X]

Esempio Completo: Sia X il numero di figli maschi in una famiglia con due figli. Trovare: $P_X, E[X], Var[X], Sd[X]$
Spazio Campionario

$$\Omega = \{MM, MF, FM, FF\}$$

Densità Discreta

$$P_X(2) = \frac{1}{4} \qquad P_X(1) = \frac{1}{2} \qquad P_X(0) = \frac{1}{4}$$

Valore Medio

$$E[X] = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

Momento Secondo

$$E[X^2] = 0^2 \cdot P(X = 0) + 1^2 \cdot P(X = 1) + 2^2 \cdot P(X = 2) = 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = 1 \cdot \frac{3}{2}$$

Varianza

$$\text{Var[X]} = E[X^2] - E[X]^2 = \frac{3}{2} - 1^2 = \frac{1}{2}$$

Deviazione Standard

$$\sqrt{Var[X]} = \frac{1}{\sqrt{2}}$$

Variabili Indipendenti

Siano X e Y due variabili aleatorie. Il valore della loro somma dipende da come sono "legate":

· Consideriamo Y = X. Allora Var[Y] = Var[X], quindi

$$Var[X + Y] = Var[2X] = 4 \cdot Var[X]$$

· Consideriamo invece Y = -X. Allora

$$Var[X + Y] = Var[X - X] = Var[0] = 0$$

Gli esempi precedenti sono casi estremi di dipendenza. Definiamo invece X e Y come variabili **indipendenti** se gli eventi {X = x} e {Y = y} sono indipendenti, ovvero

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

Definizione: Se X e Y sono indipendenti allora

$$Var[X + Y] = Var[X] + Var[Y]$$

3.2 Distribuzioni Notevoli Discrete

Consideriamo una variabile aleatoria X in un certo esperimento aleatorio: $X : \Omega \to \mathbb{R}$. Possiamo calcolare P($X \in A$) per ogni $A \subseteq \mathbb{R}$. L'insieme di tali probabilità definisce la **Distribuzione della variabile aleatoria X**.

Osservazione: Per variabili aleatorie discrete, la distribuzione di X è determinata dalla densità discreta e quindi, con abuso di notazione, si può dire che la distribuzione è la densità discreta.

3.2.1 Bernoulli

Variabile aleatoria X che può assumere soltanto i valori 0 e 1. Scriveremo X $\sim$ Be(p). Sia p := (X = 1). Dato che

$$\sum_{i=1}^{N} P_X(x_i) = P_X(0) + P_X(1) = 1$$

si ottiene:

$$P(X = x) = \begin{cases} p & \text{se x} = 1 \\ (1 - p) & \text{se x} = 0 \end{cases} \tag{3.1}$$

Allora, nel caso della variabile aleatoria di Bernoulli troviamo che:

- Il **valore medio** E[X] = p.
- La **varianza** Var[X] = p(1 - p)

3.2.2 Binomiale

Consideriamo un esperimento aleatorio costituito da prove ripetute ed indipendenti dove abbiamo solo due esiti. Siano n il numero di prove e p la probabilità di successo di ciascuna e X il numero di successi.
La distribuzione di X è detta **binomiale** di parametri n e p indicata con X $\sim$ Bin(n, p). La densità discreta è data da:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

dove:

- P(X = k) è esattamente il numero k di successi in n prove
- p^k è la probabilità di k successi fissati
- $(1-p)^{n-k}$ è la probabilità di (n - k) insuccessi fissati
- $\binom{n}{k}$ scelte di quali prove hanno successo

Introduciamo le variabili:

$$X_i = \begin{cases} 1 & \text{se la i-esima prova ha successo} \\ 0 & \text{se la i-esima prova non ha successo} \end{cases} \tag{3.2}$$

Possiamo allora scrivere:

$$X = \sum_{i=1}^{N} X_i$$

Avremo che ogni $X_i \sim Be(p)$ e quindi sappiamo valore medio e varianza. Otteniamo allora:

- **Valore medio** E[X] $=n \cdot p$
- **Varianza** Var[X] $= n \cdot p \cdot (1-p)$

Esempio: Sia X numero di figli maschi in una famiglia con 2 figli. Abbiamo $\Omega = \{MM, MF, FM, FF\}$ e avremo :

$$P_X(0) = \frac{1}{4} \qquad P_X(1) = \frac{1}{2} \qquad P_X(2) = \frac{1}{4}$$

Allora X $\sim$ Bin(2; $\frac{1}{2}$):

$$P_X(k) = \binom{2}{k} \cdot \frac{1}{2}^k \cdot \frac{1}{2}^{2-k} = \frac{2}{k! \cdot (2-k)!} \cdot \frac{1}{2^2} = \frac{1}{2} \cdot \frac{1}{k! \cdot (2-k)!}$$

Infatti, se sostituiamo k=0, ritroveremo $P_X(0) = \frac{1}{4}$, se k=1 avremo $P_X(1) = \frac{1}{2}$ e con k=3 avremo $P_X(2) = \frac{1}{4}$

3.2.3 Poisson

Una variabile aleatoria X si dice di Poisson di parametro $\lambda \in (0, \infty)$ ovvero X $\sim Pois(\lambda)$ se X(Ω) = $\mathbb{N}_0$ e

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

Approssimazione di Poisson: Possiamo ottenere $X \sim Pois(\lambda)$ attraverso una variabile aleatoria binomiale $Y \sim Bin(n, p)$ quando:

$$n \to \infty \qquad p \to 0 \qquad \text{allora} \qquad n \cdot p = \lambda$$

Allora avremo che:

- Il **Valore medio** E[X] = λ
- La **Varianza** Var[X] = λ

Osservazione: Le variabili aleatorie di Poisson sono approssimazioni per variabili aleatorie che contano il numero di successi quando si considera una grande quantità di prove la cui probabilità di successo è piccola.

Esempio: In un ospedale nascono mediamente 2,2 bambini ogni giorno. Qual è la probablità che nessun bambino nasca in ogni giorno. E qual è la probabilità che ne nascono più di 3?
Sia X il numero di nascite in un giorno. Supponiamo $X \sim Pois(\lambda)$. Sappiamo che E[X] = λ e per ipotesi E[X] = 2,2. calcoliamo:

- P(X = 0) = $e^{-\lambda} \cdot \frac{e^0}{0!} = e^{-2,2} \simeq 11\%$
- P(X > 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3) = $1 - e^{-\lambda}(1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{6}) = 18\%$

3.2.4 Geometrica

Una variabile aleatoria X si dice **Geometrica** di parametro p $\in$ (0, 1] e si scrive X $\sim$ Geo(p) se

$$P(X = k) = p \cdot (1 - p)^{k-1}$$

Osservazione: Se p è 1, allora possiamo dire che, per quanto piccola sia la probabilità, l'evento prima o poi accadrà.

Possiamo ottenere una variabile aleatoria geometrica partendo da una successione di prove ripetute dove consideriamo T l'istante del primo successo. Avremo che

$$P(T = k) = P(X_1 = 0, ..., X_{K-1} = 0, X_K = 1) = (1 - p)^{k-1} \cdot p$$

Allora troviamo:

- il **Valore medio** E[X]= $\frac{1}{p}$
- la **Varianza** Var[X] = $\frac{1-p}{p^2}$

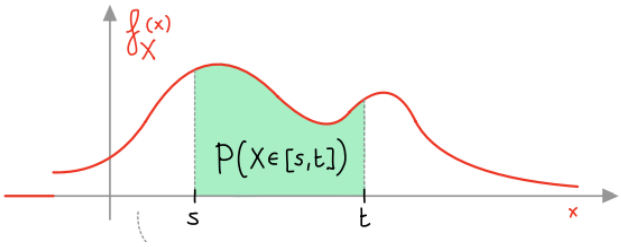
Osservazione: Possiamo calcolare la probabilità di coda:

$$P(T > n) = (1 - p)^n \qquad P(T \leq n) = \sum_{k=1}^{N} P_T(k) = (1 - p)^n$$

3.3 Var. Aleatorie Assolutamente Continue

Consideriamo una classe complementare di variabili aleatorie, dette **assolutamente continue**, che assumono un insieme *infinito più che numerabile di valori*, come ad esempio un intervallo in $\mathbb{R}$.
Una variabile X è assolutamente continua se la sua distribuzione è determinata da una funzione $f_X(x)$ a valori positivi detta **densità della variabile aleatoria X** nel modo seguente:

$$P(X \in [s,t]) = \int_s^t f_X(x)dx$$



Osservazione: Troviamo una analogia tra una variabile aleatoria discreta ed una assolutamente continua:

$$\text{V.A. discreta: } \sum_{x \in [s,t]} P_X(x_i) \qquad \text{V.A. assolutamente continua: } \int_s^t f_X(x)$$

Notiamo anche delle differenze importanti: se X è assolutamente continua $\forall x \in \mathbb{R} : P(X = x) = 0$. In particolare, tranne per $f_X(x) = 0$:

$$f_X(x) \neq P(X = x)$$

Densità

La **densità** di una variabile aleatoria assolutamente continua X è una funzione $f_X : \mathbb{R} \to \mathbb{R}$ integrabile tale che:

$$f_X(x) \geq 0 \quad \forall x \in \mathbb{R} \qquad \int_{-\infty}^{+\infty} f_X(x)dx = 1$$

Valore medio e varianza di v.a. assolutamente continue

Le definizioni di valore medio e varianza di v.a. assolutamente continue ricalcano quelle delle v.a. discrete:

- $E[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx$
- $Var[X] = E[X^2] - E[X]^2$
- $Sd[X] = \sqrt{Var[X]}$

Anche le proprietà definite per le variabili discrete continuano a valere.

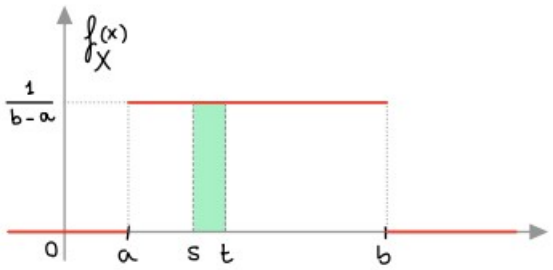
3.3.1 Uniforme Continua

Una variabile aleatoria X è **uniforme continua in [0,1]** e si indica con $X \sim U(0,1)$ se è definita da una funzione:

$$f_X(x) = \begin{cases} c = 1 & \text{se } x \in [0,1] \\ 0 & \text{se } x \notin [0,1] \end{cases} \tag{3.3}$$

Una variabile aleatoria X è **uniforme continua in [a,b]** e si indica con $X \sim U(a,b)$ se è definita da una funzione:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} = 1 & \text{se } x \in [a,b] \\ 0 & \text{se } x \notin [a,b] \end{cases} \tag{3.4}$$



Dato un intervallo $[s,t] \subseteq [a,b]$ avremo che

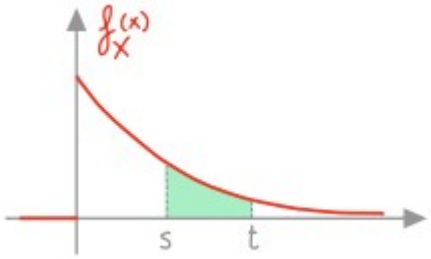
$$P(X \in [s,t]) = \int_s^t f(x)dx + \frac{1}{b-a} = \int_s^t 1 \; dx = \frac{t-s}{b-a}$$

Osservazione: Poiché la variabile aleatoria uniforme continua assume un'infinita quantità di valori in un intervallo continuo [0,3], la probabilità di ottenere un valore specifico x è zero. In altre parole, la probabilità di ottenere un risultato esatto in una variabile aleatoria continua è sempre zero.

3.3.2 Esponenziale

Una variabile aleatoria **esponenziale** è spesso utilizzata per descrivere il tempo tra gli arrivi di eventi casuali indipendenti di un processo di Poisson. Misuriamo il tempo medio di evento come τ, avremo che $\lambda = \frac{1}{\tau}$. La funzione che la descrive è

$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases} \tag{3.5}$$



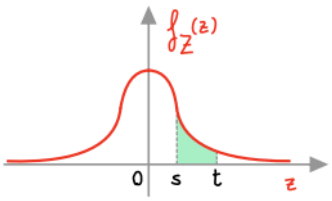
Una variabile aleatoria X con tale densità è detta esponenziale di parametro $\lambda \in (0, \infty)$ e si scrive X $\sim$ Exp(λ). Avremo che:

- $P(X \in [s,t]) = \int_s^t f_X(x)dx = e^{-\lambda \cdot s} - e^{-\lambda \cdot t}$
- $E[X] = \frac{1}{\lambda}$
- $Var[X] = \frac{1}{\lambda^2}$

3.3.3 Normale

Abbiamo già visto due classi notevoli assolutamente continue: uniforme continua ed esponenziale. Consideriamo ora la più importante: **variabili aleatorie normali** (o gaussiane).
Una variabile aleatoria X si dice **normale standard** e si indica con $Z \sim N(0,1)$ se è assolutamente continua e ha densità:

$$f_Z(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{z^2}{2}\right)}$$



Una variabile aleatoria normale standard avrà:

- $\mathrm{E}[\mathrm{Z}] = 0$
- $\mathrm{Var}[\mathrm{Z}] = 1$
- $\mathrm{P}(\mathrm{Z} \in [s,t]) = \int_s^t f_Z(z)dz$

Purtroppo l'integrabile non è calcolabile, allora si introduce la **funzione di ripartizione di Z** indicata con:

$$\Phi(z) := F_Z(z) = P(Z \leq z) = \int_s^t f_Z(t)dt$$

Anche questa non è calcolabile esplicitamente ma i valori che può assumere sono riportati in una tabella.

z	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
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	0.99995	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	0.99997									
5.0	0.9999997									
6.0	0.9999999990									

Osservazione: per valori negativi si applica la formula:

$$\Phi(z) = 1 - \Phi(-z)$$

Una variabile aleatoria X si dice **normale** con media μ e varianza σ^2 e si scrive con $X \sim N(\mu, \sigma^2)$ se X è assolutamente continua con:

$$f_X(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{(}{z^2}2)}$$

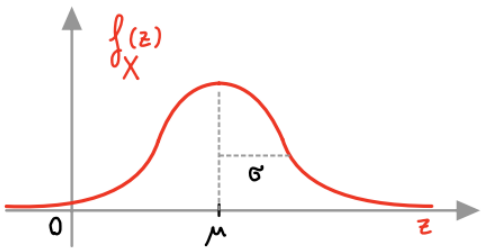


Figura 3.1: grafico a campana centrato in μ di ampiezza σ^2

Osservazione: Ci si può sempre ricondurre da una variabile normale ad una normale standard e viceversa:

- $X \sim N(\mu, \sigma^2) \to Z := \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- $Z \sim N(0, 1) \to X := \sigma \cdot Z + \mu \sim N(\mu, \sigma^2)$

Osservazione: Da questi trasformazioni, se abbiamo X v.a. normale standard, allora Z v.a. normale ha media μ e varianza σ^2:

$$X \sim N(\mu, \sigma^2) \to E[Z] = \mu \quad Var[Z] = \sigma^2$$

Teoremi

Se X è normale, $Y = a \cdot X + b$ è normale.
Se X e Y sono normali indipendenti allora X + Y è normale

Esempio: $X \sim N(0, 1)$, $Y \sim N(0, 1)$ allora

$$X + Y \sim N(0, 2) \qquad X - Y = X + (-1) \cdot (Y) \sim N(0, 2)$$

Funzione di ripartizione

Finora abbiamo studiato le v.a. discrete e assolutamente continue. Introduciamo un nuovo oggetto per v.a. generica: **funzione di ripartizione**

$$F_X(x) := P(X \leq x)$$

- F_X è ben definita per ogni v.a.
- F_X determina la distribuzione della v.a. X:

$$P(X \in (s,t]) = F_X(t) - F_X(s)$$

- F_X è legata alla densità discreta/densità di X:

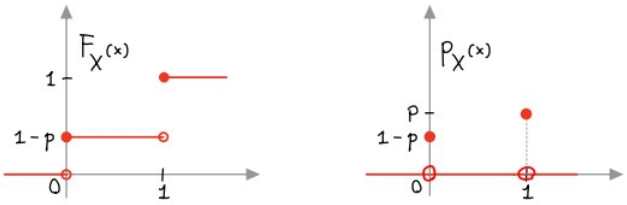
$$F_X(x) = \begin{cases} \sum_{x_i \in (-\infty, x]} P_X(x_i) & \text{se x è discreta} \\ \int_{-\infty}^{x} f_X(t)dt & \text{se x è assolutamente continua} \end{cases} \tag{3.6}$$

Esempio: Funzione di ripartizione per variabile discreta: Bernoulli. Sia $X \sim Be(p)$ con $p \in (0,1)$:

$$X(\Omega) = \{0,1\} \qquad p_X(0) = 1-p \qquad p_X(1) = p$$

Allora $F_X(x) = P(X \leq x)$ vale:

$$f_X(x) = \begin{cases} 0 & \text{se } x < 0 \\ 1-p & \text{se } 0 \leq x < 1 \\ 1 & \text{se } x \geq 1 \end{cases} \tag{3.7}$$

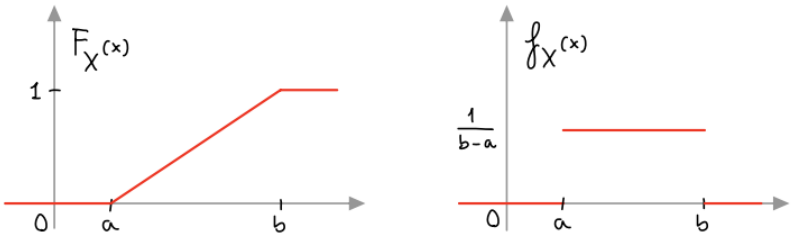


Esempio: Funzione di ripartizione per variabile uniforme continua
Sia $X \sim U(a,b)$: $X(\Omega) = [a,b]$

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{se } x < a \text{ o } x > b \end{cases} \tag{3.8}$$

Allora $F_X(x) = P(X \leq x)$ vale:

$$F_X(x) = \begin{cases} 0 & \text{se } x < a \\ \dfrac{x-a}{b-a} & \text{se } a \leq x \leq b \\ 1 & \text{se } x > b \end{cases} \tag{3.9}$$



Teorema di ripartizione di v.a. discrete

- X è una variabile aleatoria discreta $\iff$ F_X è costante a tratti
- Valori assunti x_i $\iff$ punti di discontinuità di F_X
- Densità discreta $\iff$ ampiezze dei salti
- $P_X(x_i) = F_X(x_i) - F_X(x_{\bar{i}})$ dove $x_{\bar{i}} = \lim_{t \to x_{\bar{i}}} F_X(t)$

Teorema di ripartizione di v.a. assolutamente continue

- X è una v.a. assolutamente continua (con densità continua a tratti) $\iff$ F_X è una funzione continua ed è derivabile a tratti
- Densità $F_X(x) = (F_x)'(x)$

3.4 Vettori Aleatori

Abbiamo studiato le v.a. individualmente, ma spesso è interessante lo studio *congiunto* di v.a. relative allo stesso esperimento aleatorio.

$$(\Omega, P) = \begin{cases} X : \Omega \to \mathbb{R} & \text{variabile aleatoria} \\ Y : \Omega \to \mathbb{R} & \text{variabile aleatoria} \end{cases} \tag{3.10}$$

La coppia (X,Y) è detta **vettore aleatorio**:

$$(X, Y) : \Omega \to R^2$$

3.4.1 Vettori Aleatori Discreti

Un vettore (X, Y) si dice **discreto** se i valori che può assumere sono contenuti in un insieme finito o numerabile $\{(x_i), (y_i)\}$

Si definisce **densità discreta congiunta** come:

$$P_{(X,Y)}(x_i, Y_i) := P(X = x_i, Y = y_i)$$

La relazione tra densità discreta congiunta e la densità discreta delle singole variabili aleatorie è definita come **densità discreta marginale**:

$$P_X(x_i) = \sum_{y_i} P_{(X,Y)}(x_i, y_i) \qquad P_Y(y_i) = \sum_{x_i} P_{(X,Y)}(x_i, y_i)$$

Il valore medio del prodotto tra due variabili aleatorie discrete X e Y è

$$E[X, Y] = \sum_{x_i} \sum_{y_i} x_i \cdot y_i \cdot P_{X,Y}(x_i, y_i)$$

Esempio: Lancio due monete dove X è prima moneta testa e Z numero totale di teste.
$X(\Omega) = 0, 1$
$Y(\Omega) = 0, 1, 2$
$p_{(X,Z)}(x, z) = P(X = x, Z = z)$

x \ z	0	1	2
0	$\frac{1}{4}$	$\frac{1}{4}$	0
1	0	$\frac{1}{4}$	$\frac{1}{4}$

3.4.2 Vettori Aleatori Assolutamente Continui

Un vettore (X, Y) si dice assolutamente continuo se esiste una funzione $f_{(X,Y)}(x, y) \geq 0$ detta **densità congiunta** di X e Y tale che

$$P(X \in [s,t], Y \in [u,v]) = \int_s^t \left(\int_u^v f_{(X,Y)}(x,y)dy \right) dx$$

Si definiscono **densità marginali** come:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{(X,Y)}(x,y)dy \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{(X,Y)}(x,y)dx$$

Il valore medio del prodotto tra due v.a. assolutamente continue X e Y è

$$E[X \cdot Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \cdot y \cdot f_{(X,Y)}(x,y)\ dx\ dy$$

3.4.3 Indipendenza

Consideriamo due v.a. X e Y che insieme formano un vettore aleatorio (X,Y). Si dice che X e Y sono **indipendenti** se

$$P(X \in [s,t], Y \in [u,v]) = P(X \in [s,t] \cdot Y \in [u,v])$$

Da questa formula si può ricavare che X e Y sono indipendenti se conoscendo il valore che assume Y *non cambia la distribuzione* di X:

$$P(X \in [s,t] | Y \in [u,v]) = P(X \in [s,t]$$

Osservazione: Per vettori discreti o assolutamente continui, l'indipendenza ha una riformulazione equivalente:

(X, Y) discreto, X e Y sono indipendenti sse $p_{X,Y}(x_i, y_i) = p_X(x_i) \cdot p_Y(y_i)$

(X, Y) assolut. continuo, X e Y sono indipendenti sse $f_{(X,Y)}(x, y) = f(x) \cdot f(y)$

Osservazione: Quando le v.a. non sono indipendenti, le densità marginali forniscono meno informazioni della congiunta. Quando le v.a. sono dipendenti, le densità dei singoli fornisce la stessa quantità di informazioni della congiunta.

Il valore medio del prodotto tra due variabili aleatorie indipendenti X e Y è

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

3.4.4 Covarianza e Correlazione

Consideriamo due v.a. X e Y che insieme formano un vettore aleatorio (X, Y). Indichiamo i valori medi con $\mu_X = E[X]$ e $\mu_Y = E[Y]$.
Si definisce **covarianza** di X e Y:

$$Cov[X, Y] := E[(X - \mu_X) \cdot (Y - \mu_Y)]$$

La covarianza misura il *grado di associazione* tra X e Y:
$Cov[X, Y] > 0$: a valori grandi di X corrispondono valori grandi di Y
$Cov[X, Y] < 0$: a valori grandi di X corrispondono valori piccoli di Y

Osservazione: la covarianza si può riscrivere come

$$Cov[X, Y] = E[X \cdot Y] - E[X] \cdot E[Y]$$

Proprietà della covarianza

- Annullamento: $Var[X] = Cov[X, X]$
- Simmetria: $Cov[X, Y] = Cov[Y, X]$
- Costanti: $Cov[X, c] = 0$
- Bilinearità: $Cov[a \cdot X, Y] = a \cdot Cov[X, Y]$
- Bilinearità: $Cov[X + Z, Y] = Cov[X, Y] + Cov[Z, Y]$
- **Formula della somma**: $Var[X] + Var[Y] + 2 \cdot Cov[X, Y]$

Si definisce **coefficiente di correlazione lineare**:

$$\rho[X, Y] := \frac{Cov[X, Y]}{Sd[X] \cdot Sd[Y]}$$

Si tratta di una versione normalizzata della covarianza:

- $-1 \leq \rho[X, Y] \leq 1$
- $\rho = \pm 1 \longleftrightarrow Y = a \cdot X + b$

Se Cov[X,Y] = 0, X e Y si dicono **scorrelate**.

Osservazione: Se abbiamo due variabili aleatorie indipendenti, allora sicuramente sono scorrelate, ma *non vale viceversa*

Esempio: Lancio due dadi regolari a 6 facce. Siano X la somma dei risultati e Y la differenza. Allora X e Y sono scorrelate:

$$\begin{aligned} Cov[X,Y] &= Cov[A+B, A-B] \\ &= Cov[A,A] - Cov[A,B] - Cov[B,A] - Cov[B,B] \\ &= Cov[A,A] - Cov[B,B] \\ &= Var[A] - Var[B] = 0 \end{aligned} \tag{3.11}$$

Tuttavia X e Y non sono indipendenti. Per esempio:

$$P(X=12) = \frac{1}{6} \quad P(Y=5) = \frac{1}{6}$$

$$P(X=12, Y=5) = 0$$

$$0 \neq \frac{1}{6} \cdot \frac{1}{6}$$

Capitolo 4

Teoremi di Convergenza

4.1 Teoria

4.1.1 Campione Aleatorio Casuale

Il modello probabilistico fondamentale per la statistica è la successione $X_i, ..., X_n$ di variabili aleatorie **indipendenti ed identicamente distribuite** (v.a i.i.d.)

Definizione: Chiameremo **campione aleatorio casuale** di ampiezza n le $X_i, ..., X_n$ osservabili.

Osservazione: Uno dei pochi casi in cui è facile conoscere la distribuzione del campione aleatorio è il caso del **campione gaussiano** $X_i, ..., X_n \sim N(\mu, \sigma^2)$ in quanto possiamo calcolare la media campionaria $\overline{X_N}$ con la tavola della distribuzione.

Obiettivo: Per ottenere il nostro obiettivo, ovvero quello di conoscere la distribuzione di un campione aleatorio $X_i, ..., X_n$, dovremo ricondurre il campione aleatorio ad una gaussiano $N(0, 1)$

Definizione: Per standardizzare un campione aleatorio $X_i, ..., X_n$ dovremo prendere la somma di tale campione, sottrarre la sua media e dividere per la radice della varianza:

$$\frac{X_i + ... + X_n - E[X]}{\sqrt{Var[X]}}$$

Osservazioni: Guardiamo ora le distribuzioni di $X_i, ..., X_n$ per campioni non normali. Avremo che le densità si riconducono a quella di una normale *all'aumentare di n.*

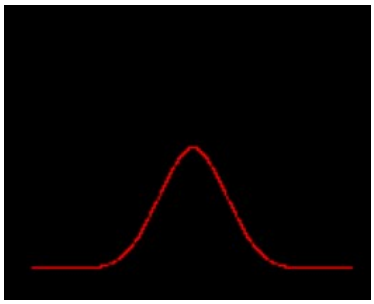


Figura 4.1: Uniforme Continua di parametro $U(0,1)$ con n=16

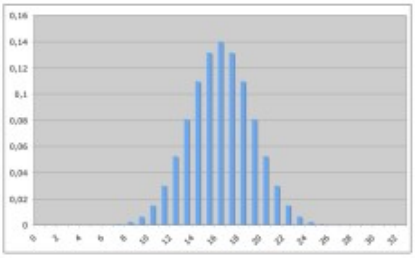


Figura 4.2: Bernoulli di parametro $Be(1/2)$ con n=32

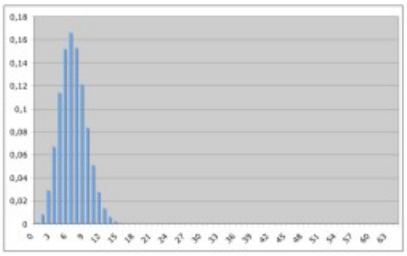


Figura 4.3: Bernoulli di parametro $Be(1/10)$ con n=64

4.1.2 Teorema del Limite Centrale

Definizione: Il **Teorema del Limite Centrale** (TLC) afferma che la somma o la media di un grande numero di v.a. i.i.d e è approssimativamente normale per n grande ($n \geq 30$). Siano $E[X_i] = \mu$ e $Var[X_i] = \sigma^2$ allora

$$\text{Media: } \mathbb{P}\left(\frac{\overline{X_N} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) \sim P(Z \leq x)$$

$$\text{Somma: } \mathbb{P}\left(\frac{X_i + ... + X_n - \mu}{\sigma} \leq x\right) \sim P(Z \leq x)$$

4.1.3 Correzioni di Continuità

Definizione: La correzione di continuità si applica quando si standardizza una distribuzione discreta ad una normale. Afferma che bisogna ampliare di $1/2$ gli estremi dell'intervallo per ottenere un valore ben approssimato

Definizione: La correzione di continuità di una **binomiale** afferma che dato se $np \geq 5$ e $n(1-p) \geq 5$ allora la sua correzione è

$$X \sim Bin(np, np(1-p))$$

Definizione: Se non ci troviamo nel caso di prima, quindi $np \leq 5$ e $n(1-p) \leq 5$ allora approssimiamo con la correzione di continuità di una **Poisson**:

$$X \sim Pois(n(1-p))$$

4.2 Pratica

Standardizzazione: $\dfrac{X_i + ... + X_n - E[X]}{\sqrt{Var[X]}}$

TLC:

$$\text{Media: } \mathbb{P}\left(\frac{\overline{X_N} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) \sim P(Z \leq x)$$

$$\text{Somma: } \mathbb{P}\left(\frac{X_i + ... + X_n - \mu}{\sigma} \leq x\right) \sim P(Z \leq x)$$

CC: Riduzione di $1/2$ del dominio per approssimare V.A. discrete

CC Binomiale: se $np \geq 5$ e $n(1-p) \geq 5$ allora: $X \sim Bin(np, np(1-p))$

CC Poisson: se $np \leq 5$ e $n(1-p) \leq 5$ allora: $X \sim Pois(n(1-p))$

4.2.1 Esercizi

Esercizio 1: TLC con assolutamente continua

Traccia: Una lampada ha un tempo di vita che segue una legge esponenziale di media 10 giorni. Non appena la lampada smette di funzionare, viene sostituita con una nuova.

a. Qual è la probabilità che 40 lampade siano sufficienti per un anno?

b. Qual è il numero minimo n di lampade comprare affinché la probabilità dell'evento "n lampade siano sufficienti per un anno" sia almeno 0.95?

Soluzione a:

1. Ricavo dalla traccia che ho 40 lampade con legge esponenziale di media 10 giorni. Ricavo dalla tabella che se $E[X] = 1/\lambda = 10$, allora $\lambda = 1/10$ Riscrivo come $X_1, ..., X_{40}$ v.a. i.i.d. $\sim exp(1/10)$

2. La richiesta è $\mathbb{P}(X_1 + ... + X_{40} > 365)$. Dato che $n = 40$ è abbastanza grande, possiamo stimare questa probabilità con il TLC.

3. Dobbiamo standardizzare la somma delle nostre v.a. Per farlo abbiamo bisogno di conoscere valore medio $E[X] = 1/\lambda$ e varianza $Var[X] = 1/\lambda^2$:

· Per la linearità della media, $E[X+Y] = E[X] + E[Y]$:

$$E[X_1 + ... + X_{40}] = E[X_1] + ... + E[X_{40}] = 10 + ... + 10 = 400$$

· Per l'indipendenza delle v.a., $Cov[X+Y] = 0$ quindi $Var[X+Y] = Var[X] + Var[Y]$:

$$Var[X_1+...+X_{40}] = Var[X_1]+...+Var[X_{40}] = 100+...+100 = 4000$$

4. Standardizziamo con la formula: $\frac{X_i + ... + X_n - E[X]}{\sqrt{Var[X]}}$:

$$\mathbb{P}(X_1 + ... + X_{40} > 365) = \mathbb{P}\left(\frac{X_1 + ... + X_{40} - 400}{\sqrt{4000}} > \frac{365 - 400}{\sqrt{4000}}\right)$$

· Semplifichiamo e cambiamo $>$ in $\leq$:

$$= 1 - \mathbb{P}\left(\frac{X_1 + ... + X_{40} - 400}{\sqrt{4000}} \leq \frac{-35}{20\sqrt{10}}\right)$$

· Essendo $\sim N(0,1)$ usando il TLC avremo:

$$\simeq 1 - \mathbb{P}\left(Z \leq \frac{-35}{20\sqrt{10}}\right) = 1 - \Phi\left(\frac{-35}{20\sqrt{10}}\right)$$

· Essendo l'argomento negativo diventa:

$$1 - 1 - \Phi\left(\frac{35}{20\sqrt{10}}\right) = \Phi(0.55) = 0.7088$$

5. La probabilità che 40 lampade siano sufficienti per un anno è 71%

Soluzione b:

1. Abbiamo $X_1 + ... + X_n \sim exp(1/10)$. L'incognita ora è n. Cerchiamo il minimo n tc $\mathbb{P}(X_1 + ...X_N > 365) \geq 0.95)$

2. Sicuramente n deve essere > 40 in quanto al punto a. dava probabilità 0.71 e a noi serve 0.95. Essendo n grande, il TLC è applicabile con buona approssimazione.

3. Dobbiamo standardizzare la somma delle nostre v.a. Calcoliamo valore medio e varianza:

· $E[X_1 + ... + X_n] = E[X_1] + ... + E[X_n] = 10n$

$\cdot$ $Var[X_1 + ... + X_n] = Var[X_1] + ... + Var[X_n] = 100n$

4. Standardizzo:

$$\mathbb{P}(X_1 + ... + X_n > 365) = 1 - \mathbb{P}\left(\frac{X_1 + ... + X_n - 10n}{\sqrt{100n}} \leq \frac{365 - 10n}{\sqrt{10\sqrt{n}}}\right)$$

5. Avrò numeratore negativo perché $n > 40$ quindi per TLC:

$$\Phi\left(\frac{10n - 365}{10\sqrt{n}}\right)$$

6. Calcoliamo il minimo n tc.

$$\Phi\left(\frac{10n - 365}{10\sqrt{n}}\right) \geq 0.95$$

7. Calcolo con la fdr lo z^* tc $\Phi(z^*) = 0.95$ e trovo $\Phi(1.645)$:

$$\Phi\left(\frac{10n - 365}{10\sqrt{n}}\right) \geq \Phi(1.645) = \frac{10n - 365}{10\sqrt{n}} \geq 1.645 = \qquad (\sqrt{n} = x)$$

$$= 10x^2 - 16.45x - 365 \geq 0 \rightarrow x = 6.92 = n \sim 47.89$$

8. Il minimo numero di lampadine da compare è 48

Esercizio 2: TLC con discreta + CC

Traccia: Qual è la probabilità di ottenere almeno 29 teste in 50 lanci di una moneta equilibrata?

Soluzione:

1. Definiamo $X_1, ..., X_{50}$ v.a. i.i.d $\sim Be(1/2)$
2. Cerco il numero di teste nei 50 lanci: $\mathbb{P}(X_1 + ... + X_{50} \geq 29)$, ovvero $Bin(50, \frac{1}{2})$ con numero di successi $k = 29$
3. Devo usare la CC di una binomiale, dunque devo rispettare due condizioni:
   (a) $np \geq 5$: $np = 50 \cdot 1/2 = 25 \geq 5$
   (b) $n(1-p) \geq 5$: $n(1-p) = 50 \cdot 1/2 = 25 \geq 5$

4. Essendo entrambe le condizioni confermate, posso approssimare con una binomiale della forma:

$$X \sim Bin(np, np(1-p)) = Bin(25, {}^{25}/_{2})$$

5. Approssimo con la CC ed ottengo:

$$\mathbb{P}(X_1 + ... + X_{50} \geq 28.5) = \mathbb{P}\left(\frac{X_1 + ... + X_{50} - 25}{\sqrt{\frac{25}{2}}}\right)$$

6. Applico il TLC:

$$\mathbb{P}(X_1 + ... + X_{50} \geq 28.5) \sim 1 - \mathbb{P}\left(Z \leq \frac{7}{\sqrt{50}}\right) = 1 - \Phi(0.99) = 0.1611$$

7. La probabilità di ottenere almeno 29 teste in 50 lanci è del 16%

Capitolo 5

Statistica Inferenziale

5.1 Teoria

5.1.1 Introduzione

Definizione: La *statistica inferenziale* consente di dedurre particolari caratteristiche di una popolazione limitandosi ad analizzare un numero finito e preferibilmente piccolo di suoi individui.

Definizione: Quando le caratteristiche che si vogliono individuare sono esprimibili numericamente allora esse sono dette *parametri*.

Definizione: Per *stima di parametri* si intende quindi il problema della deduzione di parametri di una popolazione facendo ricorso all'analisi di un suo sottoinsieme finito opportunamente scelto, detto *campione*.

Osservazione: Diverse tecniche possono essere utilizzate per effettuare delle stime di parametri. Noi ci limiteremo a considerare quelle classiche basate sulla conoscenza delle *distribuzioni campionarie*.

Definizione: Diverse ragioni possono portare a voler determinare le caratteristiche di una popolazione facendo ricorso esclusivamente ad un numero limitato di suoi individui: tempo, costo, disponibilità ecc. In questi casi occorre allora effettuare un *campionamento*, ovvero una scelta degli individui che verranno analizzati per effettuare le inferenze sull'intera popolazione.

Osservazione: Tutte le tecniche che verranno presentate in questo capitolo sono valide solo nel caso in cui il campione sia stato scelto secondo una pro-

cedura detta *campionamento casuale*.

Definizione: Denotiamo con X il *carattere* della popolazione su cui siamo interessati a fare dell'inferenza. Penseremo ad X come ad una *variabile aleatoria* la cui *distribuzione sconosciuta* corrisponde a quella che si otterrebbe facendo ricorso alle tecniche della statistica descrittiva sull'intera popolazione, e pensare invece ai valori assunti dai singoli individui come a delle *realizzazioni* di X. In forma matematica:

- Campione casuale di numerosità n $(X_1, X_2, ..., X_n)$: è una n-pla di v.a indipendenti aventi ognuna la stessa distribuzione del carattere X della popolazione.
- I valori $(x_1, x_2, ..., x_n)$ assunti dalla n-pla sono una realizzazione di $(X_1, X_2, ..., X_n)$.

5.1.2 Stime Puntuali

Definizione: Possiamo pensare al carattere della popolazione su cui vogliamo fare delle inferenze come ad una variabile aleatoria X, avente una *funzione di ripartizione F sconosciuta*, ma corrispondente alla distribuzione di frequenza cumulata di tale carattere, che si potrebbe ottenere se fosse possibile analizzare per intero la popolazione.

Definizione: Una *stima* è una realizzazione di una statistica campionaria.

Osservazione: Per le prossime definizioni denoteremo con μ il valore atteso e con σ^2 la varianza della popolazione X con distribuzione F incognita.

Definizione: Uno stimatore si dice *non distorto* se il loro valore atteso è uguale al valore medio che vogliamo stimare:

$$E[T]_\Theta = E[g(X_i, ..., X_n)]_\Theta = \Theta$$

Questa proprietà non è stabile a trasformazioni non lineari. Uno stimatore non distorto si dice *consistente* quando ha varianza che tende a 0 con N grande.

Definizione: Considerato un campione $(X_1, X_2, ..., X_n)$ estratto da una popolazione X, con distribuzione F, media μ e varianza σ^2 incognite. Definiamo *media campionaria* la variabile:

$$\overline{X_N} : \frac{X_1 + X_2 + ... + X_n}{n}$$

Questo stimatore è non distorto in quanto $E[\overline{X_N}] = \mu$.

Definizione: Considerato un campione $(X_1, X_2, ..., X_n)$ estratto da una popolazione X, con distribuzione F, media μ e varianza σ^2 incognite. Definiamo *varianza campionaria* la variabile:

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

Questo stimatore è non distorto in quanto $E[S_n^2] = \sigma^2$.

Definizione: Considerato un campione $(X_1, X_2, ..., X_n)$ estratto da una popolazione X, con distribuzione F, media $\mu = E[X_i]$ nota e varianza σ^2 incognita. Definiamo *varianza campionaria* la variabile:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$$

Questo stimatore è non distorto in quanto $E[\overline{S_n^2}] = \sigma^2$.

5.1.3 Distribuzione delle Statistiche Campionarie

Definizione: Prendiamo una v.a. i.i.d $Z \sim N(\mu, \sigma^2)$ e $\alpha \in (0, 1)$. Si definisce z_α quel valore tale che

$$\mathbb{P}(Z > z_\alpha) = \alpha$$

Osservazione: Vale anche $z_\alpha = -z_{1-\alpha}$

Definizione: Siano $Z_1, ..., Z_n \sim N(0, 1)$. Allora introduciamo Y come una distribuzione *chi quadrato con n gradi di libertà* tale che

$$Y = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

Definizione: Per α si pone $x_{n,\alpha}^2$ quel valore tale che:

$$\mathbb{P}(Y > x_{n,\alpha}^2) = \alpha$$

Osservazioni:

· si ha $E[Y] = n, Var[Y] = 2n$

- per n = 2 è la legge di $exp(1/2)$
- per n grande vale l'approssimazione della legge con una $N(n, 2n)$

Definizione: Sia $\overline{X_N}$ un campione casuale estratto da una popolazione $N(\mu, \sigma^2)$:

1. $\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$
2. $\sum_{i=1}^{n} \left(\frac{x_i - \overline{X_N}}{\sigma} \right)^2 \sim \chi^2(n-1)$
3. se $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X_N}}{\sigma} \right)^2$ allora $(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$

Osservazione: Osservando i punti 1) e 2), posiamo notare che ogni volta che stimiamo un parametro con chi quadrato, perdiamo un grado di libertà

Definizione: Siano $Z \sim N(0,1), Y \sim \chi^2(n)$ indipendenti, definiamo T come una distribuzione *t di Student con n gradi di libertà* come:

$$T = \frac{Z}{\sqrt{Y/n}} \qquad T \sim t(n)$$

Definizione: Per α si pone $t_{n,\alpha}$ quel valore tale che:

$$\mathbb{P}(T > t_{n,\alpha}) = \alpha$$

Osservazione: T è simmetrica rispetto a 0. Quindi $t_{\alpha,n} = t_{1-\alpha,n}$

5.1.4 Stima per Intervalli

Abbiamo visto come trovare un valore approssimato di un parametro incognito della popolazione per mezzo di una stima puntuale. Tali stime però non forniscono informazioni sul grado di approssimazione delle stesse. Per questo motivo alle stime puntuali vengono preferite quando possibile determinarle le *stime per intervalli* che sono stime espresse sotto forma di *intervalli fiduciari* all'interno dei quali con buona probabilità si trova il valore vero del parametro da stimare.

Definizione: Definiamo $\alpha \in [0,1]$ come *livello di confidenza* della stima ed il corrispondente intervallo è detto *intervallo di confidenza*. Spesso α assume

come valori 0.1, 0.05 e 0.01 .

Definizione: La *stima intervallare della media* di un campione estratto da una popolazione normale con *media incognita e varianza nota pari a* σ^2 ha come intervallo di confidenza:

$$IC = \left(\overline{x_n} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{x_n} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Osservazione: L'ampiezza dell'intervallo è due volte l'errore, ovvero lo scarto dal valore centrato. Nell'esempio di stima della media con media incognita e varianza nota, l'ampiezza è:

$$2z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Osservazione: La bontà della stima dipende dal livello di confidenza: maggiore è, più affidabile è la stima; ma all'aumentar di quest'ultimo, aumenta l'ampiezza dell'intervallo e quindi meno precisa sarà la stima.

Definizione: La *stima intervallare della media* di un campione estratto da una popolazione normale con *media e varianza incognita* utilizza la t di Student con n-1 gradi di libertà. Ha come intervallo di confidenza:

$$IC = \left(\overline{x_n} - t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}, \overline{x_n} + t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \right)$$

Definizione: La stima proporzione-frequenza di una popolazione *Bernoulliana* con media e varianza incognite, valida se $n\overline{x_n} > 5$ e $n(1 - \overline{x_n}) > 5$ ha come intervallo di confidenza:

$$IC = \left(\overline{x_n} - z_{\alpha/2} \sqrt{\frac{\overline{x_n}(1 - \overline{x_n})}{n}}, \overline{x_n} + z_{\alpha/2} \sqrt{\frac{\overline{x_n}(1 - \overline{x_n})}{n}} \right) \qquad \overline{x_n}(1 - \overline{x_n}) \leq \frac{1}{4}$$

Definizione: Stima intervallare della *varianza* su un campione estratto da una popolazione normale con media e varianza incognite è:

$$IC = \left(\frac{(n-1)s_n^2}{\chi_{n-1,\alpha/2}}, \frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha/2}} \right)$$

Definizione: Stima intervallare della *varianza* su un campione estratto da una popolazione normale con media nota e varianza incognita utilizza una χ con n-1 gradi di libertà ed è:

$$IC = \left(\frac{n\overline{s_n^2}}{\chi_{n,\alpha/2}}, \frac{n\overline{s_n^2}}{\chi_{n,1-\alpha/2}} \right)$$

5.2 Pratica

5.2.1 Esercizi

Esercizio 1: Intervalli di Confidenza, Stime Media

Traccia: La concentrazione di PCB nel latte materno ha approssimativamente una distribuzione Normale con media μ e varianza σ^2 entrambe incognite. Si misura un campione di 20 individui, ottenendo $\overline{x_n} = 5.8$ e $s_n = 5.085$

1. IC per μ a livello 95%
2. IC per μ a livello 99%

Soluzione punto 1:

1. *Trovo la formula da usare*: sono nel caso di media e varianza incognite e voglio trovare l'intervallo di μ, controllo nel formulario delle stime per intervalli e trovo:

$$IC = \left(\overline{X_N} - t_{n-1,\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}, \overline{X_N} + t_{n-1,\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\right)$$

2. *Calcolo* α: Abbiamo $100(1-\alpha)\% = 95\%$. Trovo $\alpha = 0.05$
3. *Riscrivo i miei dati*: Ho $n = 20$, $\overline{x_n} = 5.8$, $s_n = 5.085$, $\alpha = 0.05$. Riguardando la formula mi manca conoscere $\frac{\alpha}{2} = 0.025$
4. *Uso la tavola di t di Student*: Incrocio $n - 1 = 19$ e $\frac{\alpha}{2} = 0.025$ e trovo 2.093
5. *Riscrivo la formula*:

$$IC = \left(5.8 - 2.093\frac{5.085}{\sqrt{20}}, 5.8 + 2.093\frac{5.085}{\sqrt{20}}\right) \simeq (3.12, 8.18)$$

6. *Conclusione*: L'intervallo di confidenza per μ a livello 95% è $(3.12, 8.18)$

Soluzione punto 2:

1. *Trovo la formula da usare*: Sono nello stesso caso di prima in quanto cambia solo il livello di confidenza

$$IC = \left(\overline{X_N} - t_{n-1,\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}, \overline{X_N} + t_{n-1,\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\right)$$

2. *Calcolo* α: A differenza del precedente punto, adesso ho un livello di 99% quindi $100(1-\alpha)\% = 99\%$. Trovo $\alpha = 0.01$

3. *Riscrivo i miei dati*: Ho $n = 20$, $\overline{x_n} = 5.8$, $s_n = 5.085$, $\alpha = 0.01$. Riguardando la formula mi manca conoscere $\frac{\alpha}{2} = 0.005$

4. *Uso la tavola di t di Student*: Incrocio $n - 1 = 19$ e $\frac{\alpha}{2} = 0.005$ e trovo 2.861

5. *Riscrivo la formula*:

$$IC = \left(5.8 - 2.861\frac{5.085}{\sqrt{20}}, 5.8 + 2.861\frac{5.085}{\sqrt{20}}\right) \simeq (2.55, 9.05)$$

6. *Conclusione*: L'intervallo di confidenza per μ a livello 95% è $(2.55, 9.05)$

Esercizio 2: Intervalli di Confidenza, Stime Proporzioni

Traccia: Voglio stimare la proporzione di donne tra gli insegnanti della scuola secondaria. Su un campione di 1000 insegnanti ci sono 518 donne.

1. Stima puntuale della popolazione tramite uno stimatore non distorto

2. IC al 95% della proporzione

3. IC al 99% la cui ampiezza non sia maggiore di 0.03. Quanto dovrebbe essere numeroso il campione?

Soluzione Punto 1: Siamo nel caso campione numero estratto da una popolazione Bernoulliana Be(p) con p incognito e quindi media e varianza incognite. Uno stimatore non distorto di una Be(p) è la media p. Dunque la stima puntuale richiesta è $\overline{x_n} = {}^{518}/_{1000} = 0.518$

Soluzione Punto 2:

1. *Trovo la formula da usare*: Dal formulario,

$$IC = \left(\overline{x_n} - z_{\alpha/2}\sqrt{\frac{\overline{x_n}(1-\overline{x_n})}{n}}, \overline{x_n} + z_{\alpha/2}\sqrt{\frac{\overline{x_n}(1-\overline{x_n})}{n}}\right)$$

2. *Trovo* α: $z_{\alpha/2}$ è il $100(1 - {}^{\alpha}/_{2})$esimo percentile quindi $\alpha = 0.05$ e ${}^{\alpha}/_{2} = 0.025$

3. *Tavola Gaussiana*: A differenza della t Student, non troviamo la coda, ma la fdr. Quindi dobbiamo trovare $P(Z \leq z_{0.025}) = 1 - 0.025 = 0.975$ e quindi trovo che $z_{0.025} = 1.96$

4. *Riscrivo la formula*:

$$IC = \left(0.518 - 1.96\sqrt{\frac{0.518(1-0.518)}{1000}}, 0.518 + 1.96\sqrt{\frac{0.518(1-0.518)}{1000}}\right)$$

$$\simeq (0.487, 0.549)$$

5. *Conclusione*: L'intervallo di confidenza per la proporzione a livello 95% è $(0.487, 0.549)$

Soluzione Punto 3:

1. *Formula Ampiezza*: L'ampiezza di un IC per definizione è 2 volte lo scarto:

$$2z_{\alpha/2}\sqrt{\frac{\overline{x_n}(1-\overline{x_n})}{n}}$$

2. *Ricavare i dati*: Non conosciamo $\overline{x_n}(1-\overline{x_n})$, ma è sicuramente $\leq {}^1\!/_4$ (vedi da formulario). La formula diventa

$$2z_{\alpha/2}\sqrt{\frac{{}^1\!/_4}{n}} = z_{\alpha/2}\frac{1}{\sqrt{n}}$$

3. *Tavola Gaussiana*: Dato ${}^\alpha\!/_2 = 0.005$ devo trovare nella tavola 0.995 e lo incastro tra i valori $(2.57, 2.58)$ quindi $z_{0.005} = 2.575$

4. *Riscrivo la formula*:

$$2.575\frac{1}{\sqrt{n}} \leq 0.03 \rightarrow \sqrt{n} \geq \frac{2.575}{0.03} \rightarrow n \geq 7373.08$$

5. *Conclusione*: Per avere IC al 99% con ampiezza non maggiore di 0.03 ho bisogno di 7374 professori

Esercizio 3: Intervalli di Confidenza, Stime Varianza

Traccia: Si considera il campione

$$1.752.251.92.32.11.7$$

proveniente da una legge normale con media e varianza incognite.

1. Stima puntuale della varianza usando stimatore non distorto
2. IC al 99% per σ^2
3. Come cambiano le risposte se μ è nota pari a 2

Soluzione Punto 1: Stimatore non distorto di σ^2 con media e varianza incognite è

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2$$

Conosco n=6, $\overline{x_n} = 2$ e quindi la stima puntuale di σ^2 è:

$$s_n^2 = \frac{1}{5}[(1.75-2)^2 + ... + (1.7-2)^2] = 0.065$$

Soluzione Punto 2:

1. *Trovo la formula da usare*: Stima intervallare della varianza con media e varianza incognita utilizza una χ con n-1 gradi di libertà ed è:
$$IC = \left(\frac{(n-1)s_n^2}{\chi_{n-1,\alpha/2}}, \frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha/2}}\right)$$
2. *Calcoliamo e Usiamo la Tavola del chi quadro*: Abbiamo $\alpha = 0.01$ e quindi cerchiamo $\chi^2_{5,0.005}$ e $\chi^2_{5,0.995}$ e trovo rispettivamente 16.75 e 0.412
3. *Riscrivo la formula*:
$$IC = \left(\frac{5 \cdot 0.065}{16.75}, \frac{5 \cdot 0.065}{0.412}\right) \simeq (0.019, 0.79)$$

Soluzione Punto 3:

1. *Trovo lo stimatore non distorto*: Stimatore di σ^2 incognito conoscendo $\mu = 2$ è $\overline{s_n^2} = \frac{1}{n}\sum_{i=1}^{2}(x_i - \mu)^2 = 0.054$
2. *Trovo la formula da usare*: Stima intervallare varianza con media nota e varianza incognita è:
$$IC = \left(\frac{n\overline{s_n^2}}{\chi_{n,\alpha/2}}, \frac{n\overline{s_n^2}}{\chi_{n,1-\alpha/2}}\right)$$
3. *Calcoliamo e Usiamo la Tavola del chi quadro*: $\alpha = 0.01$ e cerchiamo $\chi_{6,0.005}$ e $\chi_{6,0.995}$ e trovo rispettivamente 18.548 e 0.676
4. *Riscrivo la formula*:
$$IC = \left(\frac{6 \cdot 0.054}{18.548}, \frac{6 \cdot 0.054}{0.676}\right) \simeq (0.017, 0.479)$$

Capitolo 6

Verifica di Ipotesi

6.1 Teoria

6.1.1 Test di Ipotesi

Definizione: Un'affermazione relativa ad una caratteristica di una popolazione è detta *ipotesi statistica* quando essa viene formulata sulla base dell'esperienza o sulla base di considerazioni teoriche.

Osservazione: Il problema di un ipotesi è la *verifica* della validità di un'ipotesi statistica. Per effettuare tali verifiche si utilizzano procedure statistiche dette *test di ipotesi* e si dividono in:

- *test parametrici*: si riferiscono ad ipotesi relative a parametri di distribuzione della popolazione (media e varianza)
- *test non parametrici*: si riferiscono al tipo di distribuzione ipotizzabile per la popolazione non esprimibili come parametri

Definizioni: Ogni test di ipotesi è caratterizzato da:

- una *popolazione statistica* X sulla quale viene effettuata il test
- un'*ipotesi nulla* H_0 da convalidare o rifiutare sulla base dei valori assunti da un campione $(X_i, ..., X_n)$ estratto da X
- un'*ipotesi alternativa* H_1 da considerare valida quando si rifiuta H_0
- una *statistica campionaria* $T = T(X_i, ..., X_n)$ di cui è nota la distribuzione quando H_0 è vera

- una *regione di accettazione* $\overline{C}$ che è l'insieme di valori assumibili dalla statistica T che portano ad un'accettazione dell'ipotesi H_0
- una *regione critica* C che è l'insieme di valori assumibili dalla statistica T che portano ad un rifiuto dell'ipotesi H_0
- un *livello di significatività* α che permette di individuare la regione di accettazione, tale che se H_0 vero allora T assume valori nella regione critica con probabilità α

Definizione: Se $(X_i, ..., X_n) \in C$, ossia si rifiuta H_0, diremo che i dati sperimentali sono in **contraddizione significativa** con H_0. Altrimenti i dati non sono in contraddizione significativa con H_0

Osservazione: La regione si chiama critica perché è improbabile che H_0 sia vera quando la statistica campionaria assume valori appartenenti ad essa ma non possiamo comunque escluderlo.

Definizione: Nei test di ipotesi si possono commettere due tipi di errori:

1. *Errore di prima specie*: si rifiuta H_0 quando è vera. Coincide con il livello di significatività α
2. *Errore di seconda specie*: si accetta H_0 quando è falsa. In genere non è nota la probabilità β

Definizione La procedura per la formulazione di un test di ipotesi solitamente prevede nell'ordine:

1. Individuazione dell'ipotesi nulla H_0 e dell'ipotesi alternativa H_1
2. La scelta del livello di significatività α
3. La scelta della statistica campionaria T
4. La determinazione delle regioni di accettazione $\overline{C}$ e critica C
5. L'accettazione o il rifiuto dell'ipotesi nulla H_0

Definizione: Una statistica è detta *semplice* se il sottoinsieme di valori che essa assegna ad un parametro è costituito da un solo elemento, altrimenti è detta *composta*. Considerando μ il parametro incognito, un esempio di statistica semplice è $H_0 : \mu = x$ mentre le composte sono nella forma $H_0 : \mu \in (x, y) \quad H_0 : \mu > x$

Definizione: L'appartenenza di $X_i, ..., X_n \in \overline{C}$ dipende dal livello di significatività α. Esiste un valore $\overline{\alpha}$ detto *p-value* tel test t.c. :

- per $\alpha > \overline{\alpha}$ si rifiuta H_0
- per $\alpha \leq \overline{\alpha}$ si accetta H_0

Osservazione: Più piccolo è il p-value, più i dati sono in contraddizione con H_0

Definizione: Un test statistico è detto *bidirezionale* se la regione critica è costituita dall'unione di due sottoinsieme disgiunti mentre diremo che è *unidirezionale* se è costituita da un solo sottoinsieme. Negli esempi del punto precedente, i primi due sono bidirezionali, mentre il terzo è unidirezionale.

6.1.2 Considerazioni sugli errori

Nel descrivere le caratteristiche di un test di ipotesi la probabilità di compiere errori di II specie va pensata come una funzione anziché come uno specifico valore numerico

Definizione: Sia Θ il parametro a cui si riferisce il test e sia Θ^* il valore specificato dall'ipotesi nulla. Denotiamo l'errore di II specie come

$$\beta(\hat{\Theta}) = \mathbb{P}(\text{accettare } H_0 | H_0 \text{ è falsa e } \Theta = \Theta^*)$$

allora viene detta *curva di potenza del test* la funzione

$$\pi(\hat{\Theta}) = 1 - \beta(\hat{\Theta})$$

Un test risulta tanto migliore quanto più la funzione $\pi(\hat{\Theta})$ si avvicina ad 1 al variare di $\hat{\Theta}$

Osservazione: Se nella costruzione di un test si desidera diminuire il livello di significatività α (errori di I specie) si può ampliarne la regione di accettazione $\overline{C}$. In questo modo però è facile osservare che diminuisce anche la potenza del test $\pi(\hat{\Theta})$,ovvero aumenta la probabilità di compiere errori di II specie.

6.1.3 Test sulla Media di una Popolazione

Vediamo come si costruisce un test sulla media μ di una popolazione X quando si formula l'ipotesi nulla che tale media sia un valore fissato: $H_0 : \mu = \mu_0$

e quando si dispone di un campione casuale $(X_i, ..., X_n)$ estratto da X. Vediamo tre distribuzioni per la media, una normale con varianza nota, una normale con varianza incognita e una non normale.

Popolazione normalmente distribuita e varianza σ^2 nota: Sia μ il valore del parametro. Avremo che

$$Z = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Z non è una statistica in quanto μ non è noto. Allora

$$Z_n = \frac{\overline{X_n} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

è una statistica essendo μ_0 fissata. Il test rifiuta H_0 se Z_n assume valori poco probabili per una $N(0,1)$. Avremo 3 ipotesi alternative:

1. $H_1^{'} : \mu \neq \mu_0$
2. $H_1^{''} : \mu < \mu_0$: in tal caso μ_0 è una sovrastima della media per cui Z_n tende ad assumere valori negativi $Z_n < Z \sim N(0,1)$
3. $H_1^{'''} : \mu > \mu_0$: in tal caso μ_0 è una sottostima della media per cui Z_n tende ad assumere valori positivi essendo $Z_n > Z \sim N(0,1)$

La regione critica risulta essere

1. $C^{'} = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, +\infty)$
2. $C^{''} = (-\infty, -z_{1-\alpha})$
3. $C^{'''} = (z_{1-\alpha}, +\infty)$

Osservazione: Nel formulario, ci sono 3 ipotesi nulle

$$H_0 = \mu_0 \quad H_0 \geq \mu_0 \quad H_0 \leq \mu_0$$

invece che la singola $H_0 = \mu_0$

Popolazione normalmente distribuita e varianza σ^2 incognita: Si procede come nell'esempio precedente, ma essendo varianza incognita useremo la t di Student.

Popolazione non normalmente distribuita: Per poter definire un test occorre avere un campione di numerosità sufficientemente elevata ($n \geq 30$) e deve valere $np_0 \geq 5$ e $n(1-p_0) \geq 5$ Allora, possiamo ricondurci ad una normale con varianza nota.

6.1.4 Test sulla Differenza delle Medie di due Popolazioni

Molti problemi in statistica consistono nel confronto tra 2 variabili, come 2 popolazioni. Vediamo due test, uno sulla differenza di media di due campioni normali accoppiati, ed uno sulla differenza di media di due campioni normali indipendenti.

Differenza Campioni Accoppiati: Prendiamo due campioni della stessa numerosità $(X_i, ..., X_n)_x \quad (Y_1, ..., Y_n)_y$. Supponiamo essi siano normali, quindi con valore medio μ_x e μ_y. Definiamo una differenza $D_i = X_i - Y_i$ e denotiamo con $\overline{D_n}$ la media campionaria e con S_d^2 la varianza campionaria del campione $D_1, ..., D_n$ Allora troviamo la statistica come:

$$T = \frac{\overline{D_n} - \mu_0}{S_d}\sqrt{n}$$

Di conseguenza posso trovare le tre sezioni critiche. (non le riporto, sono presenti nella tabella)

Differenza Campioni Indipendenti: Prendiamo due campioni della stessa numerosità $(X_i, ..., X_n)_x \quad (Y_1, ..., Y_n)_y$. Supponiamo essi siano normali, quindi con valore medio μ_x e μ_y. Siano $\overline{X}$ e S_x^2 media e varianza campionaria di $(X_i, ..., X_n)_x$ e siano $\overline{Y}$ e S_y^2 media e varianza campionaria di $(Y_1, ..., Y_n)_y$. Possiamo calcolare la varianza campionaria combinata come

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

Allora ottengo la mia statistica come:

$$T = \frac{\overline{XY}}{S_p\sqrt{\frac{1}{n_x}\frac{1}{n_y}}}$$

Di conseguenza posso trovare le tre sezioni critiche. (non le riporto, sono presenti nel formulario)

La parte sui test non parametrici e in particolare del test chi-quadro di buon adattamento è trattata nel prossimo capitolo per questioni di parallelismo teoria-esercitazione

6.2 Pratica

6.2.1 Esercizi

Esercizio 1: Test z per popolazione gaussiana, varianza nota

Traccia: Azienda produce anelli. Il diametro di questi anelli è normalmente distribuito ed ha una deviazione standard pari a $\sigma = 0.001mm$. Campione di n=15 anelli, si ricava $\overline{x} = 74.036mm$

1. Si testi l'ipotesi che la media del diametro sia uguale a $\mu = 74.035mm$ ad un livello di significatività pari a $\alpha = 5\%$ e si calcoli il p-value del test.

2. Probabilità che a livello $\alpha = 5\%$, l'ipotesi che la media sia $\mu = 74.035mm$ non venga rifiutata quando il valore della media è 74.034

Soluzione punto 1: Essendo il campione normalmente distribuito ed avendo varianza nota troviamo dal formulario il Test z sulla media di una popolazione normale con varianza nota pari a σ^2

1. *Scrivo l'ipotesi nulla e l'ipotesi alternativa*: $H_0 : \mu = 74.035$, $H_1 : \mu \neq 74.036$

2. *Ricavo dalla tabella la regione critica e la calcolo*: Regione critica $\overline{C}$ del test a livello α:

$$\left|\frac{\overline{x_n - \mu}}{\sigma}\sqrt{n}\right| > z_{\alpha/2} \rightarrow \left|\frac{74.036 - 74.035}{0.001}\sqrt{15}\right| \simeq 3.873 > z_{\alpha/2}$$

3. *Confronto con il percentile della gaussiana*: $\alpha = 0.05 \rightarrow z_{0.025}$ quindi cerco sulle tavole della normale 0.975 e trovo $z_{0.025} = 1.96$

4. *Confronto*: Dopo aver calcolato regione critica e percentile della gaussiana ottengo $3.873 > 1.96$. Essendo questo vero, rifiuto H_0 a livello 5%. I dati mi permettono di dimostrate statisticamente H_1

5. *Formula p-value*: Calcolo $\overline{\alpha}$, ovvero la probabilità che la regione critica sia esattamente il percentile della gaussiana, ma $\overline{\alpha}$ è incognita: $\overline{C}$ del test a livello α:

$$\left|\frac{\overline{x_n - \mu}}{\sigma}\sqrt{n}\right| = z_{\overline{\alpha}/2} \rightarrow 3.873 = z_{\overline{\alpha}/2}$$

6. *Calcolo p-value*: Per trovare il valore di $\overline{\alpha}$ uso la fdr della gaussiana standard:

$$\Phi(3.873) = \Phi(z_{\overline{\alpha}/2}) \to \Phi(3.873) = 1 - z_{\overline{\alpha}/2} \to \overline{\alpha} = 2(1 - \Phi(3.873)) \to \overline{\alpha} \simeq 0$$

7. *Conclusione*: Più il p-value è piccolo, più i dati sono in forte contraddizione con H_0.

Soluzione punto 2: Stiamo lavorando con un errore di II specie: "accetto H_0 quando è falsa"

1. *Formula statistica*: Ricavo dalla tabella la formula

$$\mathbb{P}_{\mu=74.034} = \left(\left| \frac{\overline{x_n - \mu}}{\sigma} \sqrt{n} \right| > z_{\alpha/2} \right) = \left(\left| \frac{\overline{x_{15}} - 74.035}{0.001} \sqrt{15} \right| > 1.96 \right)$$

2. *Riscrivere come normale standard*: Essendo sotto ipotesi che $\mu \neq \overline{x_n}$ il nostro corpo della probabilità non è approssimabile come una $N(0,1)$. Aggiungo allora (+ 0.001 - 0.001):

$$\left(-1.96 \leq \frac{\overline{x_{15}} - 74.035 + 0.001 - 0.001}{0.001} \sqrt{15} \leq 1.96 \right)$$

3. *Faccio i calcoli*:

$$\left(-1.96 \leq \frac{\overline{x_{15}} - 74.034}{0.001} \sqrt{15} - \frac{0.001}{0.001} \sqrt{15} \leq 1.96 \right) =$$

$$= \left(-1.96\sqrt{15} \leq \frac{\overline{x_{15}} - 74.034}{0.001} \sqrt{15} \leq 1.96\sqrt{15} \right) \simeq$$

$$\simeq \mathbb{P}(1.91 \leq z \leq 5.83) = \Phi(5.83) - \Phi(1.11) = 0.0281$$

Traccia: Un produttore di batterie ha messo sul mercato un nuovo modello sostenendo che la durata media è superiore a quella del vecchio modello che era pari a 14 ore. Su un campione di 10 batterie sono state osservate le seguenti durate: 18 15 14 16 15 12 13 15 17. Supponendo che il tempo dio durata sia normale:

1. Sottoporre a verifica l'affermazione de produttore a livello $\alpha = 5\%$
2. Calcolare il p-value del test

Soluzione punto 1: Essendo il campione normalmente distribuito ed avendo varianza incognita troviamo dal formulario il Test t sulla media di una popolazione normale con varianza incognita

1. *Ricavo l'ipotesi nulla e l'ipotesi alternativa*: In questo caso la traccia non specifica direttamente qual'è l'ipotesi nulla. Possiamo scegliere che l'affermazione del produttore sia l'ipotesi alternativa $H_1 : \mu > 14$, in quanto se rifiuto $H_0 : \mu \leq 14$ posso dire con certezza che il produttore abbia torto. Se avessimo messo l'ipotesi del produttore come ipotesi nulla H_0 e la rifiutassimo i dati non ci permetterebbero di escludere che il produttore abbia ragione (conclusione meno forte).

2. *Ricavo dalla tabella la regione critica e la calcolo*: Regione critica $\overline{C}$ del test a livello α:
$$\left| \frac{\overline{x_n - \mu}}{s_n} \sqrt{n} \right| > t_{n-1,\alpha}$$

3. *Trovo i dati e risolvo*: $\alpha = 0.05$, $n = 10$, $\overline{x_n} = \frac{18+...+13}{10} = 15$, $s_n = \frac{1}{9}((18-15)^2 + ... + (13-15)^2) = 4$, $s_n^2 = \sqrt{4} = 2$ dunque la mia regione critica sarà:
$$\frac{15-14}{2}\sqrt{10} \simeq 1.58 > t_{n-1,\alpha}$$

4. *Confronto con t di Student e conclusioni*: trovo che $t_{9,0.005} = 1.833$ quindi $1.58 > 1.833$. Essendo questo falso accetto H_0 a livello 5%, ovvero i miei dati non mi permettono di rifiutare H_0, ovvero non mi permettono di dimostrare che il produttore ha ragione. I dati non sono in contraddizione significativa con H_0

Soluzione punto 2: P-value, so già che $\overline{\alpha} > 5\%$ dal punto prima, quindi p-value alto. So che $1.58 = t_{9,\alpha}$ quindi cerco i numeri che includono 1.58 e li trovo ai valori $0.05 < \overline{\alpha} < 0.1$ della tabella, ovvero $5\% < \overline{\alpha} < 10\%$

Esercizio 3: Test sulla proporzione

Traccia: Un'inserzione pubblicitaria per un prodotto contro il mal di testa dichiara che almeno un 90% delle persone che soffrono di questo disturbo otterrebbe beneficio se lo usasse. L'associazione dei consumatori, considerando tale pubblicità tendenziosa, ottiene un campione di 100 individui di cui 88 dichiarano che il prodotto è stato efficace. Siete d'accordo con l'associazione?
Soluzione: Essendo il campione una proporzione usiamo il Test z sulla proporzione. Le ipotesi ($n \geq 30$), $np_0 \geq 5$ e $n(1-p_0) \geq 5$ sono verificate con i nostri dati $p_0 = 0.9$ e $n = 100$ quindi posso procedere.

1. *Trovo ipotesi nulla ed alternativa*: Metto come ipotesi alternativa $H_1 : p < 0.9$ in modo tale che se rifiuto $H_0 : p \geq 0.9$ posso dire con certezza che la pubblicità sta mentendo.

2. *Scelta di svolgimento*: Per i test sulla proporzione possiamo scegliere se prendere un livello di significatività a piacere oppure trovare il p-value. Facciamo il primo.

3. *Trovo Area Critica*: Scelgo $\alpha = 0.05$ e la mia area critica usando il formulario è

$$\frac{\overline{x_n} - p_0}{\sqrt{p_0(1-p_0}}\sqrt{n} < -z_\alpha \rightarrow \frac{0.88 - 0.9}{\sqrt{0.9(1-0.1)}}\sqrt{100} < -1.645$$

4. *Conclusioni*: Ottengo $-0.667 < -1.645$ che è falso quindi accetto H_0 a livello 5% e quindi i dati non sono in contraddizione significativa con H_0 e quindi non mi permettono di dimostrare statisticamente che la pubblicità sia tendenziosa. Se calcolassi il p-value ottengo 0.2524 quindi p-value grande

Esercizio 4: Dati Accoppiati

Traccia: I dati seguenti mettono in relazione la frequenza cardiaca di 12 individui prima e dopo aver masticato tabacco. Verifica a livello 5% l'ipotesi che masticare tabacco non provochi un aumento della frequenza cardiaca e calcolare p-value

Masticare Tabacco		
Individuo	Frequenza prima	Frequenza Dopo
1	73	77
2	67	69
3	68	73
..	..	..
12	78	80

Soluzione: Usiamo il test t sulla differenza della media di 2 campioni normali accoppiati $X_i, ..., X_n$ di media μ_x e $Y_1, ..., Y_n$ di media μ_y. $D_1 = X_1 - Y_1$, $D_n = X_n - Y_n$, media $\overline{D_n}$, varianza S_d^2.

1. *Scelgo ipotesi nulla ed ipotesi alternativa*: $H_0 : \mu_d \leq 0$, $H_1 : \mu_d > 0$ ovvero metto come ipotesi alternativa che ci sia differenza così se rifiuto l'ipotesi nulla posso dire con certezza che c'è differenza tra prima e dopo masticare tabacco.

2. *Regione Critica*: Dal formulario:

$$\frac{\overline{d_n} - \mu_0}{S_d}\sqrt{n} > t_{n-1,\alpha}$$

3. *Calcolo i valori*: Prendo i valori $n = 12$, $\mu_0 = 0$, $\overline{d_n} = \frac{(77-73)+...+(80-78)}{12} = 3.75$, $s_d^2 = \frac{1}{11}((4 - 3.75)^2 + ... + (2 - 3.75)^2 = 9.477$, $s_d = 3.078$, $t_{11,0.05} = 1.796$

4. *Confrontiamo*:
$$\frac{3.75 - 0}{3.078}\sqrt{12} = 4.22 > 1.796$$
I dati sono in contraddizione significativa con H_0 dimostro statisticamente che masticare tabacco provoca un aumento della frequenza cardiaca.

5. *P-value*: $4.22 = t_{11,\overline{\alpha}}$ e trovo $0.0005 < \overline{\alpha} < 0.001$ forte evidenza empirica contro H_0 in quanto p-value molto piccolo

Capitolo 7

Regressione Lineare

7.1 Teoria

7.1.1 Test Non Parametrici

Questa sezione è trattata in questo capitolo nonostante a livello teorico appartenesse al precedente perché sulla regressione ci sono pochi esercizi e quindi l'esercitatrice ha unito i due argomenti

Nei test parametrici non si fanno inferenze sui parametri ma sulla distribuzione di una o più popolazioni. Un esempio è il test Chi-quadro di buon adattamento, dove vogliamo sapere se la popolazione ha una certa distribuzione assegnata.

Test χ^2 di buon adattamento: prendiamo un insieme di dati che vogliamo analizzare e supponiamo una certa distribuzione che ci aspettata se i dati fossero distribuiti in modo casuale. Il test confronta la distribuzione effettiva dei dati con quella attesa e ci dice se ci sono differenze significative tra le due. In conclusione, se i dati che abbiamo raccolto sono simili a quelli che ci aspetteremmo in base alla nostra ipotesi, allora il test ci darà un valore basso, altrimenti ci darà un valore alto.

Osservazione: Il test chi-quadro può essere effettuato in 3 contesti diversi:

1. Caso in cui la popolazione ha distribuzione discreta finita
2. Caso in cui la popolazione ha distribuzione discreta infinita
3. Caso in cui la popolazione ha distribuzione continua

4. Confronto con una distribuzione assegnata a meno di parametri incogniti

Osservazione: Per poter applicare il test nei casi continui o con dati infiniti, sarà fondamentale raggruppare i dati.

Definizione: Per poter applicare la procedura del test per una distribuzione discreta, affinché ci sia buona approssimazione devono essere soddisfatte le due condizioni della *regola empirica*:

1. $f_k > 1 \quad \forall k = 1, ..., j$
2. $f_k \geq 5$ per l'80% dei K

Osservazione: Se i dati non verificano la regole empirica andrebbero raggruppate le classi.

7.1.2 Richiami di Statistica Descrittiva - Dati accoppiati

Abbiamo un campione in cui osserviamo due variabili accoppiate $(x_1, Y_1), ..., (x_n, y_n)$. Abbiamo 2 metodi per misurare il legame:

· graficamente: diagramma di dispersione

· numericamente: coefficiente di correlazione lineare:

$$r_{x,y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \overline{x}(y_i - \overline{y}}{s_x s_y}$$

Osservazione: Il coefficiente di correlazione lineare assume valori r$\in [-1, 1]$. Nello specifico diremo che:

· $|r_{x,y}|$ vicino a 1: avremo forte legame tra x e y

· $|r_{x,y}|$ vicino a 0: avremo poco legame tra x e y.

Obiettivo: L'idea ora è di trovare la retta di equazione $y = a + bx$ che*meglio approssima* i dati nel diagramma di dispersione, ovvero che minimizza le distanze al quadrato tra y_i valore osservato e $a + bx_i$ valore previsto fissata la retta.

Definizioni: $(y_i - a - bx_i)^2$ è lo *scarto quadratico* tra il valore osservato y_i e il valore previsto $a + bx_i$. Lo *scarto quadratico totale* è la sua sommatoria:

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

Osservazione: Per trovare a e b in modo da minimizzare lo scarto quadratico totale. Troveremo:

- $b = \dfrac{r_{x,y}}{s_x} s_y$
- $a = \overline{y} - b\overline{x}$

Definizione: Con questi valori di a e b la retta prende il nome di *retta di regressione di y su x*

7.1.3 Modello di Regressione Lineare Semplice

Obiettivo: Finora abbiamo confrontato 2 variabili, ora si tratta di studiare la *dipendenza* tra esse.

Definizione: Vogliamo studiare la *distribuzione di y* per ogni fissato valore della variabile x:

- x è detta *ingresso* o predittore o input
- y è detta *uscita* o risposta o output

Osservazione: Si assume che il legame tra input ed output sia lineare a meno di un errore:

$$Y_i = \alpha + \beta x_i + e_i \qquad e1, ..., e_n \text{ i.i.d}$$

Dunque avremo x_i, α, β fissati, mentre e_i essendo l'errore è una variabile aleatoria, e di conseguenza Y_i è una variabile aleatoria.

Definizione: Si assume che $e_i \sim N(0, \sigma^2)$ e quindi

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

con α, β, σ^2 parametri incogniti. L'obiettivo è di fare inferenze sui parametri incogniti.

Definizione: Supponiamo di non aver fatto ancora osservazioni, vogliamo trovare degli stimatori di α, β. Usiamo A e B, detti *stimatori dei minimi quadrati di* α *e* β. Denotiamo con:

- S_{xx} somma dei quadrati di x: $\sum_{i=1}^{n} x_i^2 - n\overline{x}^2$
- S_{YY} somma dei quadrati di Y: $\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$
- S_{xY} somma dei prodotti incrociati tra x e Y: $\sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}$

Allora definiamo:

$$B = \frac{S_{x,Y}}{S_{xx}} \qquad A = \overline{Y} - B\overline{x}$$

Definizione: Siamo riusciti a dare una stima puntuale, per poter dare la stima intervallare e il test di ipotesi serve conoscere la *legge di A e B*:

$$B = \frac{1}{S_{xx}} \left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{Y} \right) \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

$$A = \overline{Y} - B\overline{x} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right)$$

Definizione: Queste due formule ci permettono di trovare la *stima per* σ^2. Consideriamo i *residui*: $R_i = Y_i - A - Bx_i$, vogliamo trovare la loro legge. Prendo la formula di $Y_i : \alpha + \beta x_i + e_i$, la standardizzo, la elevo al quadrato e trovo la Chi-quadro con n gradi di libertà. Se però uso due stime per α, β perdo due gradi di libertà. Definisco quindi la somma dei quadrati residui come $SS_R = \sum_{n=1}^{n} (Y_i - A - Bx_i)^2$ e trovo la stima di σ^2:

$$\frac{\sum_{n=1}^{n} (Y_i - A - Bx_i)^2}{\sigma^2} \sim \chi^2(n-2)$$

Infine, dato che $E\left(\frac{SS_R}{\sigma^2} = n-2\right)$ allora $\left(\frac{SS_R}{n-1} = \sigma^2\right)$. In conclusione, uno *stimatore non distorto di* σ^2 è

$$\frac{SS_R}{n-2}$$

7.1.4 Intervallo di Confidenza per β

Dal fatto che $B \sim N(\beta, \frac{\sigma^2}{s_{xx}})$ segue che se tolgo beta e standardizzo ottengo una gaussiana:

$$\frac{B - \beta}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

Essendo σ^2 incognito devo usare il suo stimatore $\frac{SS_R}{n-2}$. Allora ottengo:

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t(n-2)$$

Siamo riusciti ad ottenere una statistica con legge che non dipende dai parametri incogniti.

Definizione: Troviamo l'intervallo di confidenza di β di livello $(100-\gamma)$:

$$IC = \left(\hat{\beta} - \sqrt{\frac{SS_R}{S_{xx}(n-2)}}t_{n-2,\gamma/2}, \hat{\beta} + \sqrt{\frac{SS_R}{S_{xx}(n-2)}}t_{n-2,\gamma/2}\right)$$

Osservazione: La semiampiezza dell'IC per β dipende dai dati campionari perché è data dalla somma dei quadrati residui SS_R che a sua volta dipende da $y_1, ..., y_n$

7.1.5 Verifica dell'ipotesi $\beta = 0$

L'ipotesi nulla $H_0 : \beta = 0$ equivale a dire che l'output *non dipende* dall'input. Se i dati sperimentali conducono al rifiuto di H_0 è possibile concludere che la dipendenza tra input ed output sia significativa.

Definizione: Se vale H_0 e quindi $\beta = 0$ allora ottengo

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-0) \sim t(n-2)$$

Dunque la mia *regione critica* sarà:

$$\left|\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\hat{\beta}\right| > t_{n-2,\gamma/2}$$

7.1.6 Regressione verso la Media

Esiste un fenomeno di regressione verso la media, dove si nota che i caratteri più "estremi", passando di padre in figlio diventano meno "estremi".

Definizione: Per concludere con un test statistico che i dati denotano regressione verso la media, si sceglie come ipotesi alternativa $H_1 : \beta < 1$, in modo tale che se i dati permettono di rifiutare l'ipotesi nulla, ovvero accettiamo H_1, allora troviamo *regressione verso la media*. Dato che l'ipotesi nulla $H_0 : \beta \geq 1$ ma sappiamo che $0 \leq \beta \leq 1$ allora la nostra ipotesi nulla sarà $H_0 : \beta = 1$, permettendoci di calcolare la sua regione critica come:

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(\hat{\beta} - 1) < -t_{n-2,\gamma}$$

7.1.7 Coefficiente di Determinazione

Obiettivo: misurare la variazione degli output corrispondenti agli input.

Osservazione Le Y_i variano perché gli input non sono tutti uguali e per il fatto che abbiamo nella formula di Y_i un errore $e_i \sim N(0, \sigma^2)$

Definizione: Si ha che $S_{YY} - SS_R$ è la variazione dell'output quantificata dai valore dell'input, in quanto SS_R è la quantità residua di variazione della risposta. Allora troviamo il *coefficiente di determinazione* come

$$R^2 = 1 - \frac{SS_R}{S_{YY}}$$

definita come la proporzione della variazione dell'output che è qualificata dagli input: $0 \leq R^2 \leq 1$.

- R^2 vicino a 1: maggior parte della variazione delle risposte giustificata dai produttori
- R^2 vicino a 0: solo una piccola parte della variazione delle risposte è giustificata dai produttori

7.2 Pratica

7.2.1 Esercizi

Test chi-quadro di adattamento - Discreta Finita

Traccia: Un campione casuale di 100 assenza di studenti sono ripartiti tra i giorni della settimana come segue: Lunedì=27, Martedì=19 Mercoledì=13 Giovedì=15 Venerdì=26. Verifica l'ipotesi che un'assenza abbia la stessa probabilità di capire in un qualunque dei 5 giorni.

Soluzione:

1. *Distinzione discreta/continua*: Siamo nel caso di una distribuzione discreta:
$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) = (1/5, 1/5, 1/5, 1/5, 1/5)$$

2. *Calcolo frequenze assolute attese*:
$$f_i = n\pi(i) = 100\frac{1}{5} = 20$$

3. *Controllo Regola empirica*: RE soddisfatta: $f_k > 1 \quad \forall k = 1, ..., j \quad f_k \geq 5$ 80% dei K

4. *Formulo ipotesi nulla ed alternativa*: Le nostre ipotesi sono:
$$H_0 : \text{la popolazione ha distribuzione } \pi$$
$$H_1 : \text{la popolazione non ha distribuzione } \pi$$

5. *Trovo statistica e regione critica*: Guardo dal formulario, la statistica è sempre quella, la regione critica cambia se abbiamo parametri incogniti. In questo caso avremo:
$$\text{Statistica: } Q = \sum_{j=1}^{5} \frac{(N_j - f_j)^2}{f_j} \sim \chi^2(4)$$
$$\text{Regione Critica: } q = \sum_{j=1}^{5} \frac{(n_j - f_j)^2}{f_j} \sim \chi^2_{4,\alpha}$$

6. *Calcolo e Concludo*: Scelgo $\alpha = 5\%$, trovo la regione critica:

$$q = \frac{(27-20)^2}{20} + ... + \frac{(26-20)^2}{20} = 8$$

Trovo nelle tabelle $\chi^2_{4,0.05} = 9.49$ e quindi dato che $8 \not> 9.49$ accetto H_0, ovvero la popolazione ha distribuzione *pi* a livello 5%. I dati non sono in contraddizione significativa con H_0

Test chi-quadro di adattamento - Discreta Infinita

Traccia: Da un'indagine eseguita qualche anno fa si sa che il numero di difetti di un macchinario prodotto dalla fabbrica PSI segue una distribuzione di Poisson di parametro $\lambda = 1.76$. Vengono esaminati oggi 50 macchinari della fabbrica PSI ottenendo i dati relativi ai difetti: 10 senza, 14 ne mostrano uno, 15 ne mostrano due, 5 ne mostrano tre, 3 ne mostrano quattro e 3 ne mostrano 5 o più.

Soluzione

1. *Distinzione discreta/continua*: Abbiamo una v.a. discreta, nella precisione una Poisson.

2. *Calcolo le* π: Utilizziamo la formula della densità discreta della Poisson e troviamo:

   · $\pi_0 = e^{-176} \sim 0.172$
   · $\pi_1 = e^{-176} 1.76 \sim 0.172$
   · $\pi_2 = e^{-176} (1.76)^2 / 2! \sim 0.172$
   · $\pi_3 = e^{-176} (1.76)^3 / 3! \sim 0.172$
   · $\pi_4 = e^{-176} (1.76)^4 / 4! \sim 0.172$
   · $\pi_5 = 1 - P(Y < 5) = 0.034$

3. *Calcolo frequenze assolute attese*: Ottengo le frequenze assolute moltiplicando i precedenti valori per il numero di osservazioni, ovvero 50:

   · $f_0 = 8.6$
   · $f_1 = 15.15$
   · $f_2 = 13.3$
   · $f_3 = 7.8$
   · $f_4 = 3.45$

- $f_5 = 1.7$

4. *Controllo regola empirica*: I casi 4 e 5 non soddisfano il secondo punto della regola empirica, dunque li raggruppiamo ed otteniamo i nuovi valori: $\pi_4 \sim 0.103$ e $f_4 = 5.15$

5. *Formulo ipotesi nulla ed alternativa*: Le nostre ipotesi sono:

$$H_0 : \text{la popolazione ha distribuzione } \pi$$

$$H_1 : \text{la popolazione non ha distribuzione } \pi$$

6. *Statistica e Regione Critica*: Calcolo la regione critica $q = \sum_{j=1}^{4} \frac{(n_j - f_j)^2}{f_j}$ e trovo 1.677, calcolo il chi-quadro $\chi^2_{4,0.05}$ e trovo 9.49.

7. *Calcolo e concludo*: Accetto H_0 in quanto i dati sperimentali non sono in contraddizione significativa.

Test chi-quadro di adattamento - Continua e Stima Parametro

Traccia: Una ditta di martinetti idraulici vuole analizzare la durata del modello PSI in commercio da 5 anni. Si analizza un campione di 10 pezzi ottenendo i tempi intercorsi dalla consegna al guasto:

$$[0, 299] = 55$$
$$[300, 599] = 25$$
$$[600, 899] = 10$$
$$[900, 1199] = 4$$
$$[1200, 1499] = 3$$
$$[1500, 1799] = 2$$
$$[1800, 1825] = 1$$

Si può affermare che il tempo di vita segue una legge esponenziale?
Soluzione

1. *Distinzione discreta/continua*: Siamo nel caso continua, avendo una $exp(\lambda)$. Non sappiamo il valore di λ, ma essendo $E[X] = 1/\lambda$ lo stimiamo con $1/\bar{x}_{100}$

2. *Calcolo* λ: Per calcolare il valore medio devo prendere il punto medio delle classi e le rispettive frequenze. Ottengo

$$\overline{x}_100 = \frac{1}{100}(55 \cdot 150 + ... + 1 \cdot 1813) = 403.63$$

Allora la mia stima di λ è 0.002477.

3. *Calcolo frequenze assolute attese*:

   · $[0, 299] = P(Y < 300) = 1 - e^{-0.002477 \cdot 300} = 0.524$
   · $[300, 599] = P(300 \leq Y < 600) = 0.249$
   · $[600, 899] = 0.119$
   · $[900, 1199] = 0.056$
   · $[1200, 1499] = 0.027$
   · $[1500, 1799] = 0.013$
   · $[1800, 1825] = 0.001$

4. *Controllo regola empirica*: La regola empirica non è soddisfatta, allora raggruppo gli ultimi 3 casi ottenendo 0.052

5. *Calcoli e Concludo*: Gli altri punti sono uguali ai precedenti esercizi, con l'unica osservazione che in questo esercizio avrò una $\chi^2(5 - 1 - 1)$ in quanto ho stimato il parametro λ e quindi perdo un grado di libertà. Avrò $2.6 > 7.981$ dunque ai dati non sono in contraddizione significativa con H_0

Regressione Lineare

Traccia: Si consideri il seguente campione di n=9 osservazioni relative ai caratteri X e Y:

X 17 8 36 23 19 17 2 5

Y 25 0 11 -26 -2 -18 8 26 35

Supponendo valga un modello di regressione lineare semplice

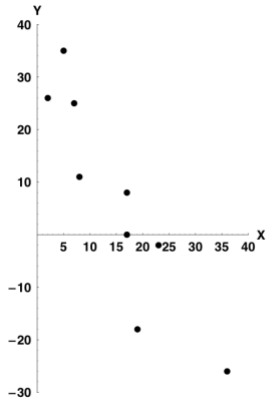
$$Y = \alpha + \beta x + e$$

1. Tracciare il grafico di dispersione dei dati, calcolare il coefficiente di correlazione lineare e la retta di regressione di Y su X.

2. Costruire un IC al livello 99% per β
3. Verificare l'ipotesi $\beta = 0$ a livello 1%
4. Calcolare una misura della bontà di adattamento del modello
5. Discutere il grafico dei residui della regressione di Y su X

Soluzione punto 1

· Il grafico di dispersione sarà:



· La formula per calcolare il coefficiente di correlazione lineare è

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$$

. Calcolo allora $\overline{x}_9 = 14.89$ e $\overline{y}_9 = 6.56$ e trovo

$$r = \frac{-1576.44}{\sqrt{910.89 \cdot 3328.22}} = -0.9$$

Troviamo quindi una forte correlazione lineare negativa.

· Trovo la retta di regressione di Y su X con la formula $y = \hat{\beta}x + \hat{\alpha}$. Trovo $\beta = \frac{S_{xy}}{S_{xx}} = \frac{-1576.44}{910.89} = -1.73$ e $\alpha = \overline{y} - \hat{\beta}\overline{x} = 6.56 - (-1.73) \cdot 14.89 = 32.32$ e trovo quindi la retta

$$y = -1.73x + 32.32$$

Soluzione punto 2: Dal formulario è $\left(\hat{\beta} \pm \sqrt{\frac{SS_R}{S_{xx}(n-2)}} t_{n-2,\alpha/2}\right)$ Conosco n=9, $\alpha/2 = 0.005$, $\hat{\beta} = -1.73$, $S_{xx} = 910.89$, mi manca trovare, attraverso formulario e tabelle $SS_R = 599.94$ e $t_{7,0.005} = 3.499$ Dunque l'intervallo è

$$-1.73 \pm \sqrt{\frac{599,94}{7 \cdot 910.89}} = (-2.80, -0.66)$$

Soluzione punto 3: $H_0 : \beta = 0 \quad H_1 : \beta \neq 0$. Rifiutiamo l'ipotesi nulla se (dal formulario)

$$\left| \sqrt{\frac{S_{xx}(n-2)}{SS_R}} \hat{\beta} > t_{n-2,\alpha/2} \right|$$

Calcolo e trovo che $5.64 > 3.499$ e quindi rifiuto H_0. Questo test si poteva fare anche osservando l'IC in quanto

$$\beta \notin (-2.80, -0.66)$$

Soluzione punto 4: Per calcolare la bontà di adattamento del modello si utilizza il coefficiente di determinazione:

$$R^2 = 1 - \frac{SS_R}{S_{yy}} = 1 - \frac{599.94}{3328.22} = 0.82$$

Questo significa che l'82% della variazione delle risposte è giustificata dai predittori, quindi dal modello. Osservazione: dal coefficiente di determinazione ritroviamo il quadrato del coefficiente lineare, che possiamo trovare in modulo. Per il segno, sappiamo che il coefficiente lineare ha lo stesso segno di β

Soluzione punto 5: Per calcolare i residui posso calcolarli con la formula $\overline{e_i} = y_i - \alpha - \beta x_i$. Trovo allora I residui si dispongono in modo abbastanza casuale attorno all'asse x.

x_i	7	17	8	36	23	19	17	2	5
$\bar{e}_i$	4.79	-2.91	-7.48	3.96	5.47	-17.45	5.09	-2.86	11.33