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

  $\rightsquigarrow \quad \overline{x} = \dfrac{\sum_{j=1}^{M} z_j f_j}{\sum_{j=1}^{M} f_j}$ (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: $q_1\ (p = \frac{1}{4}), \quad q_2 = m\ (p = \frac{1}{2}), \quad q_3\ (p = \frac{3}{4})$

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

- 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: 26 maggio 2023

## Spazi di probabilità

Assiomi della probabilità:

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

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: $\quad \mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A)$

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

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

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

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

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

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

## 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: $\mathrm{P}(X \in A) = \displaystyle\sum_{x_i \in A} \mathrm{p}_X(x_i)$
- Valore medio: $\mathrm{E}[X] = \displaystyle\sum_{x_i} x_i \cdot \mathrm{p}_X(x_i)$

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

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

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

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

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

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

- 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**<br>$\mathrm{Be}(p)$<br>$p \in [0,1]$ | $\{0,1\}$ | $\begin{cases} p & \text{se } k = 1 \\ 1-p & \text{se } k = 0 \end{cases}$ | $p$ | $p(1-p)$ |
| **Binomiale**<br>$\mathrm{Bin}(n,p)$<br>$n \in \{1,2,\dots\}$ $p \in [0,1]$ | $\{0,1,\dots,n\}$ | $\binom{n}{k} p^k (1-p)^{n-k}$ | $np$ | $np(1-p)$ |
| **Poisson**<br>$\mathrm{Pois}(\lambda)$<br>$\lambda \in (0,\infty)$ | $\mathbb{N}_0 = \{0,1,\dots\}$ | $\mathrm{e}^{-\lambda}\dfrac{\lambda^k}{k!}$ | $\lambda$ | $\lambda$ |
| **Geometrica**<br>$\mathrm{Geo}(p)$<br>$p \in (0,1]$ | $\mathbb{N} = \{1,2,\dots\}$ | $p(1-p)^{k-1}$ | $\dfrac{1}{p}$ | $\dfrac{1-p}{p^2}$ |

## 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**<br>$\mathrm{U}(a, b)$<br>$a, b \in \mathbb{R}$ con $a < b$ | $[a, b]$ | $\dfrac{1}{b-a}$ | $\dfrac{x-a}{b-a}$ | $\dfrac{a+b}{2}$ | $\dfrac{(b-a)^2}{12}$ |
| **Esponenziale**<br>$\mathrm{Exp}(\lambda)$<br>$\lambda \in (0, \infty)$ | $[0, \infty)$ | $\lambda\,\mathrm{e}^{-\lambda x}$ | $1 - \mathrm{e}^{-\lambda x}$ | $\dfrac{1}{\lambda}$ | $\dfrac{1}{\lambda^2}$ |
| **Normale**<br>$\mathrm{N}(\mu, \sigma^2)$<br>$\mu \in \mathbb{R} \quad \sigma \in (0, \infty)$ | $(-\infty, +\infty)$ | $\dfrac{\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$ | $\Phi(x)$ | $\mu$ | $\sigma^2$ |

## 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 $\rightsquigarrow$ $\begin{cases} X \text{ v.a. discreta} \\ \text{valori assunti } \{x_i\} = \text{punti di discontinuità di } F_X \\ \mathrm{p}_X(x_i) = F_X(x_i) - F_X(x_i^-) \quad \big(\text{con } F_X(x^-) := \lim\limits_{t \to x^-} F_X(t)\big) \end{cases}$

## 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 $\mu$ e varianza $\sigma^2$ 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 $\mu$ 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 $\sigma^2$ ( 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

$\alpha$= livello di significatività

*rc vero-> rifiuto H0 : p-value < alpha*
*rc falso-> accetto H0: p-value > alpha*

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

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

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

- 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 $\mu_X$ e $Y_1, ..., Y_n$ di media $\mu_Y$

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

- Test $t$ sulla differenza delle media di due campioni normali indipendenti $X_1, ..., X_{n_x}$ di media $\mu_x$ e $Y_1, ..., Y_{n_y}$ di media $\mu_y$, con varianza incognita che si suppone uguale, test applicabile se : $1/2 < \dfrac{S_x^2}{S_y^2} < 2$.

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

- Test $\chi^2$ 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 $\geq 1$ e almeno l'80% di esse $\geq 5$.

- Test $\chi^2$ di indipendenza per $X$ e $Y$ ($X$ e $Y$ sono due caratteristiche che vengono osservate su uno stesso membro della popolazione)

  $H_0$: $X$ e $Y$ sono indipendenti

  $X$ assume i valori $\{1, ..., r\}$

  $Y$ assume i valori $\{1, ..., s\}$

  $N_{i,j} = |\{k : (X_k, Y_k) = (i, j)\}|$, $i = 1, ..., r$, $j = 1, .., s$ frequenza assoluta della coppia $(i, j)$

  $N_i^x = |\{k : X_k = i\}|\, i = 1, ..., r$ frequenza assoluta di $i$ ( nella popolazione $X$ )

  $N_j^y = |\{k : Y_k = j\}|\, j = 1, ..., s$ frequenza assoluta di $j$ ( nella popolazione $Y$ )

  $n_{i,j}$, $n_i^x, n_y^y$ siano i valori sulle osservazioni rispettivamente assunti da $N_{i,j}$, $N_i^x, N_j^y$

  Statistica:

$$Q = \sum_{1 \leq i \leq r,\, 1 \leq j \leq s} \frac{\left(N_{i,j} - \frac{N_i^x N_j^y}{n}\right)^2}{\frac{N_i^x N_j^y}{n}}$$

  Regione critica a livello di significatività $\alpha$:

$$q = \sum_{1 \leq i \leq r,\, 1 \leq j \leq s} \frac{\left(n_{i,j} - \frac{n_i^x n_j^y}{n}\right)^2}{\frac{n_i^x n_j^y}{n}} > \chi^2_{(r-1)(s-1),\alpha}$$

## 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 $\alpha$ e $\beta$

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

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

• Somma dei quadrati residui, con $A$ e $B$ stimatori di $\alpha$ e $\beta$:

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

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

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

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



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 $\beta$ 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 $\gamma$ se

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

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

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