Un bignami di probabilità e statistica

Branch: main 2023-06-16 14:57:05+02:00

Per l'ultima versione:

https://github.com/domenicozambella/ProbStatDispense/raw/main/PDF/dispense.pdf

Questi appunti contengono errori e refusi, quindi correzioni e commenti sono graditi

mailto:domenico.zambella@unito.it

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 $\omega \in \Omega$ dello spazio campionario come i possibili oggetti di un rilevamento, un esperimento, un sorteggio, ecc. [*outcomes of a trial or of an experiment or individuals of a pupulation*]. I sottoinsiemi $E \subseteq \Omega$ verranno chiamati eventi.

In molti casi solo particolari sottoinsiemi di Ω vengono considerati legittimi eventi. Questo però è un dettaglio tecnico che ignoreremo.

Una misura di probabilità è una funzione $\Pr : \mathcal{P}(\Omega) \to \mathbb{R}$ che soddisfa i seguenti assiomi

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

Conseguenze:

1. $\Pr(\varnothing) = 0$
2. $\Pr(\neg E) = 1 - \Pr(E)$
3. $\Pr(E_1 \setminus E_2) = \Pr(E_1) - \Pr(E_2)$ se $E_2 \subseteq E_1$
4. $\Pr(E_1 \cup E_2 \cup E_3) = \Pr(E_1) + \Pr(E_2) + \Pr(E_3)$ for mutually exclusive E_1, E_2, E_3
5. $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap \mathrm{E}_2)$ per ogni $E_1, E_2 \in \mathcal{P}(\Omega)$.

Chiaramente (4) può essere generalizzata ad un numero qualsiasi di eventi $E_1, \ldots, E_n$ mutualmente esclusivi. Però a volte serve reneralizzare questa proprietà ad un numero qualsiasi di eventi.

3′. $\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i)$ per ogni $E_1, E_2, \ldots \subseteq \Omega$ mutualmente esclusivi.

Per ragioni tecniche, (3′) non è una conseguenza degli assiomi (1)-(3) e dev'essere aggiunto come assioma indipendente.

2 Variabili aleatorie

Esempi: Urna con biglie di dimensioni diverse

Una variabile aleatoria (v.a.) [*random variable (r.v.)*] è una funzione $X : \Omega \to R$, dove Ω è uno spazio campionario ed R un insieme qualsiasi. Spesso si scrive $X \in R$ omettendo il riferimento ad Ω (è una notazione che può dar luogo a malintesi, in realtà si intende $X(\omega) \in R$ per ogni $\omega \in \Omega$).

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 continue

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

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

La funzione $\Pr(X = x)$ si chiama distribuzione di probabilità [*probability mass function (p.m.f.)*]. Spesso indicata con p_x, $f(x)$ o $f_X(x)$. La funzione $\Pr(X \leq x)$ si chiama funzione di ripartizione [*cumulative distribution function (c.d.f.)*]. Spesso indicata con $F(x)$ o $F_X(x)$.

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

$$\Pr\big(X \in A\big) = \sum_{x \in A} \Pr(X{=}x)$$

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

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

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.

Assumeremo sempre l'esistenza di una funzione $f : \mathbb{R} \to \mathbb{R}$ che chiameremo densità di probabilità [*probability density function (p.d.f.)*] tale che

$$\Pr(X \leq t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x$$

Intuitivamente $f(x)$ è la probabilità che $X \in [x, x + \mathrm{d}x]$.

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 o percentile. Questa è la funzione che dato $p \in (0, 1)$ ritorna il valore $x = Q_X(p)$ che è il minimo (o più correttamente il limite inferiore) degli x tale che $p \leq \Pr(X \leq x)\}$.

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

In prima approssimazione $Q_X(p)$ è la funzione inversa della funzione di ripartizione $F_X(x) = \Pr(X \leq x)$. Solo che può succerere che F_X non sia invertibile: più valori di x corrisponda lo stesso valore di $F_X(x)$. In questo caso prendiamo il limite inferiore di questi x.

5 Probabilità condizionata

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

$$\Pr(A \mid \Phi) \;=\; \frac{\Pr(A \cap \Phi)}{\Pr(\Phi)}$$

Questo si legge probabilità di A dato Φ. Si verifica facilmente che $\Pr(\,\cdot\,|\Phi)$ soddisfa a tutte le proprietà di $\Pr(\,\cdot\,)$ se rimpiazziamo Ω con Φ ed ogni sottoinsieme A di Ω con $A \cap \Phi$.

6 Teorema delle Probabilità Totali

Esempi: Popolazione maschile e femminile
Urna con due tipi di dadi
Modello di Hardy-Weinberg

Siano $\Phi_1, \ldots, \Phi_n$ eventi di probabilità $\neq 0$ mutuamente esclusivi ed esaustivi (la loro unione è tutto Ω). Il teorema seguente si chiama Teorema delle Probabilità Totali.

Theorem 1. Sia C è un qualsiasi altro evento, allora

$$\Pr(C) = \sum_{i=1}^{n} \Pr(\Phi_i) \cdot \Pr\left(C \mid \Phi_i\right).$$

Proof.
Verifichiamo il teorema per $n = 2$.

$$\Pr(C) = \Pr(\Phi_1) \cdot \Pr\left(C \mid \Phi_1\right) + \Pr(\Phi_2) \cdot \Pr\left(C \mid \Phi_2\right)$$

Applichiamo la definizione di probabilità condizionata.

$$\begin{aligned} \Pr(C) &= \Pr(\Phi_1) \cdot \frac{\Pr\left(C \cap \Phi_1\right)}{\Pr\left(\Phi_1\right)} + \Pr(\Phi_2) \cdot \frac{\Pr\left(C \cap \Phi_2\right)}{\Pr\left(\Phi_2\right)} \\ &= Pr\left(C \cap \Phi_1\right) + \Pr\left(C \cap \Phi_2\right) \end{aligned}$$

As $C \cap \Phi_1$ and $C \cap \Phi_2$ are disjoint

$$\begin{aligned} \Pr(C) &= \Pr\left((C \cap \Phi_1) \cup (C \cap \Phi_2)\right) \\ &= \Pr\left(C \cap (\Phi_1 \cup \Phi_2)\right) \\ &= \Pr\left(C \cap \Omega\right) \\ &= \Pr(C). \end{aligned}$$

□

7 Indipendenza stocastica

Esempi: Esercizio su v.a. indipendenti

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

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

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

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

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

Nel caso di variabili aleatorie continue la condizione diventa

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

8 Valore atteso

Il valore atteso o media (di popolazione) (in inglese expected value, population mean, più raramente, average) di una variabile aleatoria numerica discreta $X \in R$ è

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

La lettera μ viene usata quando è chiaro a quale variabile ci si riferisce. Per evitare ambiguità a volte si scrive μ_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 !

When X is a continuous r.v. the expected value is computed by an integral

$$\mathrm{E}(X) \;=\; \int_{-\infty}^{+\infty} x\, f(x)\, \mathrm{d}x$$

9 Valore atteso della somma e prodotto di v.a.

Se X is una v.a. numerica e c una costante scriviamo cX per la v.a. che mappa $\omega \mapsto c \cdot X(\omega)$.

Theorem 2. Per ogni v.a. X $\quad \mathrm{E}(cX) = c\,\mathrm{E}(X)$

Se X, Y sono v.a. (numeriche) scriviamo e $X+Y$ per la v.a. che mappa $\omega \mapsto X(\omega)+Y(\omega)$. La v.a. $X \cdot Y$ è definita in modo simile.

Theorem 3. Per ogni v.a. X, Y $\quad \mathrm{E}(X+Y) = \mathrm{E}(X)+\mathrm{E}(Y)$

Le due proprietà qui sopra congiunte si chiamano linearità, ovvero i teoremi 2 e 3 dicono che il valore atteso è un operatore lineare.

Theorem 4. Se X, Y sono v.a. *indipendenti* allora $\quad \mathrm{E}(X \cdot Y) = \mathrm{E}(X) \cdot \mathrm{E}(Y)$

10 Variabili aleatorie di Bernoulli

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

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 1_E

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

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

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

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

11 Binomial random variables

Esempi: Estrazione ripetuta
Multiple choice quiz

We say that X is a binomial r.v. 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

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

The cumulative distribution function is

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

12 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

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

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

13 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)

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

Il fatto seguente è una conseguenza del Teorema delle Probabilità Totali. Siano $\Phi_1, \ldots, \Phi_n$ eventi mutuamente esclusivi ed esaustivi di probabilità $\neq 0$. Sia X una v.a. discreta

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

La verifica è immediata.

14 Varianza

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

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

Il seguente fatto è utile

Fact 5. Per ogni v.a. X $\qquad \mathrm{Var}(X) \;=\; \mathrm{E}(X^2) - \mathrm{E}(X)^2$

Proof.
Per semplicità verifichiamo l'uguaglianza nel caso di variabili discrete. Per la definizione di valore atteso

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{x \in R} \Big(x - \mathrm{E}(X)\Big)^2 \cdot \Pr(X = x) \\
&= \sum_{x \in R} \Big(x^2 + \mathrm{E}(X)^2 - 2x\mathrm{E}(X)\Big) \cdot \Pr(X = x) \\
&= \Big(\mathrm{E}(X^2) + \mathrm{E}(X)^2 - 2\mathrm{E}(X) \sum_{x \in R} x \Pr(X = x) \\
&= E(X^2) - E(X)^2.
\end{aligned}$$

□

La deviazione standard è la radice della varianza

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

La lettera σ viene usata quando è chiaro a quale variabile ci si riferisce. Per evitare ambiguità a volte si scrive σ_X.

15 Disequazione di Chebyshev

Enunciamo senza dimostrazione un caso particolare di famoso teorema.

Theorem 6. Sia X una v.a. con valore atteso μ e varianza σ^2. Allora per ogni $k \in \mathbb{R}^+$

$$\Pr\Big(|X-\mu| \leq k\sigma\Big) \;\geq\; 1 - \frac{1}{k^2}$$

Vediamo un caso numerico ($k = 2$)

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

Ovvero gran parte della massa di probabilità 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 disuguaglianze ancora più forti. (Per esempio, se X ha distribuzione normale possiamo scrivere 95% invece che 75%.)

16 Varianza della somma di v.a.

Chiemeremo covarianza di X ed Y la quantità

$$\text{coVar}(X, Y) \;=\; \text{E}(XY) - \text{E}(X)\text{E}(Y)$$

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

Theorem 7. $\text{Var}(X + Y) \;=\; \text{Var}(X) + \text{Var}(Y) + 2\,\text{coVar}(X, Y)$

17 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

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

L'uguaglianza è immediata da verificare usando la definizione di varianza e la linearità del valore atteso.

18 Varianza di variabili binomiali

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

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

N.B. Possiamo calcolarla così

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

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

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

19 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).

20 Regola di Bayes

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

Il fatto seguente si chiama Teorema (o regola) di Bayes. È immediato da verificare

Fact 8. 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

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

21 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 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, $\Pr(\neg D|T_-)$

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.

22 Test d'ipotesi

Esempi: Test binomiale (il più elementare test di ipotesi)
Scatola di biglie colorate
Prevalenza mancinismo 1
Prevalenza mancinismo 2

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 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 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: H_0 può essere rifiutata è sinonima di *l'esito del test è positivo*. Noi denotiamo l'evento con T_+.
- L'espressione: H_0 NON può essere rifiutata è sinonima di *l'esito del test è negativo*. Noi denotiamo l'evento con 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 $\Pr(T_-|H_A)$.

23 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 $\Pr(T_+|H_0)$ e $\Pr(T_-|H_A)$. In un caso estremo: un test che a prescindere dai dati rifiuta sempre H_0 non fa errori del I tipo, $\Pr(T_-|H_0) = 0$. Invece un test che non rifiuta mai H_0 non fa errori del II tipo, $\Pr(T_+|H_0) = 0$. La difficoltà nel progettare il test è trovare il giusto equilibrio tra i due errori.

24 Campioni e statistiche

Un campione $\{X_1, \ldots, X_n\}$ è un insieme di v.a. indipendenti e identicamente distribuite (v.a.i.i.d.). 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, la deviazione standard campionaria

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

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

Questi sono stimatori rispettivamente del valore atteso e della deviazione standard (di popolazione).

25 Il p-valore

Esempi: Prevalenza mancinismo 2
Pressione diastolica

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 ipotizza un certo valore w_0 per la statistica. Supponiamo di avere un test a una coda superiore. Il p-valore è la probabilità

$$\text{p-valore} = \Pr\left(W \geq w \mid H_0\right)$$

Se il test è a coda inferiore la disuguaglianza diventa $\leq$.

Se il test è a due code abbiamo

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

26 La distribuzione normale

Esempi: *Z*-test

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

$$\Pr(Z \leq z) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\,\mathrm{d}x$$

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

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

Importante proprietà:

Theorem 9. La somma di v.a.i. con distribuzione $\mathrm{N}(\mu, \sigma^2)$ è ancora una v.a. con distribuzione $\mathrm{N}(\mu, \sigma^2)$.

Una conseguenza importante è che se $X_1, \ldots, X_n$ sono v.a.i. con distribuzione $\mathrm{N}(\mu, \sigma^2)$ allora la v.a. media compionaria

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

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

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

27 The Central Limit Theorem

Let $X_1, \ldots, X_n, \ldots$ be independent and identically distributed r.v. with mean μ and variance σ^2.

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

We know that $\mathrm{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$. Let

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

Moreover, if the sample size n is *sufficiently large* then

$$\lim_{n\to\infty} Z_n \sim \mathrm{N}(0,1)$$

In words, when n is *sufficiently large* Z_n has *approximately* distribution $\mathrm{N}(0,1)$.

28 La distribuzione t di Student

Esempi: T-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.

La rilevanza in statistica di questa distribuzione si deve al seguente fatto. 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

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

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

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

29 Intervallo di confidenza, varianza nota

Esempio:

Consideriamo v.a. $X \sim \mathrm{N}(\mu, \sigma^2)$ con σ nota e μ ignota. Sia $\bar{X}$ la media campionaria. Fissiamo un $\varepsilon > 0$ e calcoliamo la probabilità che $\bar{X}$ sia a distanza inferiore a ε da μ

$$\begin{aligned} \Pr(\mu - \varepsilon < \bar{X} < \mu + \varepsilon) &= \Pr(-\varepsilon < \bar{X} - \mu < \varepsilon) \\ &= \Pr\left(-\frac{\varepsilon}{\sigma/\sqrt{n}} < Z < \frac{\varepsilon}{\sigma/\sqrt{n}} \right) \end{aligned}$$

Chiamiamo questa probabilità $= 1 - \alpha$ livello di confidenza.

Se σ è assunta nota, possiamo calcolare α dato ε. Viceversa possiamo calcolare ε dato α. Notiamo che

$$\begin{aligned} \Pr(\mu - \varepsilon < \bar{X} < \mu + \varepsilon) &= \Pr(-\bar{X} - \varepsilon < -\mu < -\bar{X} + \varepsilon) \\ &= \Pr(\bar{X} - \varepsilon < \mu < \bar{X} + \varepsilon) \end{aligned}$$

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 è la sequente: se ripetiamo l'esperimento, con probabilità $1 - \alpha$ ritroveremo un intervallo che contiene μ.

N.B. L'intervallo è alearorio, μ è un numero fissato, anche se ignoto.

30 Intervallo di confidenza, varianza ignota

Esempio:

Consideriamo v.a. $X \sim \mathrm{N}(\mu, \sigma^2)$ con σ e μ ignote. Sia $\bar{X}$ la media campionaria ed S lo stimatore della deviazione standard. Fissiamo un $\varepsilon > 0$ e calcoliamo la probabilità che $\bar{X}$ sia a distanza inferiore a ε da μ

$$\begin{aligned} \Pr(\mu - \varepsilon < \bar{X} < \mu + \varepsilon) &= \Pr(-\varepsilon < \bar{X} - \mu < \varepsilon) \\ &= \Pr\left(-\frac{\varepsilon}{S/\sqrt{n}} < T < \frac{\varepsilon}{S/\sqrt{n}} \right) \end{aligned}$$

Sia s è il valore osservato di S e assumiamo che questa probabilità venga approssimato da

$$\simeq \Pr\left(-\frac{\varepsilon}{s/\sqrt{n}} < T < \frac{\varepsilon}{s/\sqrt{n}} \right).$$

Chiamiamo questa probabilità $= 1 - \alpha$ livello di confidenza.

Ora possiamo calcolare α dato ε. E viceversa possiamo calcolare ε dato α.

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 $1 - \alpha$.

31 Regressione lineare semplice

In un esperimento misuriamo coppie di valori (x_i, y_i) per $i = 1, \ldots, n$.

Per esempio: (tempo, spazio), (temperatura, volume), ecc.

In teoria questi valori dovrebbero essere correlati linearmente:

$$y_i = \beta_0 + \beta_1 x_i$$

Però la misura delle y è soggetta ad errori e la vera relazione tra le misure è

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

dove e_i sono degli errori (ignoti).

Le x_i si chiamano predittori.

Le y_i si chiamano risposte.

Non esiste un metodo ovvio per inferire i parametri β_0 e β_1 dai dati (x_i, y_i). La scelta comune è di cercare quei parametri β_0 e β_1 che minimizzano le grandezze e_i. Però generalmente non è possibile minimizzare contemporaneamente tutti gli errori. Dobbiamo quindi aggregare gli errori in un unico numero che rappresenti l'errore totale. La funzione che aggrega gli errori si chiama cost function. Ci sono ininite possibililà. Per esempio, due ragionevoli cost function sono

1. $$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

2. $$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Generalmente si opta per seconda scelta perché è un'espressione più semplice da trattare in maniera analitica. Quando i calcoli andavano fatti a mano questa era una scelta obbligata e l'abitudine si è consolidata negli anni. Inoltre, l'analisi statistica che faremo più sotto è notevolmente semplificata nel secondo caso.

Terminologia: spesso si scrive RSS (residual sum of squares) per il numero

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

32 Regressione lineare semplice (2)

Calcoliamo le derivate del RSS rispetto a β_0 e β_1

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^{n} x_i y_i - x_i \beta_0 - \beta_1 x_i^2$$

Queste derivate si annullano per quei β_0 e β_1 che risolvono il sistema

$$\begin{cases} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

dividendo entrambe le equazioni per n ed usando le seguenti abbreviazioni

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \qquad \overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

riscriviamo il sistema in forma abbreviata

$$\begin{cases} \beta_0 + \beta_1 \bar{x} = \bar{y} \\ \beta_0 \bar{x} + \beta_1 \overline{x^2} = \overline{xy} \end{cases}$$

Le soluzioni sono facili da calcolare

$$\begin{cases} \beta_0 = \bar{y} - \beta_1 \bar{x} \\ \beta_1 = \dfrac{\overline{xy} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} \end{cases}$$

33 Regressione lineare semplice (3) inferenza statistica

Nell'a.a.2022/23 questo paragrafo non è programma d'esame.

Immaginiamo che gli errori e_i siano il risultato di variabili aleatorie E_i. Per semplificare, i predittori li immaginiamo come deterministici (le conclusioni valgono più in generale). Le risposte saranno comunque delle variabili aleatorie

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

Le formule ricavate nel paragrafo precedente le immaginiamo come degli stimatori dei parametri β_0, β_1

$$\begin{cases} B_0 = \bar{Y} - B_1 \bar{x} \\ B_1 = \dfrac{\bar{Y} \cdot \bar{x} - \overline{xY}}{\overline{x^2} - \bar{x}^2} \end{cases}$$

Vorremmo sapere la distribuzione di B_0, B_1 per poter calcolare significatività della stima di β_0, β_1. Dobbiamo fare altre ipotesi semplificatrici. Assumiamo che gli errori E_i siano v.a.i.i.d. di tipo $\mathrm{N}(0, \sigma^2)$. Si verifica che (approssimativamente) per $i = 0, 1$

$$\frac{B_i - \beta_i}{S_i/\sqrt{n}} \sim \mathrm{T}(n-2)$$

dove

$$S_0^2 = \frac{n}{n-2} \sum_{i=1}^{n} E_i^2$$

$$S_1^2 = \frac{S_0^2}{n(\overline{x^2} - \bar{x}^2)}$$

Chiamiamo b_0 e b_1, s_0, s_1 valori osservati di queste v.a. (quindi calcolati dai dati (x_i, y_i) con le formule al paragrafo precedente).

Per avere raggio di un intervallo di confidenza $1 - \alpha$ per β_i possiamo usare la formula ricavata al paragrafo 30

$$1 - \alpha = \Pr\left(-\frac{\varepsilon}{s_i/\sqrt{n}} < T < \frac{\varepsilon}{s_i/\sqrt{n}} \right) \qquad \text{per } i = 0, 1.$$

dove $T \sim T(n-2)$.

Un test di ipotesi frequentemente fatto in questo contesto è $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$. Si noti $\beta_1 = 0$ essenzialmente dice che le risposte non dipendono dai predittori. Il p-valore si questo test si calcola $\Pr(|B_1| \geq |b_1|) = \Pr(|T| \geq |b_1\sqrt{n}/s_1)$ dove $T \sim T(n-2)$.

34 Regressione multipla

Molto spesso le risposte dipendono da più di un predittore. Con due predittori il modello è quindi

1. $$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i$$

Più in generale potermmo avere k predittori diversi ma per chiarezza discutiamo il caso $k = 2$. I parametri $\beta_0, \beta_1, \beta_2$ si stimano minimizzando la cost function

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2$$

in modo del tutto analogo al caso $k = 1$.

Questa soluzione permette anche di trattare il caso di dipendenza non lineare. Supponiamo di voler modellare una dipendenza quadratica (per esempio, il predittore è il tempo e la risposta lo spazio di un moto rettilineo con accelerazione costante). Possiamo continuare ad usare il modello (1) con $z_i = (x_i)^2$.

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

$\Pr(\{a\}) = \Pr(\{c\}) = \Pr(\{g\}) = \Pr(\{t\}) = 1/4$

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. L'esperimento consiste nell'estrarre una biglia ed osservarne il colore.

Vediamo due scelte naturali per modellare questo esperimento. Posto $\Omega = \{r, b, n\}$ stipuliamo che

$$\begin{aligned} \Pr(r) &= 1/2 \\ \Pr(b) &= 3/8 \\ \Pr(n) &= 1/8 \end{aligned}$$

In alternativa definiamo $\Omega = \{1, \ldots, 16\}$ e stipuliamo che

$$\begin{aligned} \Pr(i) &= 1/16 \qquad\qquad \text{per ogni } i = 1, \ldots, 16. \\ R &= \{1, \ldots, 8\} \\ B &= \{9, \ldots, 14\} \\ N &= \{15, 16\} \end{aligned}$$

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 da 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 Urna con due tipi di dadi

Un urna contiene dadi a 6 facce e dadi a 4 facce nelle proporzioni di 1/3 e 2/3. Tutti i dadi sono equilibrati. Estraiamo un dado dall'urna, lo lanciamo, ed osserviamo il risultato. Qual'è la probabilità di ottenere un numero tra 1 e 3?

Soluzione.
Sia C l'evento estrarre il dado a 6 faccie (cubico). Sia T l'evento estrarre il dado a 4 faccie (teraedrico). Sappiamo che

$$\Pr\big(C\big) = 1/3$$

$$\Pr\big(T\big) = 2/3$$

Sia E_i, per $1 \leq i \leq 6$, l'evento il risultato del lancio (qualsiasi sia il dado estratto) è i.

Sappiamo che

$$\Pr\big(E_i|T\big) = \begin{cases} 1/4 & \text{per } i = 1, \ldots, 4 \\ 0 & \text{per } i = 5, 6 \end{cases}$$

$$\Pr\big(E_i|C\}\big) = 1/6 \text{ per } i = 1, \ldots, 6.$$

La consegna è calcolare

$$\Pr\left(\bigcup_{i=1}^{3} E_i\right).$$

Applichiamo il Teorema delle Probabilità Totali

$$\Pr\left(\bigcup_{i=1}^{3} E_i\right) \quad = \quad \Pr\left(\bigcup_{i=1}^{3} E_i \;\middle|\; C\right) \cdot \Pr\big(C\big) \;+\; \Pr\left(\bigcup_{i=1}^{3} E_i \;\middle|\; T\right) \cdot \Pr\big(T\big).$$

Tenendo conto che gli eventi E_i sono mutualmente esclusivi, otteniamo

$$= \quad 3 \cdot \frac{1}{6} \cdot \frac{1}{3} \;+\; 3 \cdot \frac{1}{4} \cdot \frac{2}{3}.$$

$$= \quad \frac{2}{3}.$$

□

Si noti che la forma esplicita di Ω è irrilevante. Volendo per una volta essere pedanti potremmo immaginare $\Omega = \{c, t\} \times \{1, \ldots, 6\}$ dove il primo elemento della coppia denota dado estratto, il secndo il risultato del lancio. Quindi $C = \{c\} \times \{1, \ldots, 6\}$, $T = \{t\} \times \{1, \ldots, 6\}$, ed $E_i = \{c, t\} \times \{i\}$.

3.4 Modello di Hardy-Weinberg (1)

In un locus possono occorrere due alleli a, b.

- ▷ specie dipoide
- ▷ accoppiamento casuale (random mating) nessuna selezione
- ▷ 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}

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

3.5 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)

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

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

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

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
$\Pr(F\|A) = 0.2$	frazione di fumatori tra gli affetti
$\Pr(A\|F) = \frac{\Pr(F\|A) \cdot \Pr(A)}{\Pr(F)} = \frac{0.2 \cdot 0.02}{0.1} = 0.04$	prevalenza tra i fumatori

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

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

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

4.4 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 that the test was taken in South Africa where the prevalence is estimated to be around 20%

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

Indipendenza stocastica (richiede nozione di **v.a. di Bernoulli**)

5 Indipendenza

Siano X_1, X_2, X_3 v.a.i. di Bernoulli. Dire quali delle seguenti coppie X, Y sono v.a. indipendenti.

1. $X = X_1 + X_2$ $\quad$ $Y = X_2 + X_3$
2. $X = |X_1 - X_2|$ $\quad$ $Y = |X_2 - X_3|$
3. $X = |X_1 - X_2|$ $\quad$ $Y = X_2$
4. $X = X_1 X_2$ $\quad$ $Y = |X_2 - X_3|$
5. $X = X_1 X_2$ $\quad$ $Y = X_2 X_3$
6. $X = |X_1 - X_2|$ $\quad$ $Y = X_1 + X_3$
7. $X = |X_1 - X_2|$ $\quad$ $Y = X_1 + X_2 + X_3$
8. $X = (X_1 - X_2)^2$ $\quad$ $Y = X_2 + X_3$
9. $X = X_1 - X_2$ $\quad$ $Y = (X_2 + X_3)^2$

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?

Soluzione.
Chiamiamo X la v.a. che conta il numero di biglie rosse estratte. Allora $X \sim B(6, p)$ con $p = 21/36 = 7/12$.

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

□

6.2 Multiple choice quiz

Suppose that you take a 10-question multiple-choice quiz by randomly guessing. Each question has 4 possible answers and only 1 is correct. What is the probability that answering at random you correctly guess at least 4 answers?

Soluzione.
Let $X \sim \mathrm{B}(10, 1/4)$.

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - \sum_{k=0}^{3} \binom{10}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{10-k} \\ &= 1 - \left(\frac{3}{4}\right)^{10} + \left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^{9} + 45\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right)^{8} + 120\left(\frac{1}{4}\right)^{3}\left(\frac{3}{4}\right)^{7} \end{aligned}$$

□

7 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.

7.1 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 $p = 3/4$. Questi dati vengono riassunti scrivendo

$H_0 : \quad p = 1/2$

$H_A : \quad p = 3/4$

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 $\geq n/2 + k$ 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 $\{0, \ldots, n\}$. 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. Non è possibile conoscere la distribuzione di X non conoscendo la proporzione di monete difettose $\Pr(H_A) = 1 - \Pr(H_0)$. Conosciamo solo le distribuzioni $\Pr(X{=}i \mid H_0)$ e $\Pr(X{=}i \mid H_A)$. Queste sono rispettivamente $\mathrm{B}(n, 1/2)$ e $\mathrm{B}(n, 3/4)$. Conoscendo $\Pr(H_A)$ potremmo ricavarci $\Pr(X = i)$ col teorema delle probabilità totali

$$\Pr(X = i) \quad = \quad \Pr(H_A)\Pr(X{=}i \mid H_A) \; + \; \Pr(H_0)\Pr(X{=}i \mid H_0)$$

Per il momento lavoriamo senza assumere $\Pr(H_A)$ nota.

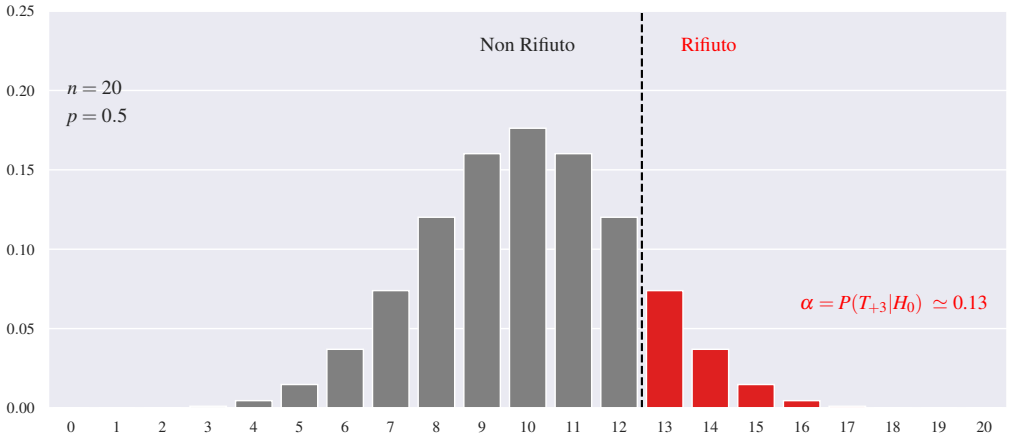
7.2 Test a una coda, errore I tipo

Indichiamo con T_k l'evento $X(\omega) \geq n/2 + k$, ovvero il risultato del test positivo. N.B. stiamo immaginando (per semplicità) che sia n fissato. Diciamo $n = 20$. Quindi l'evento *test positivo* dipende solo da k. Per il momento k lo lasciamo inteterminato per poterne scegliere quello ottimale.

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

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

Per concretezza fissiamo anche $k = 3$ quindi $\boldsymbol{T_{+k} = \{13, \ldots, 20\}}$ è la regione di rifiuto.

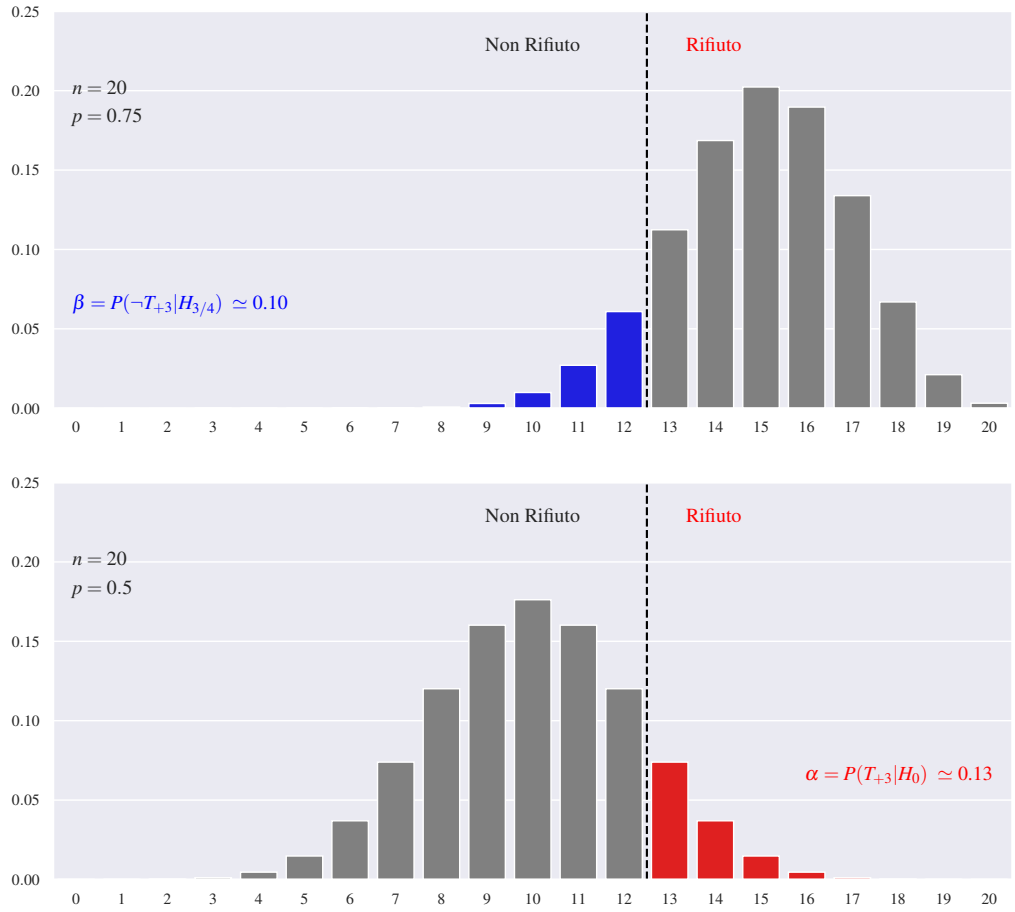


7.3 Test a una coda, errore II tipo

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

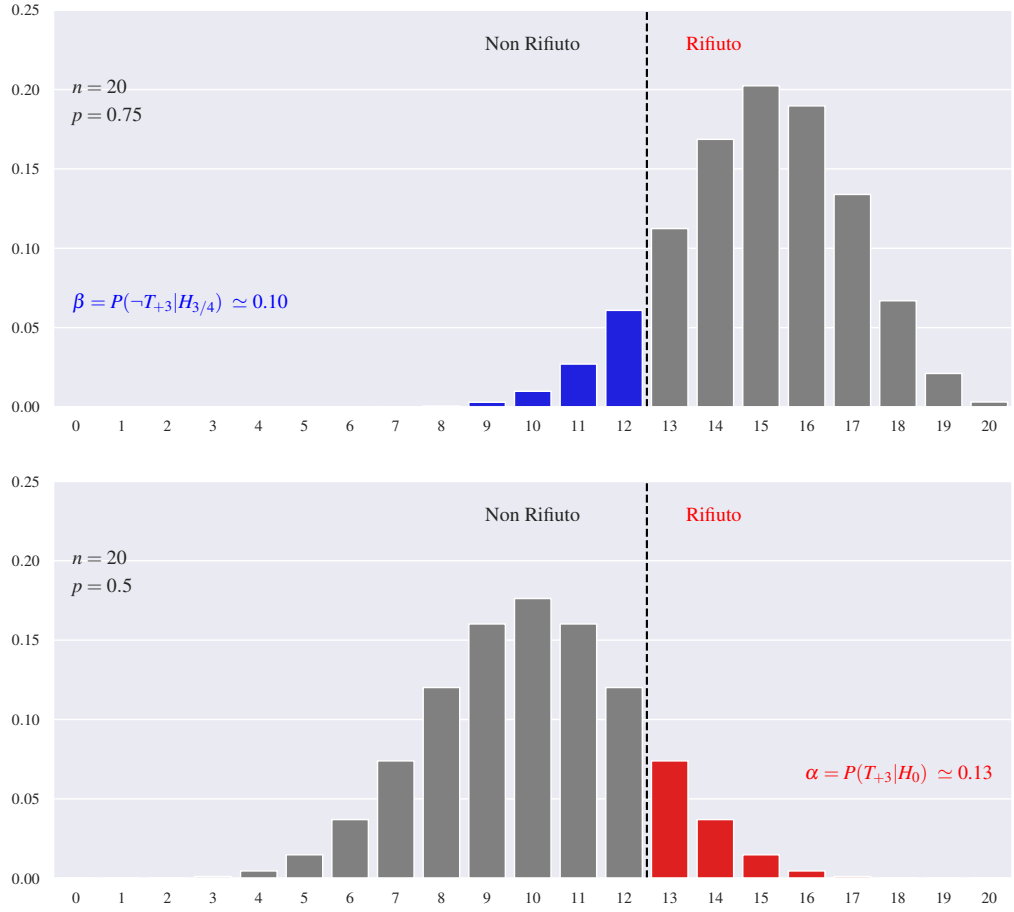
$$\Pr(\neg T_{+k} \mid H_A) \quad = \quad \Pr(X < k \mid H_A) \quad = \quad \sum_{i=0}^{n/2+k-1} \binom{n}{i} p^i (1-p)^{n-i}$$

Consideriamo come prima $k = 3$. Abbiamo $\neg T_{+k} = \{0, \ldots, 12\}$ è la zona di NON rifiuto. Rappresentiamo la distribuzione di X nel caso in cui vale H_A. Per confronto lo accostiamo al grafico del paragrafo precedente.



7.4 Effect size: δ

Cosa possiamo dire se l'ipotesi alternativa fosse stata $H_A : p > 1/2$?



Al crescere di p la distribuzione si sposta verso destra quindi la probabilità di errori del II tipo diminuisce. Di converso, se p si sposta verso sinistra e si avvicina a $1/2$ la probabilità d'errore del II tipo aumenta. Al limite quando $p \approx 1/2$ avremo $\alpha + \beta \approx 1$. Dobbiamo quindi fissare il minima differenza δ che riteniamo significativa e prendere come ipotesi alternativa $H_A : p \geq 1/2 + \delta$. Per calcolare β abbiamo comunque bisogno di un preciso valore di p. Scegliamo quindi il più sfavorevole $H_A : p = 1/2 + \delta$.

7.5 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 : $\quad p = 1/2$

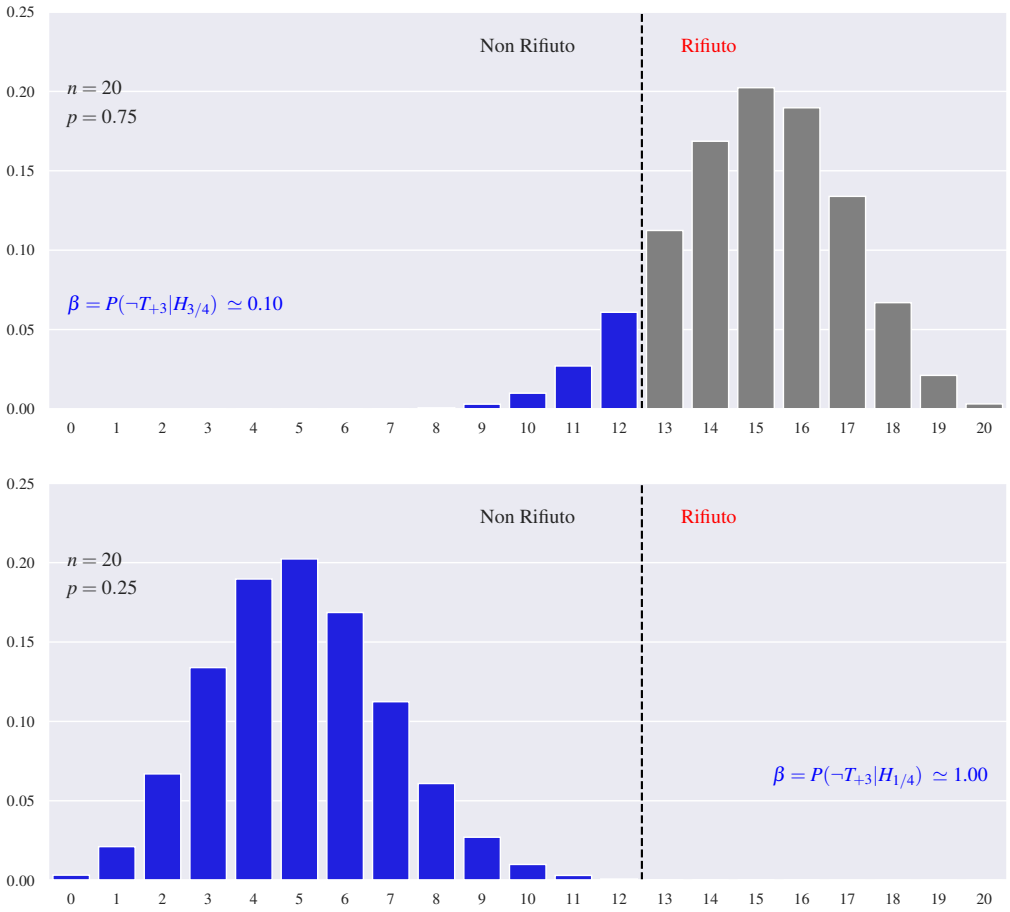
H_A : $\quad p \neq 1/2$

Verifichiamo prima che **la zona di rifiuto dei paragrafi precedenti NON è adatta** alla nuova situazione. Il grafici che descrive $\Pr(T_{+3}|H_0)$ rimane invariato (perché l'insieme H_0 non è cambiato). Le cose cambiano drammaticamente se vale H_A.

Per semplificare la discussione dell'errore del II tipo fissiamo $\delta = 1/4$ e ci mettiamo nei casi più sfavorevoli (sono due ai due lati di H_0)

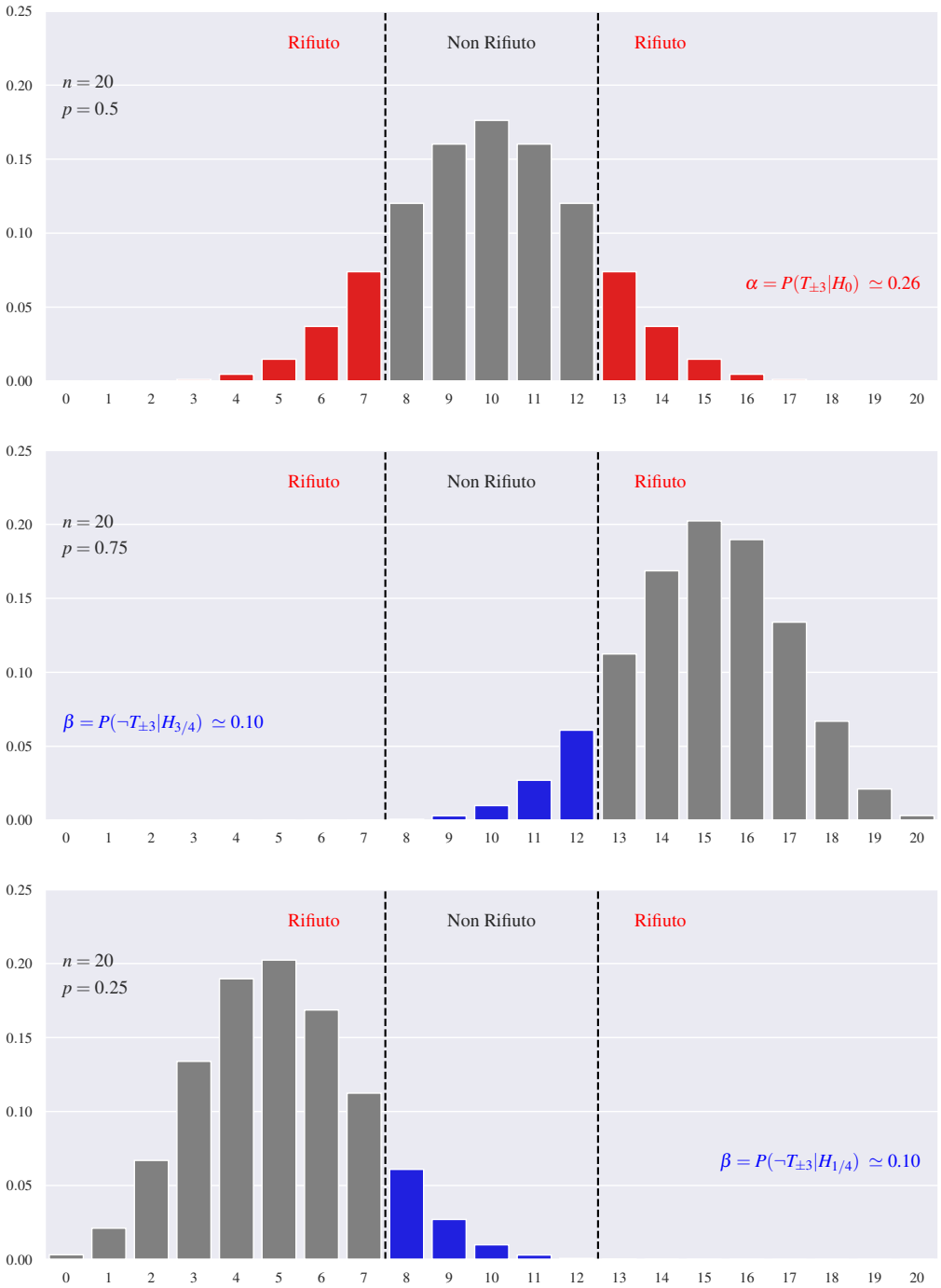
H_A : $\quad p = 3/4 \quad$ o $\quad p = 1/4$

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 $H_{1/4}$



7.6 Test a due code, errori I e II tipo

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

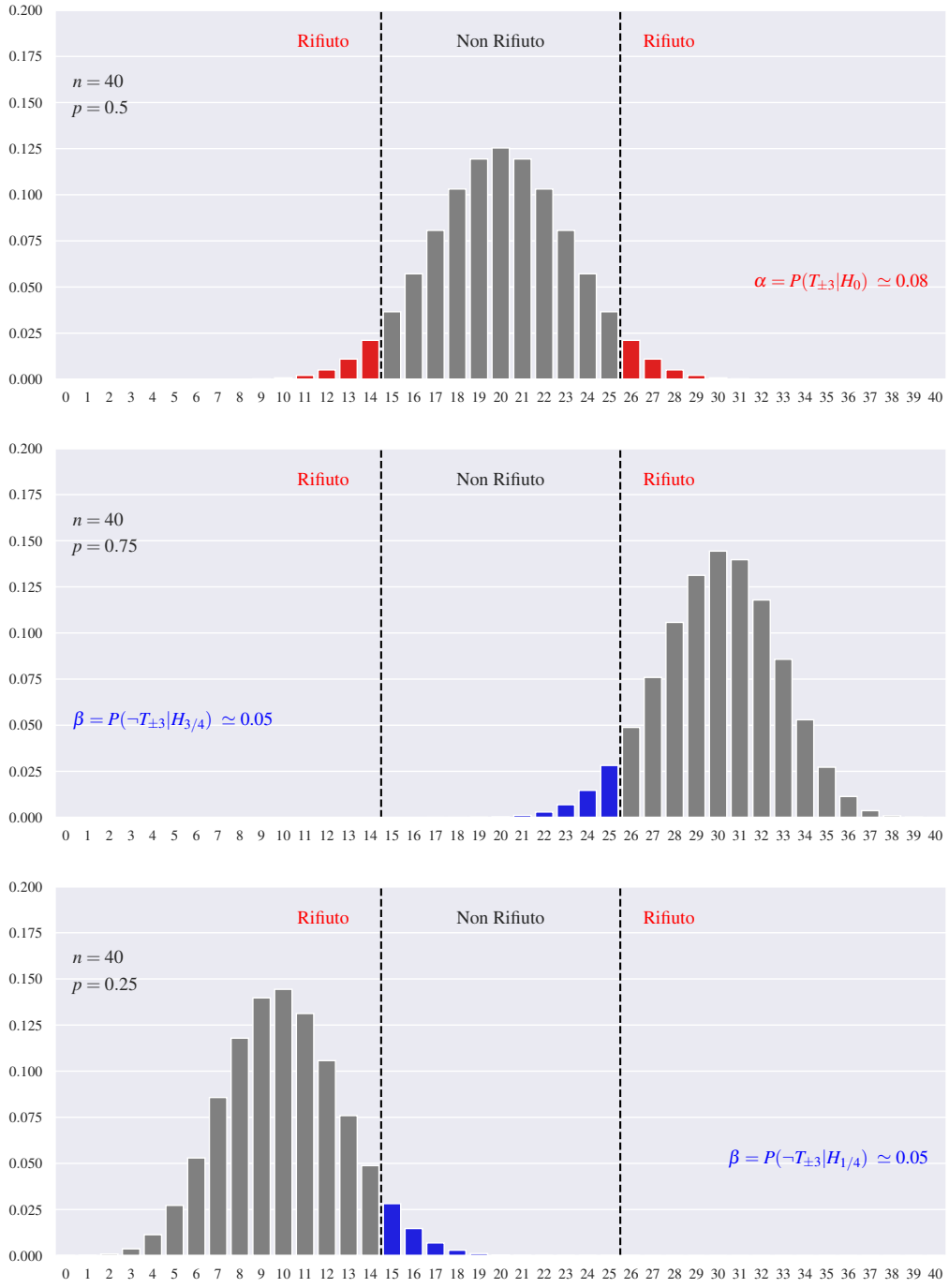


7.7 Test a due code, con campione più ampio

Supponiamo di raddoppiare la dimensione del campione ($n = 40$). Aggiustiamo la zona di rifiuto allo stesso modo ($k = 6$):

$$T_{\pm 6} = \{0, \ldots, 14 = n/2 - k\} \cup \{k + n/6 = 26, \ldots, 40\}$$

Entrambi gli errori diminuiscono.



7.8 Scatola di biglie colorate (esercizio in formato esame)

Un impianto di produzione di una fabblica riempie scatole con due tipi di biglie: rosse e blu. E stato garantito al cliente che le scatole contengono almeno il 40% di biglie rosse con una tolleranza del 5%.

Per controllare la qualità delle scatole confezionate preleviamo un campione 100 biglie. Sia N la v.a. che ritorna il numero di biglie rosse nel campione.

Domande.

1. Qual è l'potesi nulla?
2. Qual è l'potesi alternativa?
3. Qual è l'effect size?
4. Che confezioni devono essere scartate se vogliamo che il 95% delle confezioni non conformi al nostro obbiettivo vengano scartate?
5. Dato il risultato al punto precedente qual è la probabilità α di un erore del I tipo?

Si risponda assumendo note le funzioni in calce.

Risposte. Definiamo $n = 100, \quad \beta = 0.05, \quad p_0 = 0.4$.

1. H_0: $p = p_0$ dove p è la frazione di biglie rosse nella confezione in esame
2. H_A: $p < p_0$
3. $\delta = 0.05$
4. Scartiamo la confezione se $N \leq x$. Dove x è calcolato sapendo che
   $1 - \beta = \Pr\left(N \leq x \,\middle|\, p = p_0 - \delta\right) = \Pr(X \leq x)$ dove $X \sim \mathrm{B}(n, p_0 - \delta)$
   $$x = \texttt{qbinom}(1 - \beta, n, p_0 - \delta).$$
5. $\alpha = \Pr\left(N \leq x \,\middle|\, p = p_0\right) = \Pr(X \leq x)$ dove $X \sim \mathrm{B}(n, p_0)$
   $$\alpha = \texttt{pbinom}(x, n, p_0).$$

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

`pnorm(z)`$=P(Z \leq z)$, per $Z \sim \mathrm{N}(0,1)$

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

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

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

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

7.9 Prevalenza mancinismo 1 (esercizio 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 stiamo facendo?
4. Qual è la significatività del test?
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.

Si risponda assumendo note le funzioni in calce.

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 $X \sim B(n, p_0)$ ovvero `1-pbinom(`x`,`n`,`p_0`)`
5. $\beta = \Pr(X \leq x)$ dove $X \sim B(n, p_1)$ ovvero `pbinom(`x`,`n`,`p_1`)`
6. La potenza è $1 - \beta$.

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

`pnorm(z)`$=P(Z \leq z)$, per $Z \sim \mathrm{N}(0,1)$

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

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

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

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

7.10 Prevalenza mancinismo 2 (esercizio 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?

Si risponda assumendo note le funzioni in calce.

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. p-valore $= \Pr\left(X \geq x \;\middle|\; p = p_0\right)$, dove $X \sim \mathrm{B}(n, p_0)$
   $= 1 - \texttt{pbinom}(x, n, p_0)$

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

`pnorm(z)`$=P(Z \leq z)$, per $Z \sim \mathrm{N}(0,1)$

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

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

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

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

8 Z-test

8.1 Test a una coda

Si sospetta che una certa terapia faccia aumentare la pressione diastolica. Nella popolazione generale la pressione diastolica ha distribuzione $\mathrm{N}(\mu_0, \sigma^2)$ 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

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

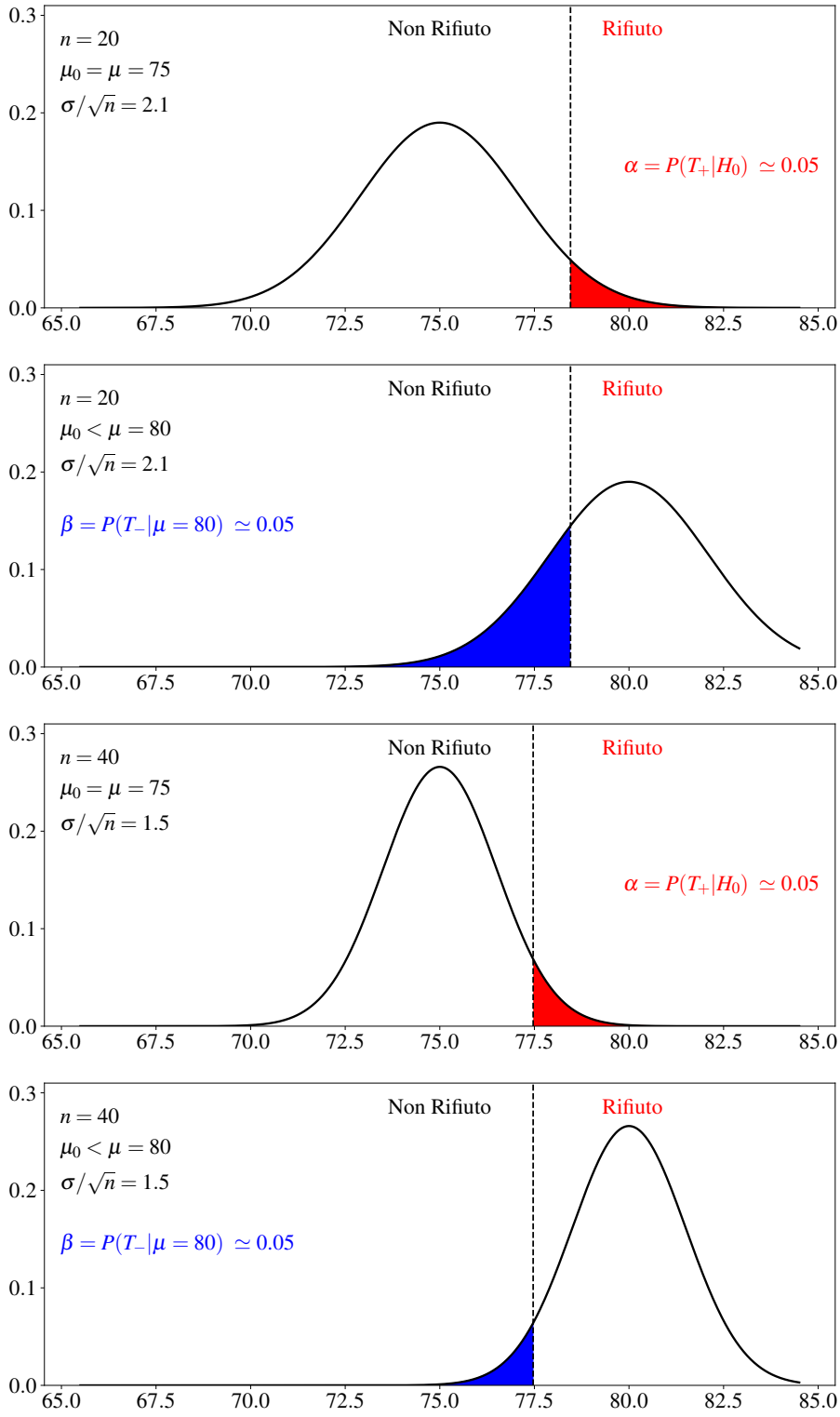
Dove X_i è la v.a. che dà la pressione dell'i-esimo paziente del campione. Rigetteremo H_0 se il valore ottenuto è suporiore ad un certo x_α che vogliamo fissare in modo che l'errore I tipo risulti uguale ad α. Quindi x_α dev'essere tale che x_α tale che $\Pr(\bar{X} > x_\alpha) = \alpha$.

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 $\mu > \mu_0$.

8.2 Una coda, errore I e II tipo

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 calcolare gli errori del II tipo prendiamo $\delta = 5$.



8.3 Pressione diastolica (esercizio formato esame)

(Testo ripetuto da sopra 8.) Si sospetta che una certa terapia faccia aumentare la pressione diastolica. Nella popolazione generale la pressione diastolica ha distribuzione $\mathrm{N}(\mu_0, \sigma^2)$ 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 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. Rifiutiamo H_0 se la misura è superiore al valore di soglia x_α. Quant'è x_α se vogliamo che la significatività sia $\alpha = 5\%$?
5. Dato il valore di soglia al punto precedente, qual è la potenza del test per un effect size $\delta = 2$?

Si risponda assumendo note le funzioni in calce.

Risposte Definiamo $\mu_0 = 75, \quad \sigma = 5.5, \quad , n = 64, \quad \delta = 2$. Sia $\bar{X}$ la v.a. che ritorna il valor medio della pressione diastolica del campione.

1. H_0: $\mu = \mu_0$ dove μ è pressione diastolica delle persone in terapia
2. H_A: $\mu > \mu_0$
3. Z-test a una coda (coda superiore)
4. x_α è tale che

   $$\alpha \;=\; \Pr\left(\bar{X} \geq x_\alpha \;\middle|\; \mu = \mu_0\right) \;=\; \Pr(Y \geq x_\alpha), \text{ dove } Y \sim \mathrm{N}(\mu_0, \sigma^2/n)$$

   standardizzando otteniamo

   $$\alpha \;= 1 - \Pr\left(Z \leq \frac{x_\alpha - \mu_0}{\sigma/\sqrt{n}}\right), \text{ dove } Z \sim \mathrm{N}(0,1).$$

   $$x_\alpha \;=\; \frac{\sigma}{\sqrt{n}}\,\texttt{qnorm}(1-\alpha) - \mu_0.$$

5. La potenza del test è $1 - \beta$ dove

   $$\beta \;=\; \Pr(\bar{X} \leq x_\alpha \mid \mu = \mu_0 + \delta) \;=\; \Pr(Y \leq x_\alpha), \text{ dove } Y \sim \mathrm{N}(\mu_0 + \delta, \sigma^2/n)$$

   standardizzando otteniamo

   $$\beta \;=\; \Pr\left(Z \leq \frac{x_\alpha - \mu_0 - \delta}{\sigma/\sqrt{n}}\right)$$

   $$\beta \;=\; \texttt{pnorm}\left(\frac{x_\alpha - \mu_0 - \delta}{\sigma/\sqrt{n}}\right)$$

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

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

8.4 Pressione diastolica (due code)

(In corsivo le differenze dal testo precedente 8.3.) Si sospetta che una certa terapia *modifichi* la pressione diastolica. Nella popolazione generale la pressione diastolica ha distribuzione $\mathrm{N}(\mu_0, \sigma^2)$ 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 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. Rifiutiamo H_0 se misura *differisce* da μ_0 più del valore di soglia x_α. Quant'è x_α se vogliamo che la significatività sia $\alpha = 5\%$?

Si risponda assumendo note le funzioni in calce.

Risposte Definiamo $\mu_0 = 75, \quad \sigma = 5.5, \quad , n = 64, \quad \delta = 2$. Sia $\bar{X}$ la v.a. che ritorna il valor medio della pressione diastolica del campione.

1. H_0: $\mu = \mu_0$ dove μ è pressione diastolica delle persone in terapia
2. H_A: $\mu \neq \mu_0$
3. Z-test a due code
4. x_α è tale che

   $$\alpha \;=\; \Pr\left(|\bar{X} - \mu_0| \geq x_\alpha \;\middle|\; \mu = \mu_0\right) \;=\; \Pr(|Y - \mu_0| \geq x_\alpha), \text{ dove } Y \sim \mathrm{N}(\mu_0, \sigma^2/n)$$

   standardizzando otteniamo

   $$\alpha \;= \Pr\left(|Z| \geq \frac{x_\alpha}{\sigma/\sqrt{n}}\right) = 2\Pr\left(Z \geq \frac{x_\alpha}{\sigma/\sqrt{n}}\right), \text{ dove } Z \sim \mathrm{N}(0,1).$$

   $$x_\alpha \;=\; \frac{\sigma}{\sqrt{n}}\,\texttt{qnorm}(1-\alpha/2).$$

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

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

8.5 Confezioni

Un certo processo produttivo confeziona scatole con biglie rosse e blue nella proporzione del 50%.

Sappiamo però che il 10% delle confezioni si discosta da questo obbiettivo per almeno il 5% (con uguale probabilità per eccesso e per difetto)

Per scartare le confezioni non conformi da ogni scatola vengono estratte 1000 biglie se la frazine di biglie rosse è $> 55\%$ o $< 45\%$ la confezione viene scartata.

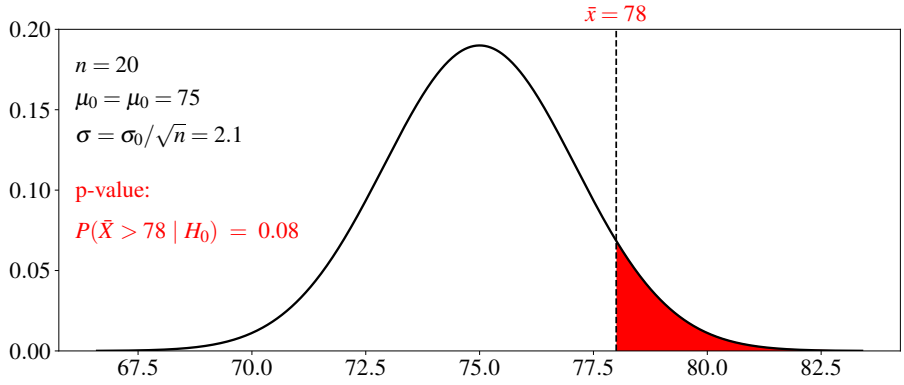
Qual è la probabilità che una scatola con esattamente 50% di biglie rosse venga scartata?

Qual è la probabilità che una scatola con più del 60% di biglie rosse non venga scartata?

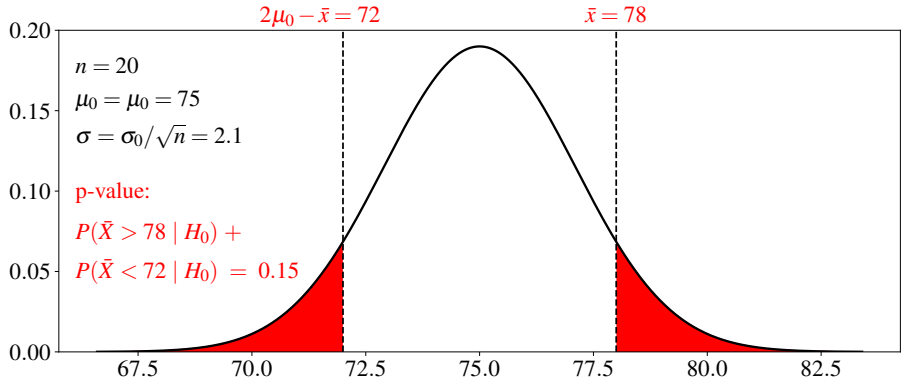
Quante confezioni che si discostano per almeno il 5% non vengono scartate.

8.6 Una coda, p-valore

(Continua l'esempio 8.) Supponiamo di ottenere $\bar{x} = 78.0$ da un campione di dimensione $n = 20$. Il p-valore di questa misura è $\Pr(\bar{X} \geq 78) = 1 - \Pr(\bar{X} \leq 78)$.



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).



8.7 Pressione diastolica cont. (esercizio formato esame)

Continua da 8.3

Domande

6. Supponiamo che il valore medio della pressione diastolica nel nostro campione sia 78. Con che p-valore possiamo rifiutare l'ipotesi nulla?

Si risponda assumendo note le funzioni in calce.

Risposte Definiamo $\bar{x} = 78$.

6. $$\begin{aligned}\text{p-valore} &= \Pr\left(Y \geq \bar{x} \;\middle|\; \mu = \mu_0\right) = \Pr(X \geq \bar{x}), \text{ dove } X \sim \mathrm{N}(\mu_0, \sigma^2/n) \\ &= 1 - \Pr\left(Z \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right), \text{ dove } Z \sim \mathrm{N}(0,1) \\ &= 1 - \texttt{pnorm}\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right).\end{aligned}$$

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

`pnorm(z)`=$P(Z \leq z)$, per $Z \sim \mathrm{N}(0,1)$

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

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

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

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

8.8 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.

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

8.9 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?

Si assumano noti i valori delle funzioni in calce

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

1. $\mu = \mu_0$.
2. $\mu > \mu_0$
3. Z-test una coda (superiore)
4. il p-valore è `1-pnorm(z)` dove $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.

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

`pnorm(z)`=$P(Z \leq z)$, per $Z \sim \mathrm{N}(0,1)$

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

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

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

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

9 T-test

9.1 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. $H_0 \quad \mu_0 = 70$
2. $H_A \quad \mu > \mu_0$
3. T-test coda superiore.
4. Il t-score è $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{79.17 - 70}{13.17/\sqrt{6}} = 1.71$
5. Il p-valore è $\Pr(T \geq t)$ dve $T \sim t(5)$. Ovvero 1-`pt(`t`, `$n-1$`)` `(0.074)`

9.2 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 $x_1, \dots, x_6 = 62, 92, 75, 68, 83, 95$ nel secondo gruppo $y_1, \dots, y_5 = 60, 95, 76, 69, 89$.

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 a due code per due popolazioni con dati accoppiati
3. Il t-score è $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_x + s_y^2/n_y}}$ dove

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

   $s_x^2 = \dfrac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2$ $\qquad$ $s_y^2 = \dfrac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \bar{y})^2$
4. Il p-valore è $2\Pr(T \geq |t|)$ dove $T \sim t(n_x + n_y - 2)$.
   Ovvero $2\texttt{pt}(-|t|,\ n_x + n_y - 2)$

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

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

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

9.3 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 $y_1, \ldots, y_6 = 60, 95, 76, 69, 89, 90$.

Domande

1. Qual è l'potesi nulla?
2. Qual è l'potesi alternativa?
3. Che test si applica?
4. Qual'è il p-valore ottenuto dai dati?

Si assumano noti i valori delle funzioni in calce

Risposte

1. H_0 $\mu = 0$ la media delle differenze è 0
2. H_A $\mu \neq 0$
3. T-test per due campioni accoppiati a due code
4. Il t-score è $t = \dfrac{\bar{z}}{s/\sqrt{n}} = -0.4264$ dove $z_i = x_i - y_i$ e
   $$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{6}(2 - 3 - 1 - 1 - 6 + 5) = -\frac{4}{6}$$
   $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = 14.67$$
4. Il p-valore è $\Pr(|T| \geq |t|)$ dove $T \sim t(n-1)$. Ovvero $2\,\texttt{pt}(-|t|, n-1)$

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

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

10 Intervallo di confidenza

Da una popolazione con distribuzione normale, media e varianza ignota, misuriamo i valori $x_1, \ldots, x_6 = 62, 92, 75, 68, 83, 95$. Si calcoli un intervallo di confidenza al 95% per la media di popolazione.

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i \tag{79.17}$$

$$s = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_i - \bar{x})^2} \tag{13.17}$$

L'intervallo di confidenza è $\bar{x} \pm \varepsilon$ dove ε è tale che

$$\begin{aligned} 0.95 &= \Pr\left(-\frac{\varepsilon}{s/\sqrt{n}} < T < \frac{\varepsilon}{s/\sqrt{n}}\right) \\ &= 1 - 2\Pr\left(T < -\frac{\varepsilon}{s/\sqrt{n}}\right) \end{aligned}$$

ovvero

$$0.025 = \Pr\left(T < -\frac{\varepsilon}{s/\sqrt{n}}\right)$$

quindi

$$\varepsilon = -\frac{\texttt{qt}(0.025, 5) \cdot \texttt{s}}{\sqrt{6}}$$

N.B. Se la deviazione standard della popolazione fosse stata nota, diciamo $\sigma = 13.17$ allora avremmo ottenuto

$$\varepsilon = -\frac{\texttt{qnorm}(0.25) \cdot \sigma}{\sqrt{6}}$$

11 Regressione lineare

Incompleto

Abbiamo le seguanti coppie di dati

$(x_i) = (1, 2, 3, 4, 5)$ $(y_i) = (2, 5, 5, 8, 9)$

Si calcoli un intervallo di confidenza al 95% per i valori dell'intercetta β_0 e del cefficente angolare β_1 della retta interpolante.

Risposta

$\bar{x} = \frac{15}{5} = 3$

$\bar{y} = \frac{30}{5} = 6$

$(x_i y_i) = (2, 10, 15, 36, 45)$

$(x_i^2) = (1, 4, 9, 16, 25)$

$\overline{xy} = \frac{108}{5}$

$\overline{x^2} = \frac{55}{5} = 11$

$\beta_1 = 1.8$

$\beta_0 = 0.6$

$(e_i) = (-0.4, 0.8, -1, 1.2, -0.6)$

$(e_i^2) = (0.16, 0.64, 1, 1.44, 0.36)$

$s_0 = \frac{5}{3} 3.6 = 6$