Chapter 1

Teoria (il minimo sindacale)

Per esempi ed esercizi seguire i link

.

1 Spazio di probabilità

Esempi: Dado a 4 facce ➦
Doppio lancio moneta ➦
Urna con biglie di 3 colori ➦

Fisssiamo un insieme non vuoto Ω che chiameremo **spazio campionario** (**sample space**) o **popolazione**. Immagimiamo gli elementi $\omega \in \Omega$ come i possibili **risultati** di un rilevamento, un esperimento, un sorteggio, ecc. I sottoinsiemi $E \subseteq \Omega$ verranno chiamati **eventi** e, intuitivamente, rapresentano proprietà osservabili.

Una **misura di probabilità** è una funzione $\Pr : \mathcal{P}(\Omega) \to \mathbb{R}$ tale che

- $\Pr(\Omega) = 1$
- $\Pr(E) \geq 0$ per ogni $E \in \mathcal{P}(\Omega)$
- $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$ per ogni coppia $E_1, E_2 \in \mathcal{P}(\Omega)$ di insiemi **disgiunti**, ovvero $E_1 \cap E_2 = \varnothing$. Si dice anche che E_1 e E_2 sono **eventi mutualmente esclusivi**.

Conseguenze:

- $\Pr(\varnothing) = 0$
- $\Pr(\neg E) = 1 - \Pr(E)$
- $\Pr(E_1 \setminus E_2) = \Pr(E_1) - \Pr(E_2)$ se $E_2 \subseteq E_1$
- $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap \mathrm{E}_2)$ per ogni $E_1, E_2 \in \mathcal{P}(\Omega)$.

.

2 Variabili aleatorie

Esempi: Urna con biglie di dimensioni diverse

Sia R un insieme quasiasi. Una **variabile aleatoria** è una funzione $X : \Omega \to R$. Spesso si scrive $\boldsymbol{X \in R}$ (una notazione che può dar luogo a malintesi) omettendo il riferimento ad Ω.

Se R è un insieme numerico (un sottoinsieme di $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{R}^2$, ecc.) diremo che X è una **variabile aleatoria numerica** o **quantitativa**. Una variabile aleatoria non numerica è detta **qualitativa** o **categorica**.

.

3 Distribuzione di probabilità discrete e contnue

Dato $x \in R$ e $A \subseteq R$ scriveremo

$$\begin{aligned} \boldsymbol{p_x} &= \boldsymbol{\Pr(X = x)} &&= \Pr\big(\{\omega \in \Omega \,:\, X(\omega) = x\}\big) \\ &\quad \boldsymbol{\Pr(X \in A)} &&= \Pr\big(\{\omega \in \Omega \,:\, X(\omega) \in A\}\big) \\ \boldsymbol{F_X(x)} &= \boldsymbol{\Pr(X \le x)} &&= \Pr\big(\{\omega \in \Omega \,:\, X(\omega) \le x\}\big) \quad \text{se } X \text{ è numerica.} \end{aligned}$$

La funzione $\Pr(X = x)$ si chiama **distribuzione di probabilità (probability mass function)**. La funzione $\Pr(X \le x)$ si chiama **funzione di ripartizione (cumulative distribution function)**, spesso indicata con $\boldsymbol{F_X(x)}$.

Le variabili numeriche possono dirsi **discrete** o **continue**. Una v.a. X è discreta se per ogni sottoinsieme $A \subseteq R$

$$\Pr\big(X \in A\big) = \sum_{x \in A} \Pr(X{=}x) \qquad \text{Per la notazione} \rightarrow$$

Ovvero la probabilità è concentrata nei punti di R.

Invece X è una **variabile aleatoria continua** se $\Pr(X{=}x) = 0$ per ogni $x \in R$. Per le variabili continue è significativa solo la probabilità in intervalli di diametro positivo

$$\Pr\big(X \in [a, b]\big) = \Pr(a \le X \le b) = \Pr(X \le b) - \Pr(X \le a)$$

Si noti che la seconda uguaglianza non sarebbe corretta se $\Pr(X{=}a) \neq 0$.

N.B. Esistono variabili aleatorie (anche in esempi concreti) che sono intermedie tra il continuo e il discreto ma per il momento non le considereremo.

.

4 Funzione quantile

La **mediana** (o il **valore mediano**) di una variabile aleatoria X è quel valore x tale che $Pr(X \leq x) = 1/2$.

In generale definiamo la **funzione quantile**. Questa è la funzione che dato $p \in (0, 1)$ ritorna quel valore $x = Q_X(p)$ tale che $p \leq \Pr(X \leq x)\}$. In altre parole $Q_X(p)$ è la *funzione inversa* della funzione di ripartizione $F_X(x) = \Pr(X \leq x)$.

N.B. La definizione qui sopra non è sempre corretta. Può succedere che a più valori di x corrisponda lo stesso valore di $F_X(x)$. (Ovvero F_X non è invertibile.) In questo caso prendiamo il limite inferiore di questi x.

Quindi il valore mediano di X è $Q_X(1/2)$.

.

5 Variabili aleatorie di Bernoulli

Una variabile aleatoria $X : \Omega \to R$ si dice di **Bernoulli** se $R = \{0, 1\}$, che spesso si abbrevia scrivendo $\boldsymbol{X \in \{0, 1\}}$.

Possiamo identificare in modo canonico eventi e variabili aletorie di Bernoulli. L'evento associato ad X è l'insieme $\{\omega : X(\omega) = 1\}$ che chiameremo **successo**.

Viceversa, la v.a. di Bernoulli associata ad un evento E è spesso denotata con $\boldsymbol{1_E}$

$$1_E(x) \quad = \quad \begin{cases} 1 & \text{se } x \in E \\ 0 & \text{se } x \notin E \end{cases}$$

Queste funzione di chiama **funzione indicatrice** (o **caratteristica**) dell'evento E.

Chiameremo $\boldsymbol{p} = \Pr(X{=}1)$ la **probabilità di successo**.

Per dire che X è una variabile aleatoria di Bernoulli con probabilità di successo p scriveremo $\boldsymbol{X \sim B(1, p)}$.

.

6 Probabilità condizionata

Dato $A, \Phi \subseteq \Omega$ tali che $\Pr(\Phi) \neq 0$ definiamo

$$\mathbf{Pr}(\boldsymbol{A} \mid \boldsymbol{\Phi}) \quad = \quad \frac{\Pr(A \cap \Phi)}{\Pr(\Phi)}$$

Questo si legge **probabilità di $\boldsymbol{A}$ dato $\boldsymbol{\Phi}$**. Si verifica facilmente che $\Pr(\,\cdot\,|\Phi)$ soddisfa a tutte le proprietà di $\Pr(\,\cdot\,)$ se rimpiazziamo Ω con Φ.

.

7 Teorema delle Probabilità Totali

Esempi: Popolazione maschile e femminile
Peso neonati
Modello di Hardy-Weinberg

Il fatto seguente si chiama **Teorema delle Probabilità Totali**: siano $A_1, \ldots, A_n$ eventi **mutuamente esclusivi** ed **esaustivi** di probabilità $\neq 0$. Sia C è un qualsiasi altro evento, allora

$$\Pr(C) \quad = \quad \sum_{i=1}^{n} \Pr(A_i) \cdot \Pr\left(C|A_i\right).$$

.

8 Indipendenza stocastica

Due eventi A e B si dicono **(stocasticamente) indipenenti** se

$$\Pr(A \cap B) \quad = \quad \Pr(A) \cdot \Pr(B).$$

Il seguente fatto è facile da verificare: se A e B sono eventi probabilità non nulla allora sono indipendenti se e solo se $\Pr(A|B) = \Pr(A)$ se e solo se $\Pr(B|A) = \Pr(B)$.

Due variabili aleatorie discrete X ed Y si dicono **(stocasticamente) indipenenti** se per ogni $x \in \operatorname{img} X$ e $y \in \operatorname{img} Y$

$$\Pr(X, Y = x, y) \quad = \quad \Pr(X = x) \cdot \Pr(Y = y).$$

Nel caso di variabili aleatorie continue la condizione diventa

$$\Pr(X \leq x \text{ and } Y \leq y) \quad = \quad \Pr(X \leq x) \cdot \Pr(Y \leq y).$$

.

9 Esperimenti ripetuti: prodotto di spazi

Sia $X : \Omega \to \{0, 1\}$ una variabile aleatoria di Bernoulli (per avere un esempio semplice). Immaginiamo che X modelli il lancio di una moneta. Per brevità definiamo $A = \{\omega \in \Omega \ : \ X(\omega) = 1\}$.

Il lancio ripetuto di una monetina è modellato usando mome spazio campionario il prodotto cartesiano di Ω con se stesso: Ω^2.

Per esempio: l'evento $A \times \Omega \subseteq \Omega^2$ corrisponde ad ottenere 1 nel primo lancio. L'evento $\Omega \times \neg A$ corrisponde ad ottenere 0 nel secondo lancio. L'intersezione di questi eventi è $A \times \neg A$. Questo corrisponde ad ottenere nei due lanci la sequenza $1\,0$.

In generale la probabilità di un evento $A \times B \subseteq \Omega^2$ è per definizione $\Pr(A) \cdot \Pr(B)$. Gli eventi $A \times \Omega$ e $\Omega \times B$ sono quindi (per costruzione) indipendenti.

Introduciamo due v.a. di Bernoulli $X_1 : \Omega^2 \to \{0, 1\}$ ed $X_2 : \Omega^2 \to \{0, 1\}$.

La v.a. X_1 guarda solo alla prima coordinata di un risultato in Ω^2 e restituisce lo valore di X. Quindi X_1 restituisce 1 per i risusltati in $A \times \Omega$. La v.a. X_2 fa lo stesso con la seconda coordinata, restituisce 1 per i risusltati in $\Omega \times A$.

.

10 Binomial random variables

Esempi: Estrazione ripetuta

We say that X is a **binomial random variable with parameters n and p**, for short $X \sim B(n,p)$, if

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

where the X_i are independent Bernoulli random variables with success probability p. So, X counts the number of successes in a sequence of n independent experiments (Bernoulli trials with success probability p). Clearly $X \in \{0,\dots,n\}$.

We may also say that X has a **binomial distribution** with parameters n and p. In fact binomial random variables are characterized by their distribution which is not difficult to compute

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

il grafico

The cumulative distribution function is

$$P(X{\leq}k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

.

11 Valore atteso

Il **valore atteso** o **media (di popolazione)** (in inglese **expected value, population mean**) di una variabile aleatoria numerica discreta X a valori in R è

$$\boldsymbol{\mu} \;=\; \mathbf{E}(\boldsymbol{X}) \;=\; \sum_{x \in R} x \cdot \Pr(X = x)$$

La lettera μ viene usata quando è chiaro a quale variabile ci si riferisce. Per evitare ambiguità a volte si scrive $\boldsymbol{\mu_X}$.

ATTENZIONE: non si confonda il concetto di media di popolazione con quello di media campionaria (che introdurremo più avanti). Entrambi vengono spesso abbreviati con **media**!

.

12 Valore atteso della somma di v.a.

Siano X ed Y due variabili aleatorie discrete. Supponiamo di conoscere $\mathrm{E}(X)$ ed $\mathrm{E}(Y)$. La somma $X + Y$ è anche una variabile aleatoria e

$$\begin{aligned} \mathrm{E}(X+Y) &= \sum_{z \in R} z \cdot \Pr(X+Y=z) \\ &= \mathrm{E}(X) + \mathrm{E}(Y) \end{aligned}$$

L'uguaglianza non è difficile da verificare (noi la prendiamo comunque per buona).

È anche interessante notare che se c è una costante qualsiasi

$$\mathrm{E}(cX) = c\,\mathrm{E}(X)$$

Le due proprietà congiunte si chiamano **linearità**, ovvero si dice che il valore atteso è un operatore **lineare**.

.

13 Valore atteso del prodotto di v.a.

Siano X ed Y due variabili aleatorie discrete. Supponiamo di conoscere $\mathrm{E}(X)$ ed $\mathrm{E}(Y)$. Il prodotto XY è anche una variabile aleatoria e

$$\begin{aligned} \mathrm{E}(XY) &= \sum_{z \in R} z \cdot \Pr(XY = z) \\ &= \mathrm{E}(X)\mathrm{E}(Y) \qquad \textbf{se } X \textbf{ e } Y \textbf{ sono indipendenti} \end{aligned}$$

L'uguaglianza non è difficile da verificare (noi la prendiamo per buona).

.

14 Valore atteso di variabili binomiali

Sia $X \sim B(1,p)$ dalla definizione di valore atteso

$$\mathrm{E}(X) = \sum_{x\in\{0,1\}} x \cdot \Pr(X = x)$$

Siccome $\Pr(X = 0) = 1 - p$ e $\Pr(X = 1) = p$

$$= 1 \cdot p + 0 \cdot (1 - p) = p$$

Sia ora $X \sim B(n,p)$, allora possiamo immaginare X come

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

per $X_i \sim B(1,p)$ quindi per la linearità del valore atteso

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i)$$

$$= \sum_{i=1}^{n} p = np.$$

.

15 Teorema delle Probabilità Totali per la media

Esempi: Peso neonati

Il valore medio può essere condizionato ad un qualsiasi evento $\Phi \subseteq \Omega$ tale che $\Pr(\Phi) \neq 0$ definendo (nel caso di una variabile discreta X a valori in R)

$$\mathbf{E}(\boldsymbol{X} \mid \boldsymbol{\Phi}) \;=\; \sum_{x \in R} x \cdot \Pr\big(X{=}x \,|\, \Phi\big).$$

Il fatto seguente è una conseguenza del Teorema delle Probabilità Totali. Siano $A_1, \ldots, A_n$ eventi mutuamente esclusivi ed esaustivi di probabilità $\neq 0$. Sia X una v.a. discreta

$$\mathrm{E}(X) \;=\; \sum_{i=1}^{n} \Pr(A_i) \cdot \mathrm{E}\big(X|A_i\big).$$

La verifica è immediata.

.

16 Varianza

La **varianza** di una variabile aleatoria numerica discreta X a valori in R è

$$\boldsymbol{\sigma^2} \;=\; \mathbf{Var}(\boldsymbol{X}) \;=\; \mathrm{E}\Big(\big(X - E(X)\big)^2\Big)$$

per la definizione di valore atteso

$$\begin{aligned} &= \sum_{x \in R} \Big(x - E(X)\Big)^2 \cdot \Pr(X = x) \\ &= E(X^2) - E(X)^2 \qquad \text{(facile da verificare).} \end{aligned}$$

La **deviazione standard** è la radice della varianza

$$\boldsymbol{\sigma} \;=\; \mathbf{SD}(\boldsymbol{X}) \;=\; \sqrt{\mathrm{Var}(X)}$$

La lettera σ viene usata quando è chiaro a quale variabile ci si riferisce. Per evitare ambiguità a volte si scrive $\boldsymbol{\sigma_X}$.

.

17 Disequazione di Chebyshev

Enunciamo senza dimostrazione un caso particolare di famoso teorema.

Per ogni variabile aleatoria numerica X con valore atteso μ e varianza σ^2.

$$75\% \quad \leq \quad \Pr\Big(\mu - 2\sigma \leq X \leq \mu + 2\sigma\Big)$$

Ovvero gran parte della massa di probailità è concentrata in un intorno di μ di raggio σ. Quindi più piccola è σ più concentrata è a distribuzione di X.

Per alcune particolari v.a. valgono anche proprietà ancora più forti. (Per esempio, se X ha distribuzione normale possiamo scrivere 95% invece che 75%.)

.

18 Varianza della somma di v.a.

Siano X ed Y due variabili aleatorie discrete *indipendenti*. Supponiamo di conoscere $\mathrm{Var}(X)$ ed $\mathrm{Var}(Y)$. La somma $X+Y$ è anche una variabile aleatoria e

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \sum_{z\in R} \big(z - \mathrm{E}(X+Y)\big)^2 \cdot \Pr(X+Y=z) \\
(*) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\big[\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)\big] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) \qquad \textbf{se } X \textbf{ e } Y \textbf{ sono indipendenti}
\end{aligned}$$

L'uguaglianza (*) la prendiamo per buona.

Chiemeremo **covarianza** di X ed Y la quantità

$$\textbf{coVar}(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$

Notare che per quando detto in ↱ se X ed Y sono indipendenti allora il termine $\mathrm{coVar}(X,Y)$ si annulla.

.

19 Varianza di un multiplo di una v.a.

Sia X una variabile aleatoria e sia c una costante. Supponiamo di conoscere $\mathrm{Var}(X)$. La v.a. cX è anche una variabile aleatoria e

$$\mathrm{Var}(cX) \;=\; c^2\,\mathrm{Var}(X)$$

.

20 Varianza di variabili binomiali

Sia $X \sim B(1,p)$ dalla definizione di varianza (e ricordando che $E(X) = p$)

$$\begin{aligned} \mathrm{E}(X) &= \sum_{x \in \{0,1\}} (x-p)^2 \cdot \Pr(X = x) \\ &= (1-p)^2 \cdot p + p^2 \cdot (1-p) \quad = \quad p\,(1-p) \end{aligned}$$

N.B. Possiamo calcolarla così

$$\begin{aligned} &= \mathrm{E}(X^2) - E(X)^2 \\ &= p - p^2 \qquad \text{(perché se } X \sim B(1,p) \text{ allora } X^2 = X\text{).} \end{aligned}$$

Sia ora $X \sim B(n,p)$, allora possiamo immaginare X come

$$X \;=\; \sum_{i=1}^{n} X_i$$

per $X_i \sim B(1,p)$ indipendenti. Quindi

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \quad = \quad \sum_{i=1}^{n} \mathrm{Var}(X_i) \\ &= \sum_{i=1}^{n} p\,(1-p) \quad = \quad n\,p\,(1-p). \end{aligned}$$

.

21 Standardizzazione

Sia X una v.a. con media μ e deviazione standard σ. La variabile aleatoria Z così definita

$$Z = \frac{X - \mu}{\sigma}$$

si dice ottenuta da X per **standardizzazione**. La variabile Z ha media nulla e deviazione standaed 1 ed è sempre adimensionale. Un valore ottenuto da Z si dice **punteggio Z** o **punteggio standard** (**Z-score**).

.

22 Regola di Bayes

Esempi: Fumatori
Hemophilia gene carrier
Rain forcasts
Diagnostic test: HIV

Il fatto seguente di chiama **Teorema (o regola) di Bayes**: per ogni coppia di eventi A e B di probabilità $\neq 0$

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}.$$

In molte applicazioni $\Pr(B)$ viene calcolato usanto il teorema delle probabilità totali. La regola diventa

$$= \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A)}.$$

.

23 Diagnostic tests

Esempi: HIV test (regola di Bayes)

Let T_+ and T_- be the events that the result of a diagnostic test is positive or negative respectively. Let D be the event that the subject of the test has the disease.

Introduciamo un po di terminologia.

- We call $\Pr(D)$ the **prevalence** of the disease. Often it is very difficult to estimate: it strongly depends on the risk category the subject belongs to.
- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $\Pr(T_+|D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $\Pr(T_-|\neg D)$
- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $\Pr(D|T_+)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $\Pr(\neg D|T_-)$
- The **prevalence of the disease** is the marginal probability of disease, $\Pr(D)$

Tipicamente la specificità e la sensitività del test sono note. I poteri predittivi positivi e negativi vengono calcolati usando la prevalenza e regola di Bayes e quindi dipendono fortemente dalla categoria di rischio del cui appartiene il soggetto.

.

24 Test d'ipotesi

Esempi: Test binomiale (il più elementare test di ipotesi)

Nei test di ipotesi la scelta tra risultato positivo/negativo viene fatta in base al valore di una statistica. Si sceglie un intervallo detto **regione di rifiuto**. Se il valore è in questo intervallo l'esito si considera positivo (N.B. rifiuto $\sim$ positivo).

Introduciamo la terminologia dei test d'ipotesi basandoci sulla notazione usata per i test diagnostici.

- L'**ipotesi nulla** denotata con $\boldsymbol{H_0}$ definisce l'insieme dei *sani* (qui H_0 è anche l'evento corrispondente, quello che denotavamo con $\neg D$).
- L'**ipotesi alternativa** denotata con $\boldsymbol{H_A}$ descrive la *patologia*, ovvero definisce l'insieme dei *malati* (qui H_A è anche l'evento corrispondente, era D).
- H_A non è semplicemente la negazione di H_0. Alcune risulttati, se ritenuti impossibili, non occorrono né in H_0 né in H_A.
- L'espressione: **$\boldsymbol{H_0}$ può essere rifiutata** è sinonima di *l'esito del test è positivo*. Noi denotiamo l'evento con $\boldsymbol{T_+}$.
- L'espressione: **$\boldsymbol{H_0}$ NON può essere rifiutata** è sinonima di *l'esito del test è negativo*. Noi denotiamo l'evento con $\boldsymbol{T_-}$.
- Nel progettare il test si decide come definire T_+ e T_- a seconda di quanti falsi positivi/negativi si vuole o può tollerare (in base ai costi/rischi che questi due errori comportano). Ci si calcola quindi $\Pr(T_+|H_0)$ e $\Pr(T_-|H_A)$.

.

25 Test d'ipotesi (tavola riassuntiva)

In questa tavola contrapponiamo la terminologia usata nei **test statistici** a quella dei *test diagnostici*. Molto comuni sono anche i simboli α e β.

$T_+ \cap H_0$ *falso positivo* **errore I tipo** $\Pr(T_+\|H_0) = \alpha$ **significatività**	$T_+ \cap H_A$ *corretto positivo* $\Pr(T_+\|H_A) = 1-\beta$ *sensibilità* **potenza**
$T_- \cap H_0$ *corretto negativo* $\Pr(T_-\|H_0) = 1-\alpha$ *sepecificità*	$T_- \cap H_A$ *falso negativo* **errore II tipo** $\Pr(T_-\|H_A) = \beta$

N.B. È vacile progettare un test che minimizza una tra $\Pr(T_+|H_0)$ o $\Pr(T_-|H_A)$. In un caso estremo: se a prescindere dai dati rifiuta sempre H_0 avrà banalmente $\Pr(T_-|H_0) = 0$; invece un test che non rifiuta mai H_0 avrà $\Pr(T_+|H_0) = 0$. La difficoltà nel progettare il test è trovare il giusto equilibrio tra i due errori.

.

26 Il p-valore

Diamo due definizioni equivalenti di **p-valore**. Sia W una statistica e sia w il valore osservato.

- Il p-value di w è il minimo α che permette di rigettare H_0.
- Il p-value di w è la probabilità di osservare un risultato almeno tanto estremo quanto w, nel caso H_0 sia vera.

La seconda definizione suona più semplice ma bisogna precisare cosa si intende per estremo. Tipicamente H_0 prevede un certo valore w_0 per la statistica. Il p-valore è la probabilità

$$\begin{aligned} \textbf{p-valore} &= \Pr\big(|W - w_0| \geq |w - w_0|\big) \\ &= \Pr\big(W \leq w_0 - |w_0 - w|\big) + \Pr\big(W \geq w_0 + |w - w_0|\big). \end{aligned}$$

Quando possibile si effettua una trasformazione di coordinate in modo tale da avere $w_0 = 0$. Questo semplifica l'espressione del p-valore che diventa

$$= \Pr\big(W \leq |w|\big) + \Pr\big(W \geq |w|\big).$$

Comunque, quando H_A indica che il test è ad una coda, quindi il p-valore si riduce a $\Pr\big(W \leq w\big)$ o $\Pr\big(W \geq w\big)$ a seconda se si tratta della coda superiore o i quella inferiore.

.

27 Campioni e statistiche

Un campione $\{X_1, \ldots, X_n\}$ è un insieme di v.a. indipendenti e identicamente distribuite. Il numero n si chiama **rango** (o **dimensione**) del campione.

Una **statistica** è una variabile aleatoria a valori in $\mathbb{R}$ ottenuta come funzione delle variabili aleatorie di un campione. Una statistica che ha come valore atteso un certo parametro di una distribuzione (vedremo solo μ e σ) si chiama uno **stimatore** di quel parametro.

Gli esempi più noti sono $\bar{X}$, la **media campionaria** ed S, lo **stimatore della deviazione standard**

$$\begin{aligned} \boldsymbol{\bar{X}} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \boldsymbol{S} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \end{aligned}$$

.

28 La distribuzione normale

Esempi: Z-test

Diremo che la v.a. X a valori reali ha **distribuzione normale standard** abbreviato con $X \sim N(0,1)$, se

$$\Pr(X \le x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, \mathrm{d}t$$

il grafico di $\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$

Più in generale X ha **distribuzione normale** con media μ e varianza σ^2, abbreviato con $X \sim N(\mu, \sigma^2)$, se

$$\Pr(X \le x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2}\, \mathrm{d}t$$

Importante proprietà: se $X_1, \ldots, X_n$ sono variabili aleatorie indipendenti con distribuzione $N(\mu, \sigma^2)$ allora la variabile media compionaria

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ha distribuzione $N\left(\mu, \frac{\sigma^2}{n}\right)$.

La deviazione standard di $\bar{X}$ si chiama **errore standard** è quindi $\sigma/\sqrt{n}$.

.

29 The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample (i.e. independent and identically distributed random variables) from *any* distribution with mean μ and variance σ^2.

Then $\bar{X}$ has mean $\mathrm{E}(\bar{X}) = \mu$ and variance $\mathrm{Var}(\bar{X}) = \sigma^2/n$. Moreover, if the sample size n is *sufficiently large* then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has *approximately* distribution $N(0, 1)$.

.

30 La distribuzione t di Student

Esempi: Z-test

Diremo che la v.a. X a valori reali ha **distribuzione t di Student** con $\nu = n - 1$ **gradi di libertà** abbreviato con $X \sim t(\nu)$, se

$$\Pr(X \leq x) \;=\; C_n \int_{-\infty}^{x} \left(1 + \frac{t^2}{n-1}\right)^{-n/2} \mathrm{d}t$$

dove C_n è un opportuna costante che dipende da n.

.

31 La distribuzione t di Student

Siano $X_1, \ldots, X_n$ un campione di v.a. normali con valore atteso μ_0 (supposto noto) e varianza σ (supposta ignota). Posto

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\boldsymbol{S} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Allora $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ ha distribuzione $t(n-1)$.

.

32 Intervallo di confidenza

32.1 Intervallo di confidenza per normale con varianza nota

Consideriamo v.a. $X \sim N(\mu, \sigma^2)$ con σ nota e μ ignota. Sia $\bar{X}$ la media campionaria. Fissiamo un $\varepsilon > 0$.

La probabilità che $\bar{X}$ sia a distanza inferiore a ε da μ è

$$\begin{aligned} \Pr(\mu-\varepsilon < \bar{X} < \mu+\varepsilon) &= \Pr(-\bar{X} - \varepsilon < -\mu < -\bar{X} + \varepsilon) \\ &= \Pr(\bar{X} - \varepsilon < \mu < \bar{X} + \varepsilon) \\ &= 1 - \alpha \qquad \text{(questa è una definizione di } \alpha) \end{aligned}$$

Se in un esperimento misuriamo $\bar{x}$, diremo che $\mu = \bar{x} \pm \varepsilon$ con confidenza $1 - \alpha$, o che $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$ è un intervallo di confidenza di livello $1 - \alpha$.

L'interpretazione 'e che se ripetiamo l'esperimento, con probabilità $1 - \alpha$ ritroveremo un intervallo che contiene μ.

N.B. L'intervallo è la variabile alearoria, μ è un numero fissato, purché ignoto.

Chapter 2

Esempi ed esercizi

.

1 Spazio di probabilità

1.1 Dado con quattro facce

Cosideriamo un dado con 4 faccie (un tetraedro regolare) le faccie sono etichettate con le lettere A, C, G, T.

Come spazio campionario è naturale usare l'insieme $\Omega = \{A, C, G, T\}$ la misura di probabilità $\Pr : \mathcal{P}(\Omega) \to \mathbb{R}$ è univocamente determinata dalle condizioni

$\Pr(\{A\}) = \Pr(\{C\}) = \Pr(\{G\}) = \Pr(\{T\}) = 1/4$

N.B. In futuro abbrevieremo $\Pr(\{A\})$ con $\Pr(A)$, ecc.

La misura su un qualsiasi altro evento $E \subseteq \Omega$ è determinata dalla condizione

$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$ per ogni $E_1, E_2 \in \mathcal{P}(\Omega)$ disgiunti.

.

1.2 Doppio lancio della monetina

Lanciamo due volte una monetina. I possibili risultati sono TT, CC, TC, CT.

Come spazio campionario è naturale usare l'insieme $\Omega = \{TT, CC, TC, CT\}$ la misura di probabilità $\Pr : \mathcal{P}(\Omega) \to \mathbb{R}$ è univocamente determinata dalle condizioni

$\Pr(TT) = \Pr(CC) = \Pr(TC) = \Pr(CT) = 1/4$

L'evento *Esce una volta T ed una volta C* corrisponde all'evento $\{TC, CT\}$.

L'evento *Esce almeno una volta T* corrisponde all'evento $\{TC, CT, TT\}$.

.

1.3 Urna con bigle di 3 colori

Estraimo una biglia da urna che contiene 16 biglie che differiscono solo nel colore.

Le biglie sono 8 rosse, 6 blu, 2 nere.

Esistono due scelte naturali per lo spazio di probabilità. Posto $\Omega = \{r, b, n\}$ stipuliamo che

$$\Pr(r) = 1/2 \qquad \Pr(b) = 3/8 \qquad \Pr(n) = 1/8$$

Oppure $\Omega = \{1, \ldots, 16\}$ e posto $B = \{1, \ldots, 8\}$, $R = \{9, \ldots, 14\}$, $N = \{15, 16\}$ stipuliamo che

$$\Pr(i) = 1/16 \qquad \text{per ogni } i = 1, \ldots, 16.$$

In questo modo diamo una misura di probabilità a molti eventi che non sono in realtà osservabili.

.

2 Variabili aleatorie

2.1 Urna con biglie di dimensioni diverse

Un urna Ω che contiene biglie che differiscono per peso diametro e colore. Possiamo immaginarci tre variabili aletorie:

X peso della biglia	variabile quantiativa
Y diametro della biglia	variabile quntiativa
Z colore della biglia	variabile qualitativa

Se l'urna contiene un piccolo numero di biglie X ed Y sono variabili discrete.

L'urna potrebbe contenere un umero così grande di biglie d rendere più semplice immaginarsi che ce ne siano in numero infinito. In questo caso potrebbe essere ragionevole considerare X ed Y come variabili continue. Ma non sempre. Se per esempio le misure non hanno grossa precisione potrebbe convenire interpretarle come variabili discrete.

.

3 Pobabilità totali

3.1 Maschi e femmine patologia

In letteratura è riportata la prevalenza di una certa patologia nella popolazione femminile (5%) e nella popolazione maschile (3%). Che prevalenza dovremo aspettarci in una popolazione composta dal 60% di maschi e dal 40% di femmine?

F evento: femmina

M evento: maschio

A evento: affetto dalla patologia

$P(F) = 0.4$

$P(M) = 0.6$

$P(A|F) = 0.05$

$P(A|M) = 0.03$

$P(A) = P(A|F)P(F) + P(A|M)P(M) = 0.05 \cdot 0.4 + 0.03 \cdot 0.6 = 0.038 = 3.8\%$

.

3.2 Peso neonati

Il peso medio dei neonati (non prematuri) in Europa è di 3.5kg per i maschi e di 3.4 per le femmine. Assumendo in prima approssimazione che il rapporto tra i sessi sia $1:1$, qual è il peso medio dei neonati alla nascita (indistintamente dal sesso)?

X variabile aleatoria peso del neonato

M evento neonato maschio, $\Pr(M) = 1/2$, $E(X|M) = 3.5$

F evento neonato femmina, $\Pr(F) = 1/2$, $E(X|F) = 3.4$

$$E(X) = E(X|M) \cdot \Pr(M) + E(X|F) \cdot \Pr(F) = 3.5 \cdot \frac{1}{2} + 3.4 \cdot \frac{1}{2} = 3.45$$

.

3.3 Modello di Hardy-Weinberg (1)

In un locus possono occorrere due alleli a, b.

- specie dipoide
- accoppiamento casuale (random mating) nessuna selezione
- specie monoica (??)
- le generazioni non si sovrapongono

Alla generazione 0-*esima* la popolazione è composta da individui con genotipo aa, bb e ab nelle proporzioni p_{aa}, p_{bb}, e $p_{ab} = 1 - p_{aa} - p_{bb}$.

Quali saranno le proporzioni alla *prima generazione*? (Chiamiamole q_{aa}, q_{bb}, e $q_{ab} = 1 - q_{aa} - q_{bb}$)

Denotiamo con aa l'evento *ha genotipo aa alla prima generazione*. Analgamente per gli altri genotipi. Denotiamo con aa-ab l'evento: i genitori hanno genotipo aa-ab. Analgamente per le altre coppe di genotipi.

Osserviamo che $\Pr(aa|aa\text{-}bb) = \Pr(aa|ab\text{-}bb) = \Pr(aa|bb\text{-}bb) = 0$ quindi

$$\begin{aligned} q_{aa} &= p_{aa}^2 \Pr(aa|aa\text{-}aa) + 2\,p_{aa}p_{ab} \Pr(aa|aa\text{-}ab) + p_{ab}^2 \Pr(aa|ab\text{-}ab) \\ &= p_{aa}^2 \cdot 1 + 2\,p_{aa}p_{ab} \cdot \frac{1}{2} + p_{ab}^2 \cdot \frac{1}{4} \\ &= p_{aa}^2 + p_{aa}p_{ab} + \frac{1}{4}\,p_{ab}^2\,; \end{aligned}$$

per simmetria con q_{aa}

$$q_{bb} = p_{bb}^2 + p_{bb}p_{ab} + \frac{1}{4}\,p_{ab}^2\,;$$

.

3.4 Equilibrio di Hardy-Weinberg (2)

Mostriamo ora che q_{aa}, q_{bb}, e q_{ab} dipendono solo dalla frazione di copie dell'allele a presenti nella popolazione.

Se n è il numero di individui, $2n$ è il numero totale di copie dei due alleli. La frazione di a sul totale è (la indichiamo con p_a)

$$\# \qquad p_a \;=\; \frac{2\,n\,p_{aa} + n\,p_{ab}}{2n} \;=\; p_{aa} + \frac{1}{2}\,p_{ab} \;=\; \frac{1}{2}\left(1 + p_{aa} - p_{bb}\right)$$

È immediato verificare che (sorprendentemente) p_a determina q_{aa}, q_{bb}, e q_{ab}

$$q_{aa} \;=\; p_a^2; \qquad\qquad q_{bb} \;=\; (1-p_a)^2; \qquad\qquad q_{ab} \;=\; p_a(1-p_a).$$

A questo punto è intuitivo che la frazione di copie dell'allele a non cambia di generazione in generazione. Verifichiamolo numericamente applicando la formula $\#$ alle frequenze nella prima generazione.

Indichiamo con q_a la frazione di copie dell'allele a alla prima generazione

$$q_a \;=\; \frac{1}{2}\left(1 + q_{aa} - q_{bb}\right) \;=\; \frac{1}{2}\left(1 + p_a^2 - (1-p_a)^2\right) \;=\; p_a$$

Quindi $q_a = p_a$ e diconseguenza q_{aa}, q_{bb}, e q_{ab} sono le proporzioni dei diversi genotipi anche nelle generazioni successive alla prima.

.

4 Regola di Bayes

Regola di Bayes

4.1 Fumatori e non fumatori

Tra le persone affette da una certa patologia, il 20% è fumatore. La prevalenza nella popolazione generale è del 2%. Il 10% della popolazione fuma. Calcolare la probabilità che un fumatore sia affetto da questa patologia.

F insieme dei fumatori

A insieme persone affette da A

$\Pr(A) = 0.02$ prevalenza nella popolazione generale

$\Pr(F) = 0.1$ frazione di fumatori nella popolazione generale

$\Pr(F|A) = 0.2$ prevalenza tra i fumatori

$$\Pr(A|F) = \frac{\Pr(F|A) \cdot \Pr(A)}{\Pr(F)} = \frac{0.2 \cdot 0.02}{0.1} = 0.04 \quad .$$

4.2 Hemophilia gene carrier

Hemophilia is a disease that exhibits X-chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her two X-chromosomes is not affected. The disease is generally fatal for women who inherit two such genes.

Consider a woman who has an affected brother, which implies that her mother must be a carrier of the hemophilia gene. We are also told that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene.

Suppose she has a son (from a man who is not affected) that is not affected. What is the probability that she is a carrier?

Ω set of women

C set of women that are carrier

S_{na} set of women that have one son, and this is not affected

$$\Pr(C) = 1/2$$

$$\Pr(S_{na}|C) = 1/2$$

$$\Pr(C|S_{na}) = \frac{\Pr(S_{na}|C)\cdot\Pr(C)}{\Pr(S_{na})} = \frac{\Pr(S_{na}|C)\cdot\Pr(C)}{\Pr(S_{na}|C)\cdot\Pr(C)+\Pr(S_{na}|\neg C)\cdot\Pr(C)}$$

$$= \frac{1/4}{1/4+1/2} = 1/3$$

.

4.3 Rain forcasts

Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years it has rained only 5 days each year. But the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the times. What is the probability that it will rain on the day of Marie's wedding?

R	event: it rains on Marie's wedding
T_+	event: the weatherman predicts rain
$\Pr(R) = 5/365$	it rains 5 days out of the year
$\Pr(\neg R) = 1 - \Pr(R) = 360/365$	
$\Pr(T_+\|R) = 0.9$	when it rains, 90% of the times rain is predicted
$\Pr(T_+\|\neg R) = 0.1$	when it does not rain, 10% of the times rain is predicted

We want to know

$$\Pr(R\,|T_+) = \frac{\Pr(R)\cdot\Pr(T_+|R)}{\Pr(T_+)}$$

$$= \frac{\Pr(R)\cdot\Pr(T_+|R)}{\Pr(T_+|R)\cdot\Pr(R) + \Pr(T_+|\neg R)\cdot\Pr(\neg R)}$$

.

5 Indipendenza

Lanciamo una moneta $2n$ volte. Modelliamo l'esperimento con una sequenza $X_0, \ldots, X_{2n-1}$ di variabili di Bernoulli. N.B. cominciamo ad enumerare da 0. Dire quali delle seguenti coppie di variabili aleatorie X, Y sono indipendenti.

1. $X = \sum_{i=0}^{n-1} X_i \qquad Y = \sum_{i=n}^{2n-1} X_i.$

2. $X = \sum_{i=0}^{n-1} X_{2i} \qquad Y = \sum_{i=0}^{n} X_{2i-1}.$

3. $X = \#\{i < n \mid X_{2i} \neq X_{2i+1}\};$
   $Y = \#\{i < n \mid X_{2i+1} \neq X_{2i}\}.$

4. $X = 0$ se $X_0 \neq X_1$ altrimenti $= 1$.
   $Y = 0$ se $X_1 \neq X_2$, altrimenti $= 1$.

.

6 Distribuzione binomiale

Binomail random variables

6.1 Estrazione ripetuta

Abbiamo un urna con 36 biglie 21 biglie rosse e 15 blu. Estraiamo una biglia, ne annotiamo il colore e la reintroduiamo nell'urna (**reimbussolamento** in inglese **replacement**) per $n = 7$ volte.

Qual è la probabilità di estrarre esattamente $k = 3$ biglie rosse? Chiamiamo X la v.a. che conta il numero di biglie rosse estratte. Allora $X \sim B(6, p)$ con $p = 21/36 = 7/12$.

$$\begin{aligned} \Pr(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \binom{7}{3} \left(\frac{7}{12}\right)^3 \left(\frac{5}{12}\right)^4 = \frac{7!}{3! \cdot 4!} \cdot \frac{7^3 \cdot 5^4}{12^7} = 20.94\% \end{aligned}$$

.

7 Diagnostic test: HIV

Regola di Bayes

A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%

Suppose that a subject, from a population with a 0.1% prevalence of HIV, receives a positive test result. What is the probability that this subject has HIV?

Mathematically, we want $\Pr(D|T_+)$ given the sensitivity, $\Pr(T_+|D) = .997$, the specificity, $\Pr(T_-|\neg D) = .985$, and the prevalence $\Pr(D) = .001$

$$\begin{aligned}
\Pr(D\,|T_+) &= \frac{\Pr(T_+|D)\Pr(D)}{\Pr(T_+)} \\
&= \frac{\Pr(T_+|D)\Pr(D)}{\Pr(T_+|D)\ \Pr(D) + \Pr(T_+|\neg D)\ \Pr(\neg D)} \\
&= \frac{\Pr(T_+|D)\Pr(D)}{\Pr(T_+|D)\Pr(D) + \left[1 - \Pr(T_-|\neg D)\right]\ \left[1 - \Pr(D)\right]} \\
&= 0.062
\end{aligned}$$

The positive predictive value is 6% for this test. In this population a positive test result only suggests a 6% probability that the subject has the disease.

The low positive predictive value is due to low prevalence of disease and the somewhat modest specificity

Suppose it was known that the subject was an intravenous drug user and routinely had intercourse with HIV infected partners that the test was taken in South Africa where the prevalence is estimated to be around 20%

$$\Pr(D\,|T_+) = 0.943$$

.

8 Test Binomiale

Test di ipotesi

Nella pratica questo test è spesso sostituito da un test sulle proporzioni (uno Z-test o, a volte, un χ^2-test). In questa versione il test è computazionalmente più pesante ma concettualmente più semplice.

8.1 Test a una coda

Un'urna contiene monete equilibrate e monete difettose. Le monete equilibrate hanno probabilità di successo $p = 1/2$ le monete difettose hanno probabilità di successo ignota $p > 1/2$. Non conosciamo la frazione di monete difettose. Questi dati vengono riassunti scrivendo

$H_0 : \quad p = 1/2$

$H_A : \quad p > 1/2$

Estraiamo una moneta dall'urna e, per decidere tra equilibrata o difettosa, facciamo il seguente test: la lanciamo n volte e se il numero dei successi è $\geq k$ la dichiariamo difettosa. Stiamo descivendo una famiglia di test, uno per ogni scelta dei parametri n e k. Vogliamo vedere come variano gli errori del I e del II tipo al variare di questi parametri.

Il test è una variabile aleatoria X a valori in $\{0, \dots, n\}$. Lo spazio campionario Ω è diviso in due parti: H_A e H_0. L'insieme H_A contiene quegli ω che corrispondono a n lanci fatti con una moneta difettosa mentre H_0 contiene quegli ω che corrispondono a lanci con una moneta equilibrata.

Condizionando a H_0 otteniamo $X \sim \mathrm{B}(n, 1/2)$. Condizionando a H_A otteniamo $X \sim \mathrm{B}(n, p)$ con $p > 1/2$ ignota.

.

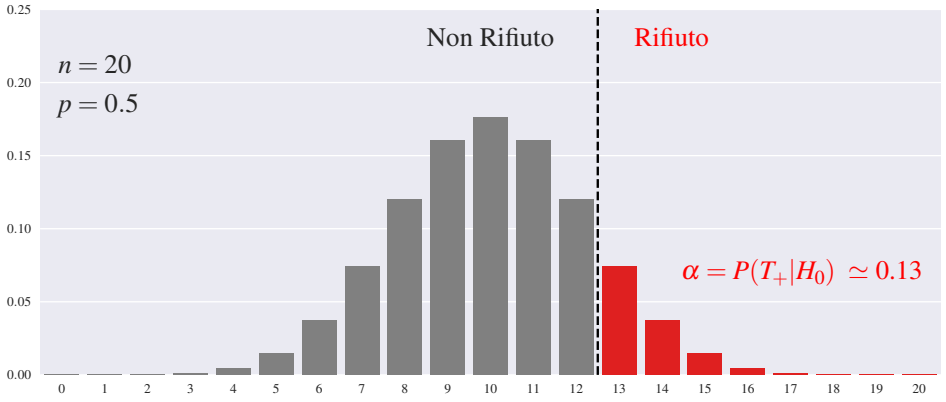
8.2 Test a una coda, errore I tipo

Indichiamo con T_+ l'evento $\{\omega \in \Omega : X(\omega) \geq k\}$, ovvero il risultato del test positivo. N.B. dipende da n e da k. si

Per quanto osservato sulla distribuzione di X, possiamo calcolare la specificità del test (probabilità di falsi positivi)

$$\begin{aligned} \Pr(T_+ \mid H_0) &= \Pr(X \geq k \mid H_0) \\ &= \sum_{i=k}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \frac{1}{2^n} \sum_{i=k}^{n} \binom{n}{i} \end{aligned}$$

Per concretezza, fissiamo $n = 20$, $k = 13$ quindi $\boldsymbol{T_+ = \{13, \ldots, 20\}}$ è la regione di rifiuto. Otteniamo



.

8.3 Test a una coda, errore II tipo

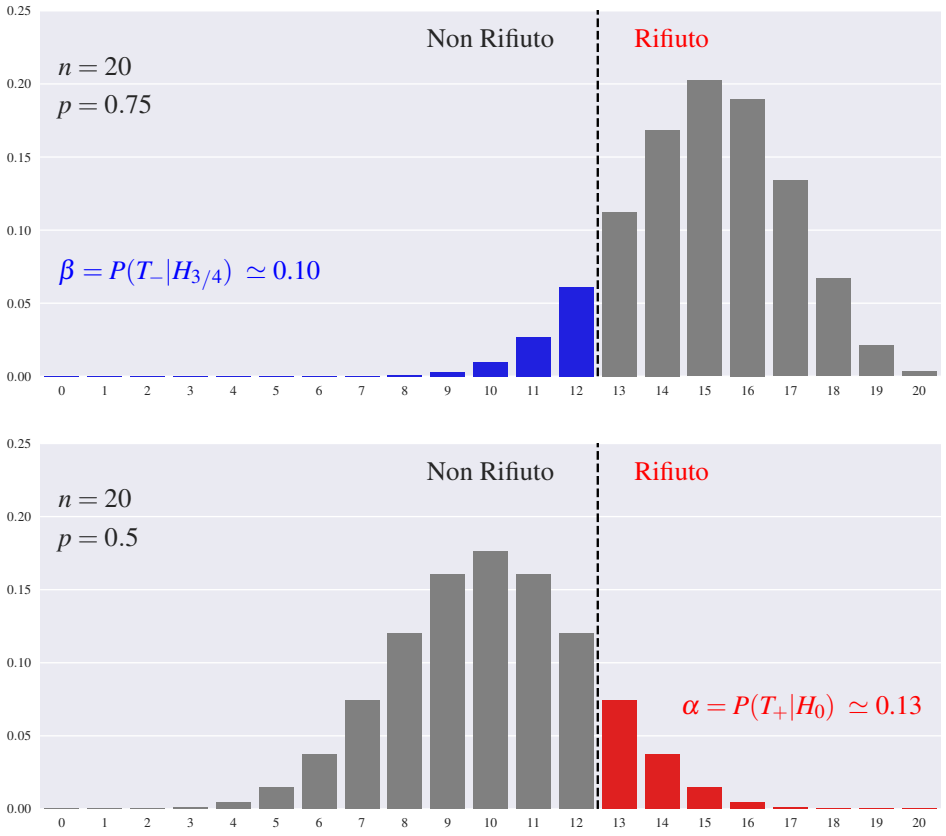
La probabilità dei falsi negativi può essere espressa in funzione di p (abbiamo solo assunto che $p > 1/2$)

$$\Pr(T_- \mid H_A) \quad = \quad \Pr(X < k \mid H_A) \quad = \quad \sum_{i=1}^{k-1} \binom{n}{i} p^i (1-p)^{n-i}$$

Se segliamo come prima $n = 20$, $k = 13$ abbiamo $\boldsymbol{T_- = \{0, \ldots, 12\}}$ è la **zona di NON rifiuto**. Ora, per semplificare la discussione supponiamo di conoscere non solo il tipo ma anche la gravità del difetto. Quindi l'ipotesi alternativa diventa

$H_{3/4}$: $\quad p = 3/4$

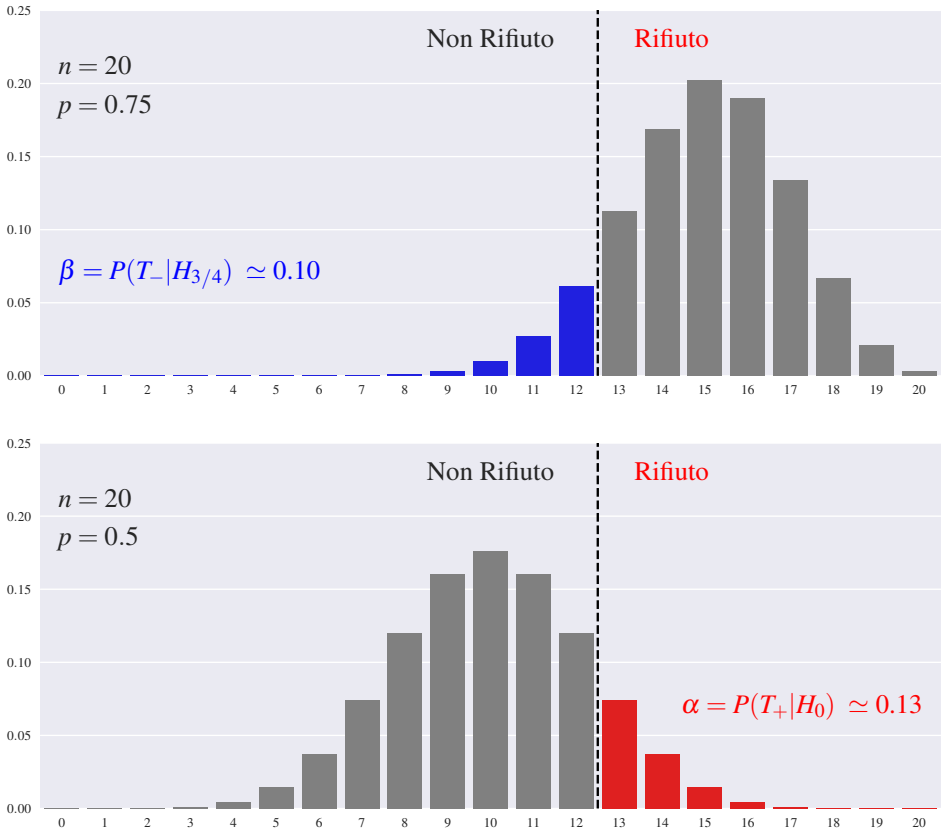
Rappresentiamo la distribuzione di X nel caso in cui vale $H_{3/4}$. Per confronto lo accostiamo al grafico del paragrafo precedente.



8.4 Effect size: δ

Cosa possiamo dire sul caso generale $H_A : p > 1/2$?

Al crescere di p la distribuzione si sposta verso destra quindi la probabilità di errori del II tipo diminuisce. Di converso, se p si avvicina a $1/2$ la probabilità d'errore aumenta. Al limite quando $p \approx 1/2$ avremo $\alpha + \beta \approx 1$. Dobbiamo quindi fissare il minima differenza δ che riteniamo significativa e calcolare β a paritire da quello.



8.5 Test a due code

Nell'esempio precedente avevamo un'informazione certa sul tipo di difetto delle monete: sapevamo che $p > 1/2$. Proviamo a fare senza, avremo quindi

H_0 : $\quad p = 1/2$

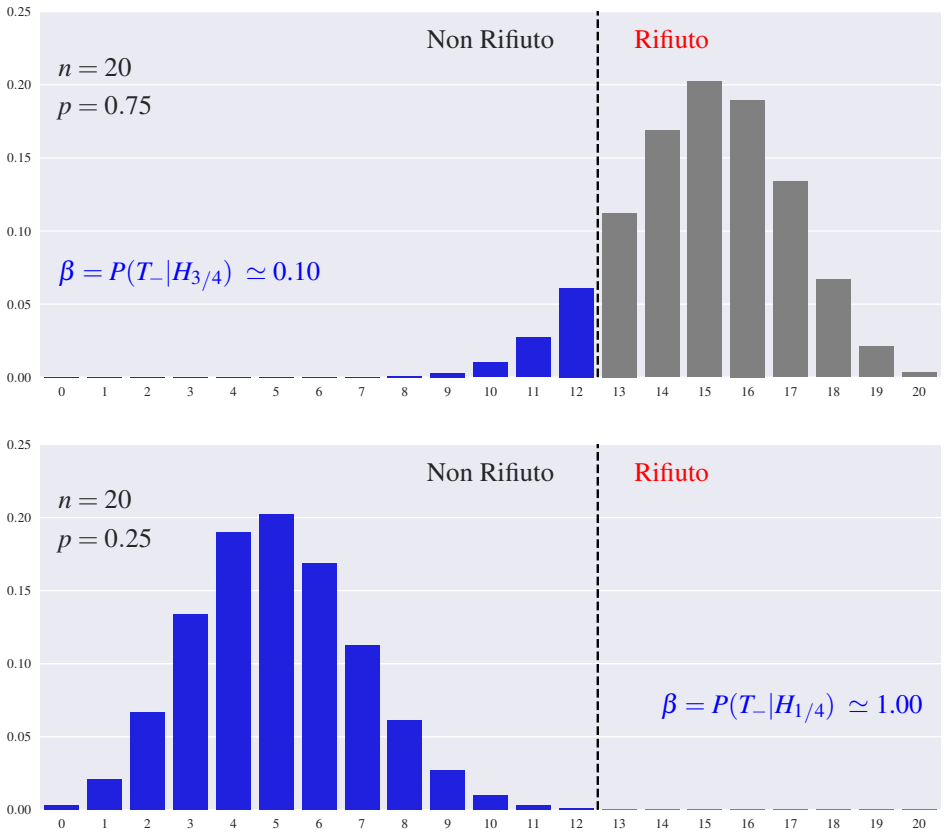
H_A : $\quad p \neq 1/2$

Verifichiamo prima che **la zona di rifiuto dei paragrafi precedenti NON è adatta** alla nuova situazione. L'analisi di $\Pr(T_+|H_0)$ rimane invariata (inaftti l'insieme H_0 non è cambiato).

Per semplificare la discussione dell'errore del II tipo supponiamo per il momento che

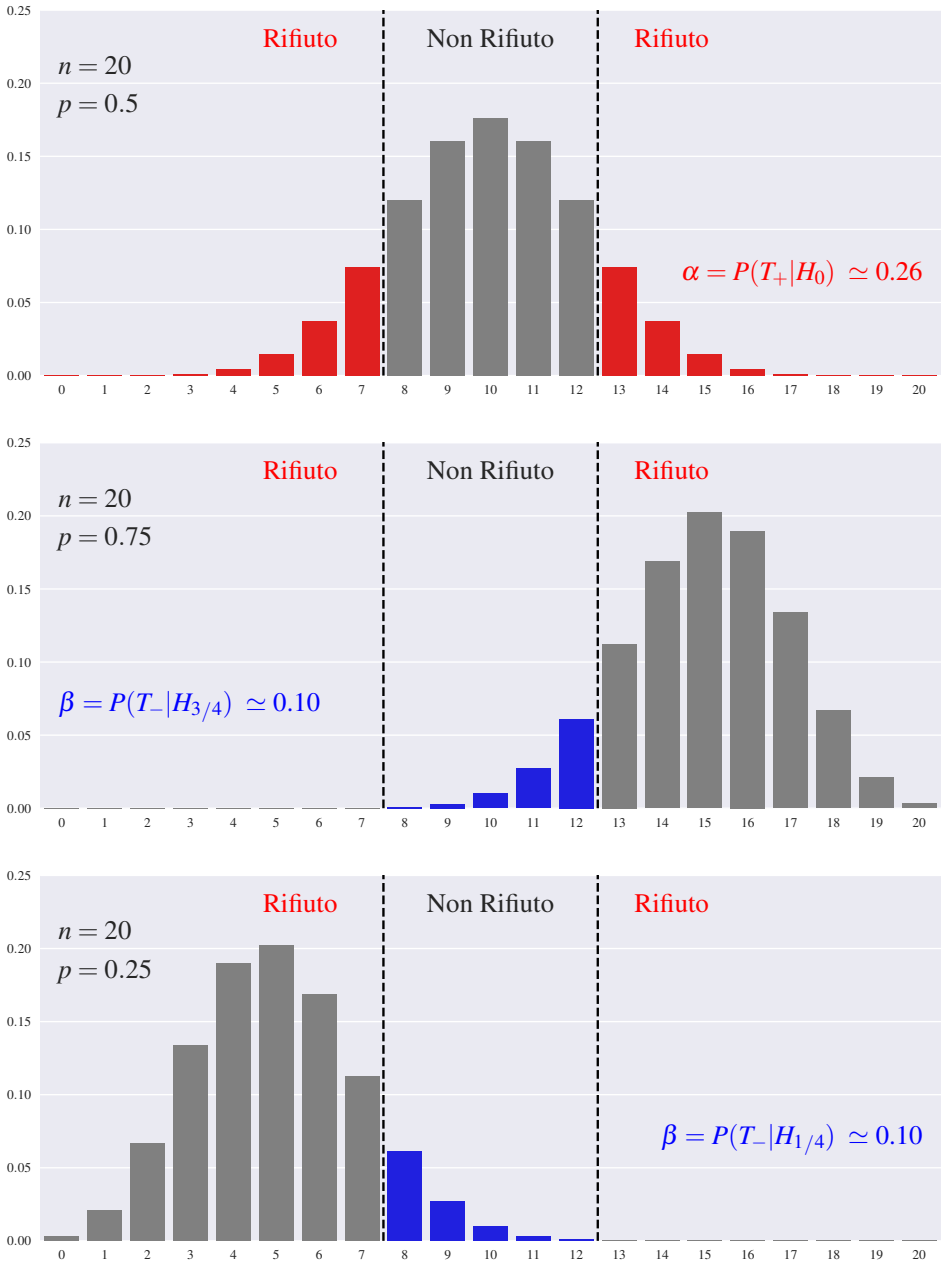
H_A : $\quad p = 3/4 \quad$ o $\quad p = 1/4$

Possiamo immaginare $H_A = H_{1/4} \cup H_{3/4}$. Se sostituiamo H_A con $H_{3/4}$ il grafico rimane come quello già discusso. Ora però dobbiamo considerare il caso il cui la moneta appartenga all'insieme $H_{1/4}$



8.6 Test a due code, errori I e II tipo

Per riparare il problema discusso al paragrafo precedente. prendiamo come zona di rifiuto $T_+ = \{0, \ldots, 7 = n - k\} \cup \{k = 13, \ldots, 20 = n\}$

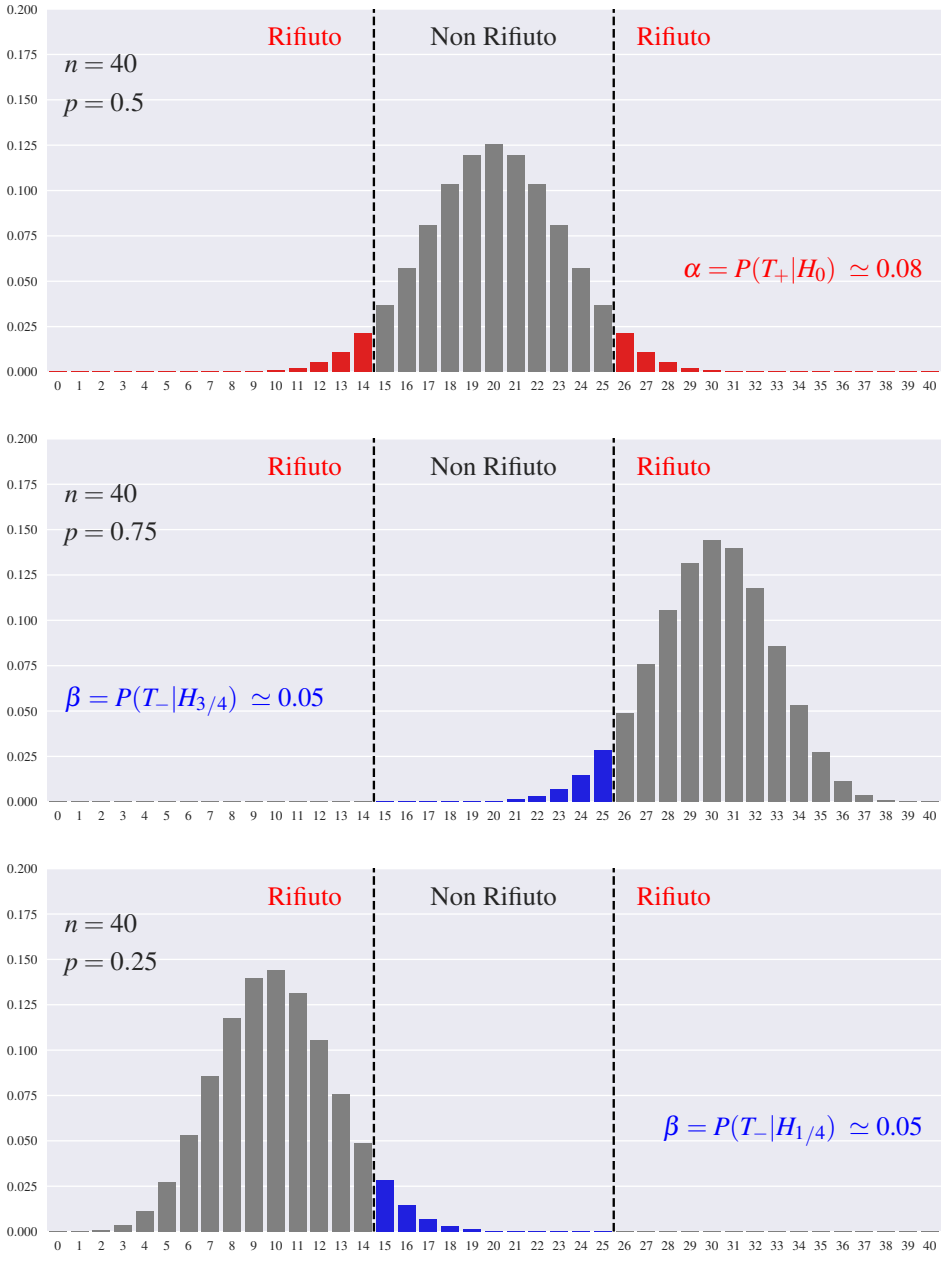


8.7 Test a due code, con campione più ampio

Supponiamo di raddoppiare la dimensione del campione ($n = 40$). Aggiustiamo la zona di rifiuto allo stesso modo ($k = 26$):

$$T_+ = \{0, \ldots, 14 = n - k\} \cup \{k = 26, \ldots, 40 = n\}$$

Entrambi gli errori diminuiscono.



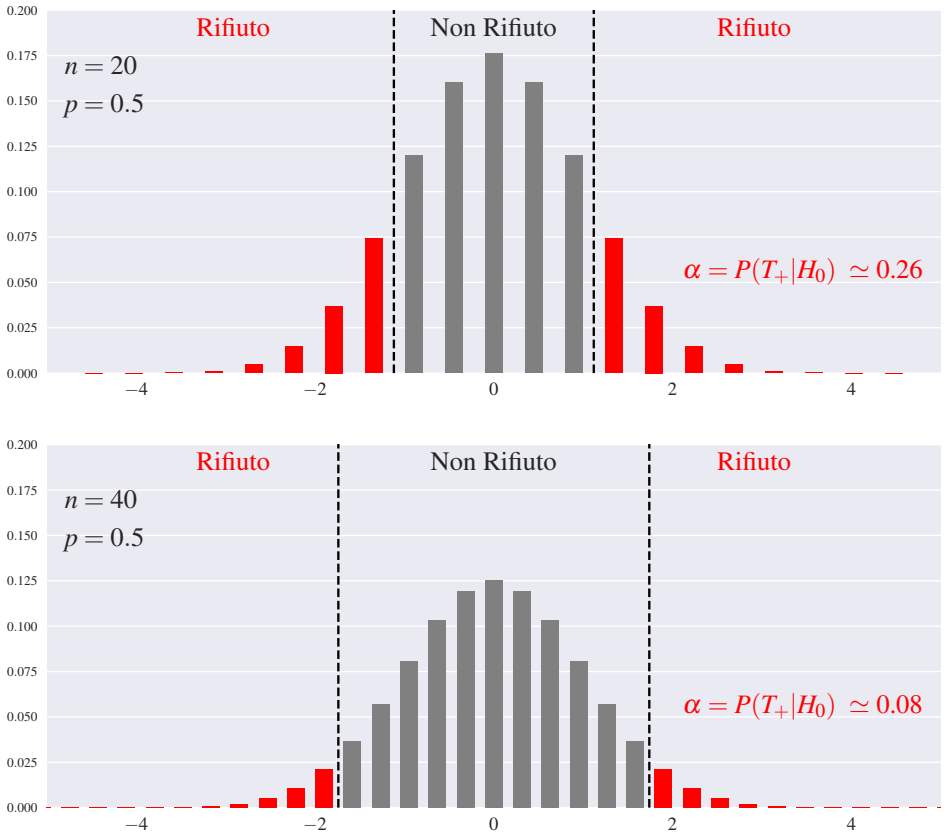
8.8 Standardizzazione (1)

Il confronto tra i test con $n = 20, 40$ non é immediato (vedi la trasformazione della zona di rifiuto). Per facilitare il confronto tipicamente la variabile viene standardizzata.

Se $X \sim B(n, p)$ allora, ricordando che la media è $\mu = np$ e la varianza è $\sigma^2 = np(1-p)$, la variabile standardizzata diventa

$$\begin{aligned} Z &= \frac{X - \mu}{\sigma} \\ &= \frac{X - np}{\sqrt{np(1-p)}} \end{aligned}$$

Gli esempi considerati (con $n = 20, 40$ e $k = 13, 26$) diventano:



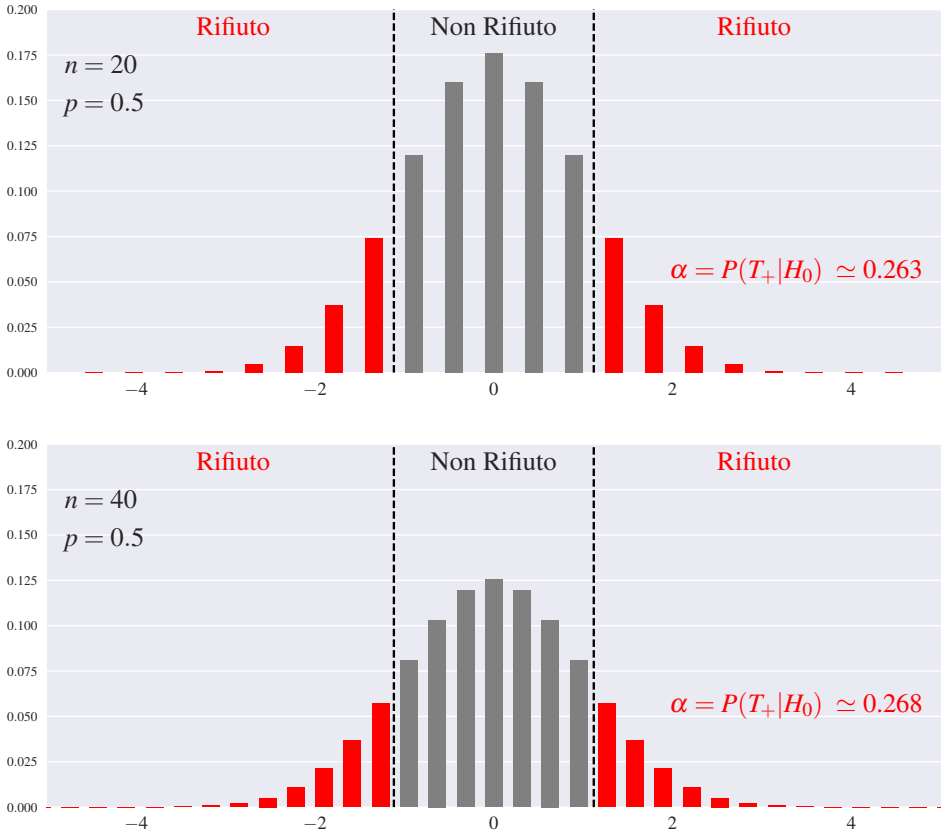
8.9 Standardizzazione (2)

Si noti che le regioni di rifiuto non sono le stesse. In effetti

$$\frac{k-np}{\sqrt{np(1-p)}} = 1.34 \qquad \text{se } n=20,\ k=13,\ p=05$$

$$= 1.90 \qquad \text{se } n=40,\ k=26,\ p=05$$

Se per entrambi i casi prendiamo la stessa regione di rifiuto misurata in punteggio Z, diciamo $(-\infty,-1.34]\cup[1.34,+\infty)$, che corrisponde a $k=24$ o 25, avremmo ottenuto praticamente lo stesso α. Infatti una volta standardizzate le due distribuzioni diventano estremamente simili. Lo scopo dela standardizzazione è rendere evidenti queste similitudini.



.

8.10 Prevalenza mancinismo 1 (domanda in formato esame)

Il 10% è delle persone sono mancine. Ci chiediamo se la caratteristica sia ereditaria. Eseguiamo il seguente esperimento. Selezioniamo un campione di 1000 persone con almeno un genitore mancino e misuriamo la frequenza di mancini. Concludiamo che la caratteristica è ereditaria se più di 115 individui sono mancini.

Domande.

1. Qual è l'potesi nulla?
2. Qual è l'potesi alternativa?
3. Che test possiamo fare?
4. Qual è la significatività del test ? (Usare le funzioni di R in calce).
5. Supponiamo che un leggero fattore ereditario renda la prevalenza del mancinismo tra i figli di un genitore mancino $\geq 12\%$. Stimare la probabilità che questa dipendenza ereditaria non venga rilevata dal test.
6. Qual è la potenza del test nel caso descritto sopra.
7. Supponiamo di volere un test con la significatività del 1% come dobbiamo sceglere la zona di rifiuto?

Risposte. Definiamo: $n = 1000$, $x = 115$, $p_0 = 0.10$, $p_1 = 0.12$, $\alpha = 0.01$

1. H_0: $p = p_0$ dove p è la prevalenza del mancinismo tra i figli di mancini.
2. H_A: $p > p_0$
3. Test binomiale a una coda.
4. $\Pr(X > x)$ dove $X \sim B(n, p)$ ovvero `1-pbinom(`x`,`n`,`p_0`)` `(0.053)`
5. $\beta = \Pr(X \leq x)$ dove $X \sim B(n, p_1)$ ovvero `pbinom(`x`,`n`,`p_1`)` `(0.033)`
6. La potenza è $1 - \beta$ `(0.67)`
7. Rifiutiamo H_0 se più di `qbinom(`$1-\alpha$`,`n`,`p_0`)` mancini `(123)`

N.B per 3-7, altre risposte corrette sono possibili (se coerenti).

Si assumano noti i valori delle seguenti funzioni

`pbinom(x,n,p)`$=P(X \leq x)$, per $X \sim B(n,p)$ `qbinom(`α`,n,p)=x`, dove $P(X \leq x) = \alpha$ per $X \sim B(n,p)$

`pnorm(z)`$=P(Z \leq z)$, per $Z \sim N(0,1)$ `qnorm(`α`)=z`, dove $P(Z \leq z) = \alpha$ per $Z \sim N(0,1)$

`pt(t,`ν`)` $= P(T \leq t)$ per $T \sim t(\nu)$ `qt(`α`,`ν`)=t`, dove $P(T \leq t) = \alpha$ per $T \sim t(\nu)$

`pchisq(q,k)` $= P(Q \leq q)$, per $Q \sim \chi^2_k$ `qchisq(`α`,k)=q`, dove $P(Q \leq q) = \alpha$ per $Q \sim \chi^2_k$

.

8.11 Prevalenza mancinismo 2 (domanda in formato esame)

Il 10% è delle persone sono mancine. Ci chiediamo se la caratteristica sia ereditaria. Eseguiamo il seguente esperimento. Selezioniamo un campione di 1000 persone con almeno un genitore mancino e misuriamo la frequenza di mancini. Otteniamo 112 mancini.

Domande

1. Qual è l'potesi nulla?
2. Qual è l'potesi alternativa?
3. Che test possiamo fare?
4. Qual è il p-valore ottenuto dai dati?
5. Possiamo rigettare l'ipotesi nulla con una significatività del $\alpha = 5\%$?

Risposte Definiamo: $n = 1000$, $x = 112$, $p_0 = 0.10$

1. $p = p_0$ dove p è la prevalenza di mancini tra i figli di genitori mancini
2. $p > p_0$
3. Test binomiale a una coda
4. il p-valore è `1 - pbinom(`x`,`n`,`p_0`)` (`0.095`)
5. no, perché $\alpha <$ p-valore.

N.B per 3-5, altre risposte corrette sono possibili (se coerenti).

Si assumano noti i valori delle seguenti funzioni

`pbinom(x,n,p)`$=P(X \leq x)$, per $X \sim B(n,p)$ `qbinom(`α`,n,p)=x`, dove $P(X \leq x) = \alpha$ per $X \sim B(n,p)$

`pnorm(z)`$=P(Z \leq z)$, per $Z \sim N(0,1)$ `qnorm(`α`)=z`, dove $P(Z \leq z) = \alpha$ per $Z \sim N(0,1)$

`pt(t,`ν`)` $= P(T \leq t)$ per $T \sim t(\nu)$ `qt(`α`,`ν`)=t`, dove $P(T \leq t) = \alpha$ per $T \sim t(\nu)$

`pchisq(q,k)` $= P(Q \leq q)$, per $Q \sim \chi^2_k$ `qchisq(`α`,k)=q`, dove $P(Q \leq q) = \alpha$ per $Q \sim \chi^2_k$

.

9 Z-test

La distribuzione normale ↱

9.1 Test a una coda

Si sospetta che una certa terapia faccia aumentare la pressione diastolica. Nella popolazione generale la pressione diastolica ha distribuzione $N(\mu_0, \sigma^2)$ con $\mu_0 = 75$ e $\sigma = 9.5$.

Assumiamo che tra i pazienti in terapia la pressione diastolica sia distribuita normalmente con media ignota μ e con la stessa deviazione standard della popolazione generale. Vogliamo testare le seguenti ipotesi:

$H_0 : \quad \mu = \mu_0$

$H_A : \quad \mu > \mu_0$

Il test consiste nel misurare la pressione ad un campione di n pazienti e di questi dati calcolare la media. Abbiamo quindi la seguente statistica

$$\bar{X} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} X_i$$

Dove X_i è la v.a. che dà la pressione dell'i-esimo paziente del campione. Rigetteremo H_0 se il valore ottenuto è suporiore ad un certo x_α che vogliamo fissare in modo che l'errore I tipo risulti uguale ad α. Quindi x_α dev'essere tale che x_α tale che $\Pr(\bar{X} > x_\alpha) = \alpha$.

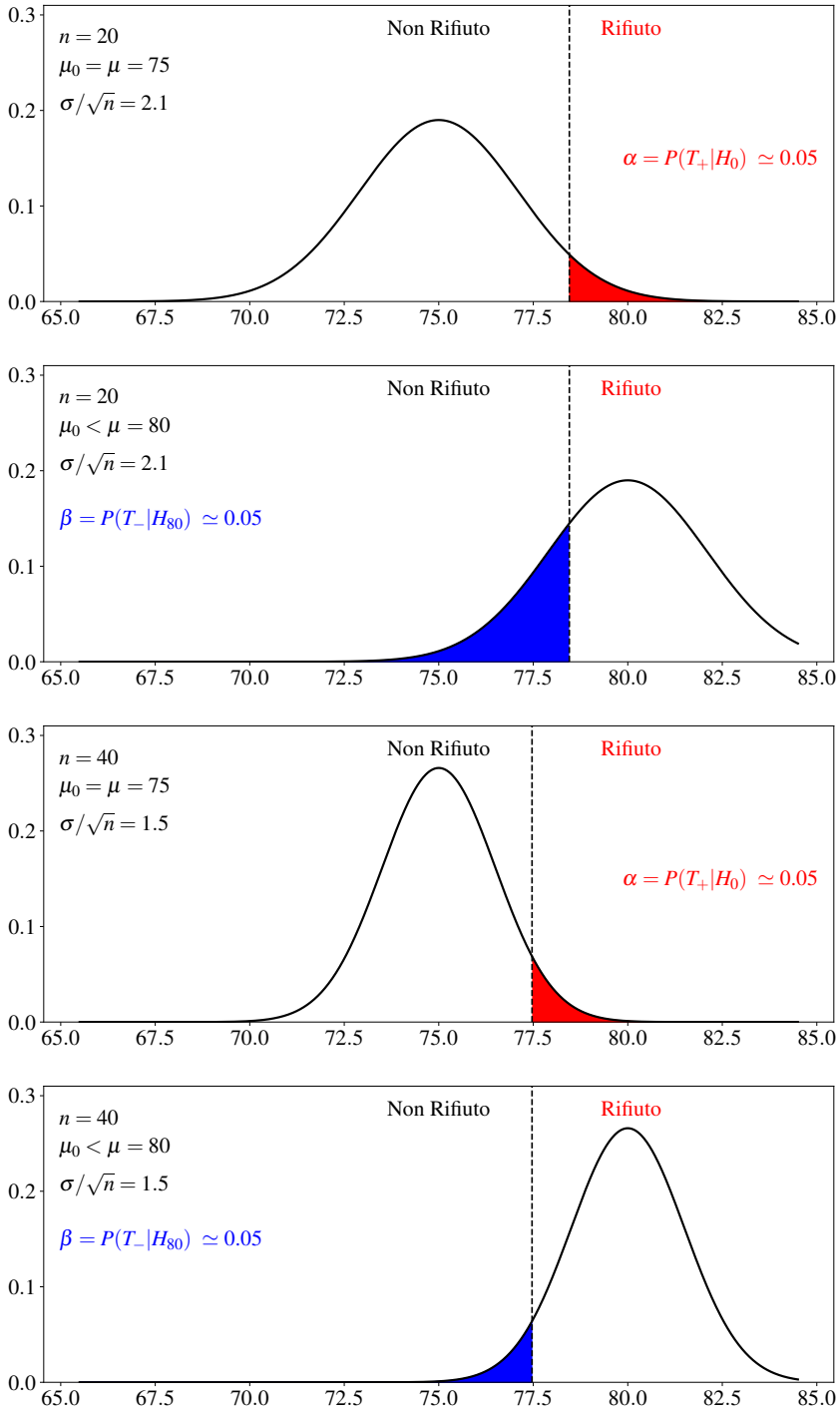
Se H_0 è vera, $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$.

Se H_A è vera, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ per qualche $\mu > \mu_0$.

.

9.2 Una coda, errore I e II tipo Z-test

Qui rappresentiamo gli errori del I e II tipo per campioni di dimensione $n = 20$ e $n = 40$ e con un x_α scelto in moto tale da avere $\alpha = 5\%$. Per gli errori del II tipo prendiamo $\delta = 5$.



9.3 Esempio di domanda in formato esame

(Ripetuto da sopra.) Si sospetta che una certa terapia faccia aumentare la pressione diastolica. Nella popolazione generale la pressione diastolica ha distribuzione $N(\mu_0, \sigma^2)$ con $\mu_0 = 75$ e $\sigma = 5.5$. Assumiamo che tra i pazienti in terapia la pressione diastolica sia distribuita normalmente con media ignota μ e con la stessa deviazione standard della popolazione generale.

Un esperimento consiste nel misurare il valor medio $\bar{x}$ della pressione diastolica di un campione di $n = 64$ pazienti in terapia.

Domande

1. Qual è l'ipotesi nulla?
2. Qual è l'ipotesi alternativa?
3. Quale test dobbiamo usare?
4. Qual è il valore di soglia x_α per cui possiamo rifiutare l'ipotesi nulla con un livello di confidenza del 95% (ovvero significatività del 5%) ?
5. Qual'è la potenza del test per un effect size $\delta = 2$?

Risposte

1. H_0: $\mu = \mu_0$
2. H_A: $\mu > \mu_0$
3. Z-test a una coda (coda superiore)
4. x_α è tale che $0.05 = \Pr(\bar{X} \geq x_\alpha \mid H_0)$ dove $\bar{X} \sim N(\mu_0, \sigma^2/n)$.
   Quindi x_α = `qnorm(0.95, `μ_0`, `$\sigma/\sqrt{n}$`)` `(76.95)`
5. La potenza del test è $1 - \Pr(\bar{X} \leq x_\alpha \mid H_A)$ dove $\bar{X} \sim N(\mu_0 + \delta, \sigma^2/n)$.
   Quindi la potenza è `1-pnorm(`x_α`, `$\mu_0 + \delta$`, `$\sigma/\sqrt{n}$`)` `(0.90)`

Si assumano noti i valori delle seguenti funzioni

`pbinom(x,n,p)`=$P(X \leq x)$, per $X \sim B(n,p)$ — `qbinom(`α`,n,p)`=x, dove $P(X \leq x) = \alpha$ per $X \sim B(n,p)$

`pnorm(x,`μ`,`σ`)`=$P(X \leq x)$, per $X \sim N(\mu, \sigma^2)$ — `qnorm(`α`,`μ`,`σ`)`=x, dove $P(X \leq x) = \alpha$ per $X \sim N(\mu, \sigma^2)$

`pt(t,`ν`)` = $P(T \leq t)$ per $T \sim t(\nu)$ — `qt(`α`,`ν`)`=t, dove $P(T \leq t) = \alpha$ per $T \sim t(\nu)$

`pchisq(q,k)` = $P(Q \leq q)$, per $Q \sim \chi^2_k$ — `qchisq(`α`,k)`=q, dove $P(Q \leq q) = \alpha$ per $Q \sim \chi^2_k$

.

9.4 Prevalenza mancinismo 3 (domanda in formato esame)

Il teorema del limite centrale ↱

Il 10% è delle persone sono mancine. Ci chiediamo se la caratteristica sia ereditaria. Eseguiamo il seguente esperimento. Selezioniamo un campione di 1000 persone con almeno un genitore mancino e misuriamo la frequenza di mancini.

Domande

1. Qual è l'potesi nulla?
2. Qual è l'potesi alternativa?
3. Che test possiamo fare?
4. Quale valore di soglia otteniamo se vogliamo rifiutare H_0 con una significatività del 5% ?

Risposte Definiamo: $p_0 = 10\%$
$p =$ prevalenza del mancinismo tra i figli di genitori mancini
$\sigma_0 = \sqrt{n\, p_0(1-p_0)}$
$n = 1000$
$x =$ frequenza relativa rilevata
$z = \dfrac{x - p_0}{\sigma_0/\sqrt{n}} = \dfrac{x - p_0}{\sqrt{p_0(1-p_0)}}$

1. H_0: $p = p_0$
2. H_A: $p > p_0$
3. Z-test a una coda (superiore). Approssimiamo binomiale con normale.
4. La soglia per z è z_α tale che $\alpha = \Pr(Z > z_\alpha)$. Quindi z_α = `qnorm(1-`α`)`
   La soglia per la frequenza è $x_\alpha = z_\alpha\sqrt{p_0(1-p_0)} + p_0$.

N.B. avremmo potuto usare il test binomiale.

Si assumano noti i valori delle seguenti funzioni

`pbinom(x,n,p)`$=P(X \le x)$, per $X \sim B(n,p)$ — `qbinom(`α`,n,p)`$=$`x`, dove $P(X \le x) = \alpha$ per $X \sim B(n,p)$

`pnorm(z)`$=P(Z \le z)$, per $Z \sim N(0,1)$ — `qnorm(`α`)`$=$`z`, dove $P(Z \le z) = \alpha$ per $Z \sim N(0,1)$

`pt(t,`ν`)` $= P(T \le t)$ per $T \sim t(\nu)$ — `qt(`α`,`ν`)`$=$`t`, dove $P(T \le t) = \alpha$ per $T \sim t(\nu)$

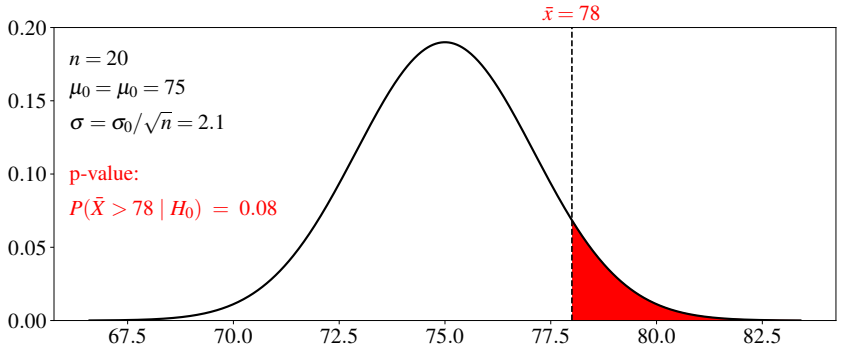
`pchisq(q,k)` $= P(Q \le q)$, per $Q \sim \chi^2_k$ — `qchisq(`α`,k)`$=$`q`, dove $P(Q \le q) = \alpha$ per $Q \sim \chi^2_k$

.

9.5 Una coda, p-valore. Z-test

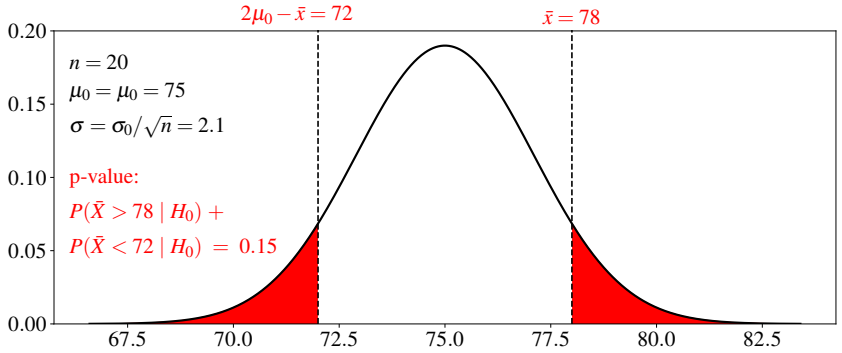
Il p-valore

(Continua l'esempio pressione diastolica.) Supponiamo di ottenere $\bar{x} = 78.0$ da un campione di dimensione $n = 20$. Il p-valore di questa misura è $\Pr(\bar{X} \geq 78) = 1 - \Pr(\bar{X} \leq 78)$.



Numericamente $\Pr(\bar{X} \leq \bar{x})$ si può calcolare usando la funzione `pnorm(x,m,s)` Nel nostro caso i valori sono `x` $= \bar{x}$, `m` $= \mu_0$, e `s` $= \sigma/\sqrt{n}$

Nel caso di un test a due code, se H_A fosse stata $\mu_0 \neq \mu$, il p-valore diventa esattamente il doppio che per il test ad una coda (qui sotto differisce numericamente a causa degli arrotondamenti).



.

9.6 Esempio di domanda in formato esame

(Ripetuto da sopra.) Si sospetta che una certa terapia faccia aumentare la pressione diastolica. Nella popolazione generale la pressione diastolica ha distribuzione $N(\mu_0, \sigma^2)$ con $\mu_0 = 75$ e $\sigma = 5.5$. Assumiamo che tra i pazienti in terapia la pressione diastolica sia distribuita normalmente con media ignota μ e con la stessa deviazione standard della popolazione generale.

(Variazione.) Misuriamo la pressione media di un campione di $n = 64$ pazienti e osserviamo la media campionaria $\bar{x} = 77$.

Domande

1. Qual è l'ipotesi nulla ?
2. Qual è l'ipotesi alternativa ?
3. Quale test dobbiamo usare ?
3. Qual è il p-valore ottenuto dalla misura $\bar{x}$ misurato ?
4. Possiamo rigettare l'ipotesi nulla con una significatività del 1% ?

Risposte

1. H_0: $\mu = \mu_0$
2. H_A: $\mu > \mu_0$
3. Z-test una coda (coda superiore)
3. p-valore = $\Pr(\bar{X} \geq \bar{x} \mid H_0) = 1 - \Pr(\bar{X} \leq \bar{x} \mid H_0)$ dove $\bar{X} \sim N(\mu_0, \sigma/\sqrt{n})$. Quindi p-valore = `1-pnorm(`$\bar{x}$`,` μ_0`,` $\sigma/\sqrt{n}$`)`
4. Si, se p-valore < 0.01

Si assumano noti i valori delle seguenti funzioni

`pbinom(x,n,p)`$=P(X \leq x)$, per $X \sim B(n,p)$ — `qbinom(`α`,n,p)`$=x$, dove $P(X \leq x) = \alpha$ per $X \sim B(n,p)$

`pnorm(x,`μ,σ`)`$=P(X \leq x)$, per $X \sim N(\mu, \sigma^2)$ — `qnorm(`α,μ,σ`)`$=x$, dove $P(X \leq x) = \alpha$ per $X \sim N(\mu, \sigma^2)$

`pt(t,`ν`)` $= P(T \leq t)$ per $T \sim t(\nu)$ — `qt(`α`,`ν`)`$=t$, dove $P(T \leq t) = \alpha$ per $T \sim t(\nu)$

`pchisq(q,k)` $= P(Q \leq q)$, per $Q \sim \chi^2_k$ — `qchisq(`α`,k)`$=q$, dove $P(Q \leq q) = \alpha$ per $Q \sim \chi^2_k$

.

9.7 Crescita media

In condizioni ottimali l'incremento di una certa cultura in una fissata unità di tempo ha media $\mu_0 = 3.1$ e deviazione standard $\sigma = 1.2$. Vogliamo progettare un test per decidere se la crecita di una data cultura sia sub-ottimale. Assumiamo che la distribuzione sia normale e che in condizioni sub-ottimali la deviazione standard sia la stessa. (Queste assunzioni sono abbastanza irragionevoli, ma portiamo pazienza.)

Domande:

1 Preleviamo $n = 9$ campioni, e misuriamo la crescita in un'unità di tempo. E calcoliamo la media campionaria $\bar{x}$. Quanto dev'essere x_α per poter affermare che con significatività $\alpha = 1\%$ che siamo in condizioni sub-ottimali quando $\bar{x} < x_\alpha$?

2 Dato x_α come sopra. Qual'è la probabilità di un errore del II tipo se l'effect size è $\delta = 0.5$?

Risposte:

1 Vogliamo $1\% = \Pr(\bar{X} \leq x_\alpha)$ con $\bar{X} \sim N(\mu_0, \sigma/\sqrt{n})$.
Quindi $x_\alpha =$ `qnorm(0.01,3.1,0.4)` $=$ `2.17`

2 $\beta = \Pr(\bar{X} \leq x_\alpha)$ con $\bar{X} \sim N(\mu_0 - \delta, \sigma/\sqrt{n})$.
Quindi $\beta =$ `pnorm(2.17,2.6,0.4)` $=$ `0.14`.

.

9.8 Mean weight (domanda in formato esame)

Boys of a certain age are known to have a mean weight of 85 pounds and standard deviation 10.6 pounds. A complaint is made that the boys living in a municipal children's home are overfed. As one bit of evidence, 25 boys (of the same age) are weighed and found to have a mean weight of 88.94 pounds. Assume the same standard deviation as in the general the population (the unrealistic part of this example).

Domande

1. Qual è l'potesi nulla?
2. Qual è l'potesi alternativa?
3. Che test possiamo fare?
4. Qual'è il p-valore ottenuto dai dati?
5. Possiamo rigettare l'ipotesi nulla con una significatività del $\alpha = 5\%$?

Risposte Definiamo:
$\mu_0 = 85$
$\sigma = 10.6$
$n = 25$
$\bar{x} = 88.94$

1. $\mu = \mu_0$.
2. $\mu > \mu_0$
3. Z-test una coda (superiore)
4. il p-valore è `1-pnorm(z)` dove $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$. `(0.03)`
   Lo stesso risultato si ottienre con `1-pnorm(`$\bar{x}, \mu_0, \sigma/\sqrt{n}$`)`, ma non è tra le possibilità elencate in calce.
5. si, perché p-valore $< \alpha$.

Si assumano noti i valori delle seguenti funzioni

`pbinom(x,n,p)`=$P(X \leq x)$, per $X \sim B(n,p)$ — `qbinom(`α`,n,p)`=x, dove $P(X \leq x) = \alpha$ per $X \sim B(n,p)$

`pnorm(z)`=$P(Z \leq z)$, per $Z \sim N(0,1)$ — `qnorm(`α`)`=z, dove $P(Z \leq z) = \alpha$ per $Z \sim N(0,1)$

`pt(t,`ν`)` = $P(T \leq t)$ per $T \sim t(\nu)$ — `qt(`α`,`ν`)`=t, dove $P(T \leq t) = \alpha$ per $T \sim t(\nu)$

`pchisq(q,k)` = $P(Q \leq q)$, per $Q \sim \chi^2_k$ — `qchisq(`α`,k)`=q, dove $P(Q \leq q) = \alpha$ per $Q \sim \chi^2_k$

.

10 T-test

La distribuzione t di Student

10.1 Una popolazione

A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90% confidence that the mean score for the class on the test would be above 70?

$n = 6$

$\mu_0 = 70$

$$\bar{x} = \frac{1}{n}\big(62 + 92 + 75 + 68 + 83 + 95\big) = 79.17$$

$$s = \sqrt{\frac{(62-\bar{x})^2 + (92-\bar{x})^2 + (75-\bar{x})^2 + (68-\bar{x})^2 + (83-\bar{x})^2 + (95-\bar{x})^2}{n-1}}$$

$$= 13.17$$

1. H_0 $\mu_0 = 70$
2. H_A $\mu > \mu_0$
3. Il t-score è $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{79.17 - 70}{13.17/\sqrt{6}} = 1.71$
4. Il p-valore è $\Pr(T \geq t)$ dve $T \sim t(5)$. Ovvero `pt(t, n-1)` `(0.074)`
   In R la funzione il risultato si ottiene direttamente con

   ```
   x = c(62,92,75,68,83,95)
   t.test(x, alternative='greater', mu=70)
   ```

   Quindi possiamo affermare con un livello di confidenza $\geq 90\%$ che il voto medio della classe in un ipotetico esame sarà del > 70.

.

10.2 Due popolazioni

Si sospetta che un certo medicinale modifichi la pressione diastolica. Prendiamo due gruppi di $n_x = 6$ e $n_y = 5$ persone. Al primo gruppo somministriamo il medicinale al secondo un placebo. Assumiamo che in entrambi i casi la pressione diastolica sia distribuita normalmente con la stessa deviazione standard (ignota). Nel primo gruppo otteniamo i valori $x_1, \ldots, x_6 = 62, 92, 75, 68, 83, 95$ nel secondo gruppo $y_1, \ldots, y_5 = 60, 95, 76, 69, 89$.

1. H_0 $\mu_x = \mu_y$ la media nelle due popolazioni è la stessa
2. H_A $\mu_x \neq \mu_y$ la media nelle due popolazioni è diversa
3. T-test per due popolazioni a due code
3. Il t-score è $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_x + s_y^2/n_y}}$ dove

   $\bar{x} = \dfrac{1}{n_x}\sum_{i=1}^{6} x_i$ $\qquad$ $\bar{y} = \dfrac{1}{n_y}\sum_{i=1}^{6} y_i$

   $s_x^2 = \dfrac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2$ $\qquad$ $s_y^2 = \dfrac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \bar{y})^2$
4. Il p-valore è $2\Pr(T \geq |t|)$ dove $T \sim t(n_x + n_y - 2)$.
   Ovvero `2 * (1 - pt(`$|t|$`, `$n_x + n_y - 2$`))`
   In R la funzione il risultato si ottiene direttamente con

```
x = c(62,92,75,68,83,95)
y = c(60,95,76,69,89)
t.test(x, y, alternative='two', var.equal = TRUE)
```

.

10.3 Dati accoppiati

Si sospetta che un certo medicinale modifichi la pressione diastolica. Prendiamo un gruppo di $n = 6$ persone. Somministriamo ad ogni individuo prima un placebo e successivamente il medicinale. Assumiamo che in entrambi i casi la pressione diastolica sia distribuita normalmente. Con il placebo otteniamo i valori $x_1, \ldots, x_6 = 62, 92, 75, 68, 83, 95$ con il medicinale otteniamo i valori $y_1, \ldots, y_6 = 60, 95, 76, 69, 89, 90$.

1. H_0 $\mu = 0$
2. H_A $\mu \neq 0$
3. T-test per due campioni accoppiati a due code
3. Il t-score è $t = \dfrac{\bar{z}}{s/\sqrt{n}}$ dove $z_i = x_i - z_i$ e
   $\bar{z} = \dfrac{1}{n}\sum_{i=1}^{n} z_i$
   $s^2 = \dfrac{1}{n}\sum_{i=1}^{n} z_i^2$
4. Il p-valore è $2\Pr(T \geq |t|)$ dove $T \sim t(n-1)$.
   Ovvero `2 * (1 - pt(`$|t|$`, `$n-1$`))`
   In R la funzione il risultato si ottiene direttamente con

```
x = c(62,92,75,68,83,95)
y = c(60,95,76,69,89,90)
t.test(x, y, alternative='two', paired=TRUE)
```

11 Esercizi vari

11.1 Placebo Esercizi

Ad un gruppo di persone vengono misurati 50 diversi parametri fisiologici (che assumiamo indipendenti) prima e dopo l'assunzione di un placebo. Per ognuno di questi parametri l'ipotesi nulla è che non ci sia differenza. Qual'è la probabilità che per almeno uno di questi parametri si ottenga p-valore < 0.02 ?

Risposta: $1 - (0.98)^{50} = 64\%$