Chapter 1

Teoria (il minimo sindacale)

Per esempi e 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 coppia $E_1, E_2 \in \mathcal{P}(\Omega)$.

2 Variabili aleatorie

Sia R un insieme quasiasi. Una **variabile aleatoria** è una funzione $X : \Omega \to R$. 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**. Una variabile aleatoria non numerica è detta **qualitativa** o **categorica**.

3 Distribuzione di probabilità discreta

Come sopra $X : \Omega \to R$ è una variabile aleatoria. Dato $x \in R$ e $A \subseteq R$ scriveremo

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

La funzione $\Pr(X = x)$ si chiama **distribuzione di probabilità (probability mass function)**. La funzione $\Pr(X \leq x)$ si chiama **funzione di ripartizione (cumulative distribution function)**.

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

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

Ovvero la probabilità è concentrata nei punti di R. Invece una variabile 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

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

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.

4 Variabili aleatorie di Bernoulli

Una variabile aleatoria X si dice di bernulli se **Bernoulli** se $\operatorname{img} X = \{0, 1\}$.

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

Viceversa, la v.a. di Bernoulli associata ad un evento E è spesso denotata con 1_E

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

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

5 Probabilità condizionata

Esempi: Popolazione maschile e femminile (probabilità totali)
Fumatori (regola di Bayes)
Rain forcasts (Bayes rule)

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

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

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

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

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

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

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

6 Indipensenza stocastica

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 $\Pr(A|B) = \Pr(A)$ 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 $y \in \operatorname{img} Y$

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

Nel caso di variabili aleatorie continue la condizione diventa

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

7 Esperimenti ripetuti: prodotto di spazi di probabilità

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 $A = X^{-1}[\,1\,]$.

Il lancio ripetuto di una monetina è modellato con lo spazio campionario Ω^2. L'insieme $A \times \Omega$ è l'evento dello spazio Ω^2 che corrisponde ad ottenere 1 nel primo lancio. L'insieme $\Omega \times \neg A$ 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 $1\,0$.

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 indipendenti.

8 Variabili aleatorie binomiali

9 Valore atteso e varianza

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

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

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

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

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

$$\boldsymbol{\sigma} \quad = \quad \sqrt{\mathrm{Var}(X)}$$

Le lettere μ e σ vengono usate quando è chiaro a quale variabile ci si riferisce. per evitare ambiguità a volte si scrive μ_X e σ_X.

10 Standardizzazione

Sia X una v.a. con media μ e deviazione standaed 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**).

11 Diagnostic tests

Esempi: HIV test (regola di Bayes)

Let $\boldsymbol{T_+}$ and $\boldsymbol{T_-}$ be the events that the result of a diagnostic test is positive or negative respectively. Let $\boldsymbol{D}$ 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.

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

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

13 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 denotavamo $\neg D$).
- 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 $\Pr(T_-|H_A)$.

14 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 $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$.

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

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.

15 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

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

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

$$= \Pr\left(W \leq w\right) + \Pr\left(W \geq w\right).$$

Comunque, quando H_A permette un test ad una coda (per esempio) una tra possibilità quindi il p-valore si riduce a $\Pr\left(W \leq w\right)$ o $\Pr\left(W \geq w\right)$.

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 coppia $E_1, E_2 \in \mathcal{P}(\Omega)$ di insiemi 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

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

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

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

In questo modo diamo una misura di probabilità a molti eventiche 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 se le nostre minure (di peso e diametro) sono molto imprecise il numero di possibili valori potrebbe essere comunque abbastnza piccolo e potrebbe concenire interpretarle come variabili discrete.

3 Pobabilità totali

3.1 Maschi e femmine

4 Regola di Bayes

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

4.1 Fumatori e non fumatori

Tra le persone affette da una certa patolopgia A, il 20% è fumatore. La prevalenza di A tra i fumatori è del 10% nella popolazione generale è del 2%. Calcolare la probabilità che un fumatore ha di essere affetto da A.

F insieme dei fumatori

A insieme persone affette da A

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

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

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

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

4.2 Hemophilia

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 affected. What is the probability that she is a carrier?

Ω set of women with mother carrier, father and husband unaffected, one son

C set of women that are carrier

S_{na} set of women whose son is not affected

$\Pr(C) = 1/2$

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

$\Pr(C|S_{na}) = \dfrac{\Pr(S_{na}|C) \cdot \Pr(C)}{\Pr(S_{na})} = \dfrac{\Pr(S_{na}|C) \cdot \Pr(C)}{\Pr(S_{na}|C) \cdot \Pr(C) + \Pr(S_{na}|\neg C) \cdot \Pr(C)} = \dfrac{1/4}{1/4 + 1/2} =$
$1/3$

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

$$
\begin{aligned}
\Pr(R\,|T_+) &= \frac{\Pr(R)\cdot\Pr(T_+|R)}{\Pr(T_+)} \\
&= \frac{\Pr(R)\cdot\Pr(T_+|R)}{\Pr(T_+|R)\cdot\Pr(R)+\Pr(T_+|\neg R)\cdot\Pr(\neg R)}
\end{aligned}
$$

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 Diagnostic test: HIV

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 $\Pr(D) = .001$

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

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 an HIV infected partner that the test was taken in South Africa where the prevalence is estimated to be around 20%

$$
\Pr(D \mid +) = 0.943
$$

7 Test Binomiale

Nella pratica questo test è sempre sostituito da un test sulle proporzioni. Comunque in questa versione è concettualmente più semplice.

7.1 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 $\{0, \ldots, n\}$. 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.

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.

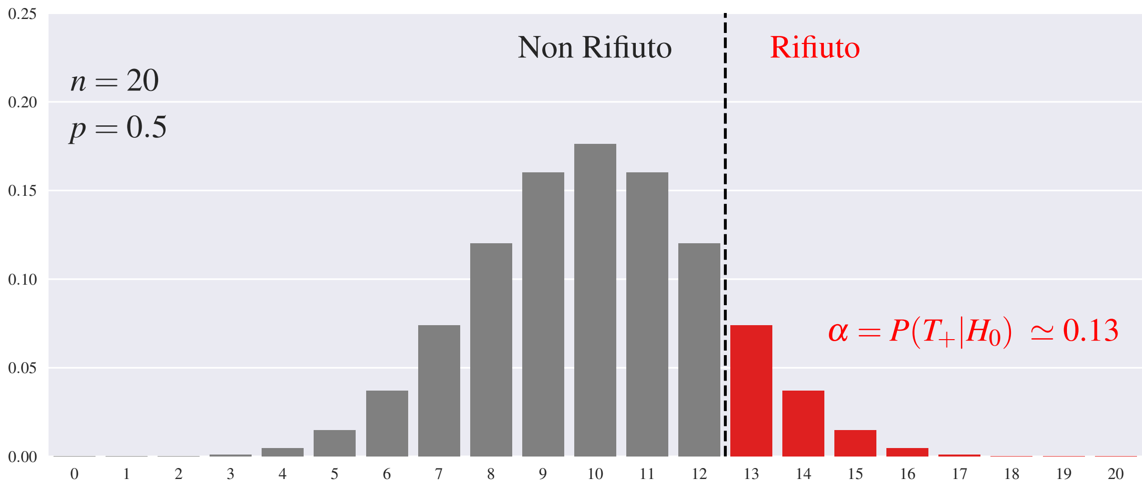
7.2 Una coda, errore I tipo

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.

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

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

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



7.3 Testa una coda, errore II tipo

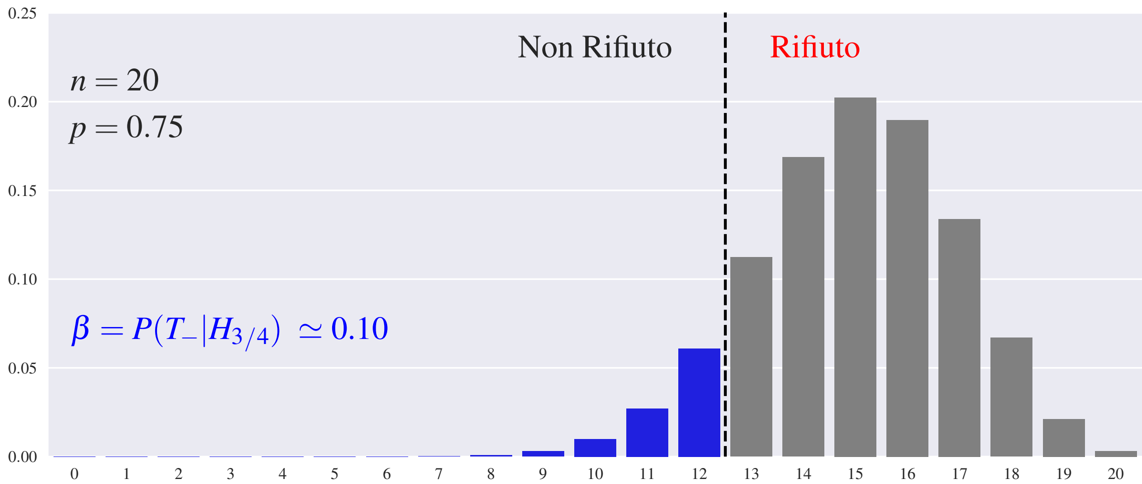
La probabilità dei falsi negativi può essere espressa in funzione di p (abbiamo solo assunto che $> 1/2$)

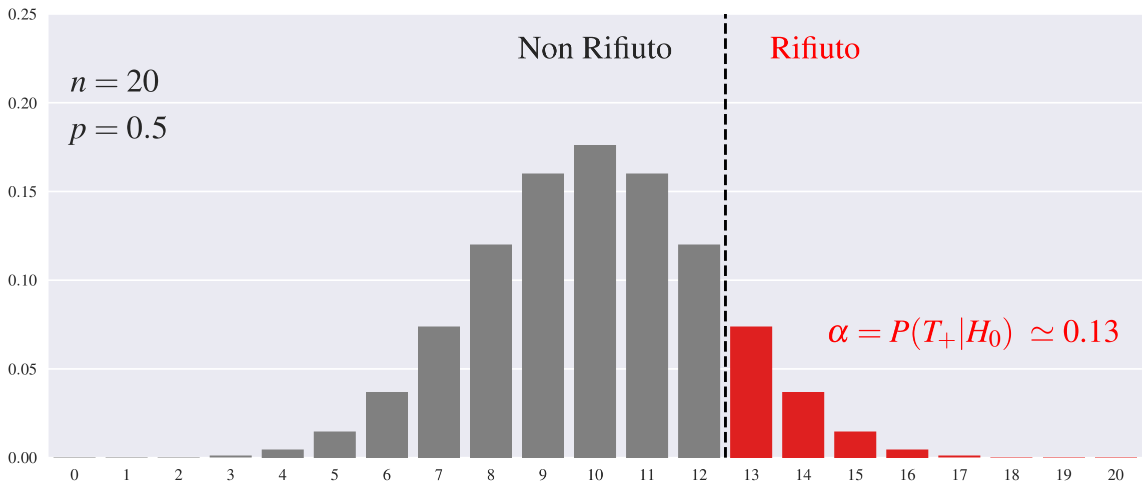
$$\Pr(T_- \mid H_A) \quad = \quad \Pr(X < k \mid H_A) \quad = \quad \sum_{i=1}^{k-1} \binom{n}{i} p^i (1-p)^{n-i}$$

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.

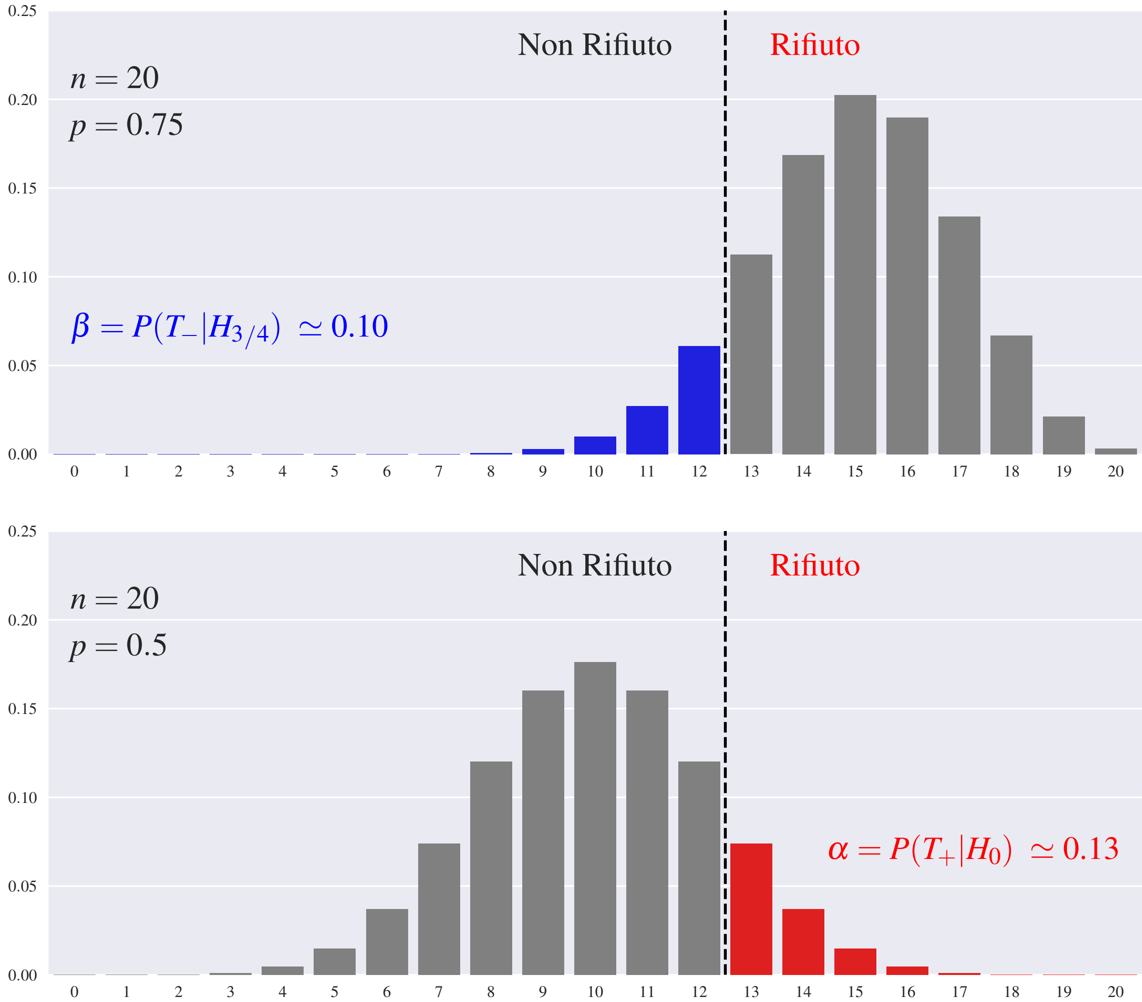




7.4 Effect size: δ

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.



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

$H_0: \quad p = 1/2$

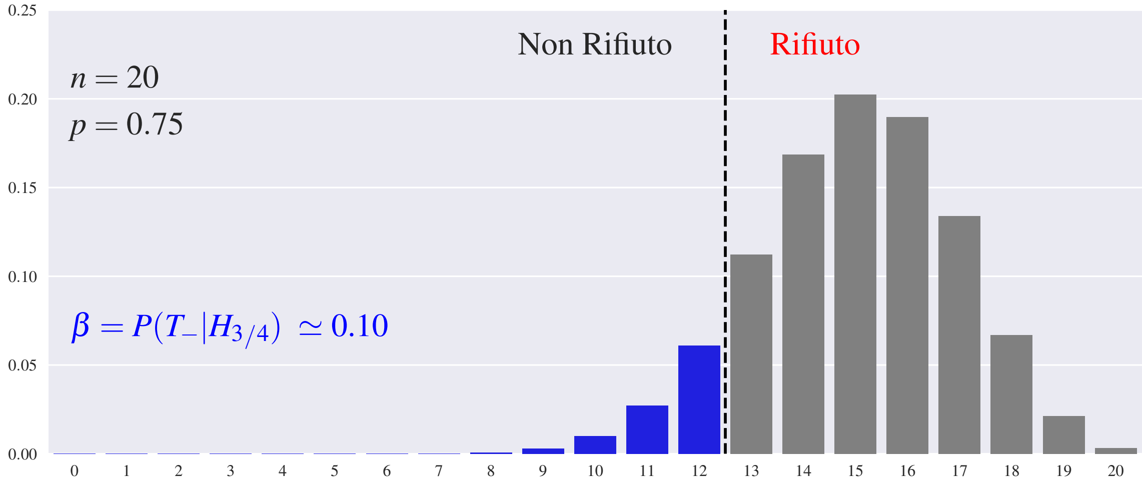
$H_A: \quad p \neq 1/2$

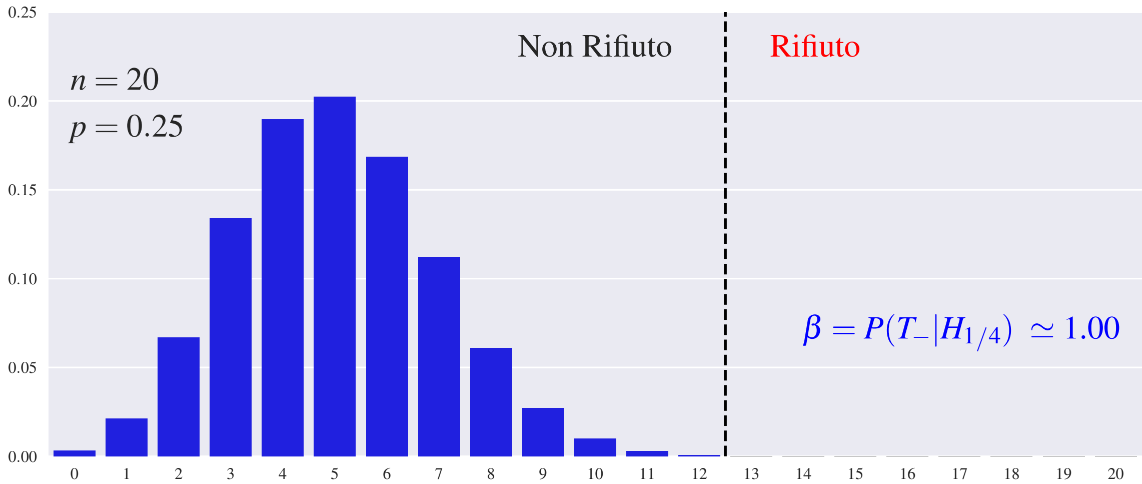
Verifichiamo prima che **la zona di rifiuto dei paragrafi precedenti NON è adatta** alla nuova situazione. L'analisi di $\Pr(T_+|H_0)$ rimane invariata (inaftti l'insieme H_0 non è cambiato).

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

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

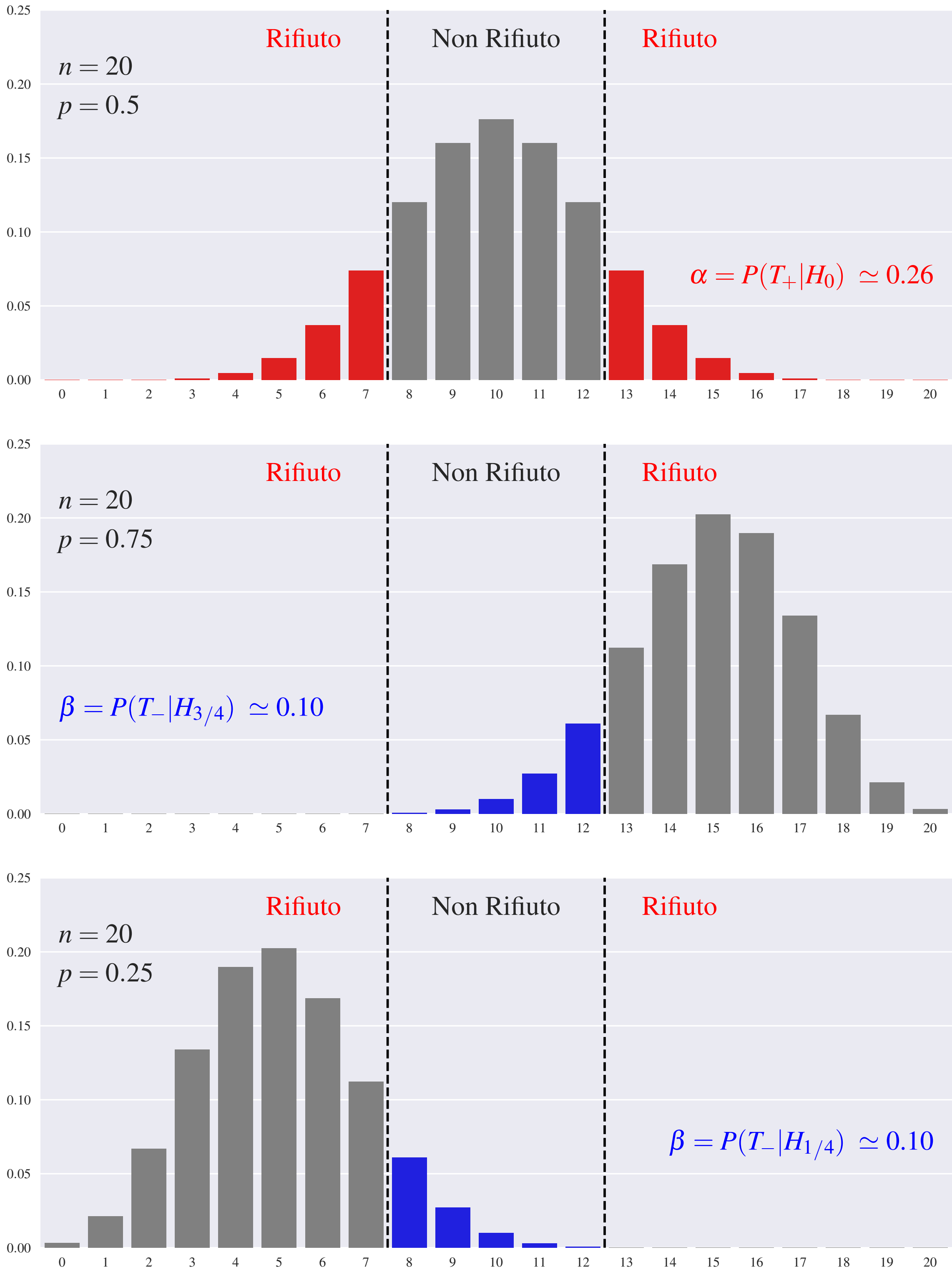
Possiamo immaginare $H_A = H_{1/4} \cup H_{3/4}$. Se sostituiamo H_A con $H_{3/4}$ il grafico rimane come quello già discusso. Ora però dobbiamo considerare il caso il cui la moneta appartenga all'insieme $H_{1/4}$





7.6 Test a due code, errori I e II tipo

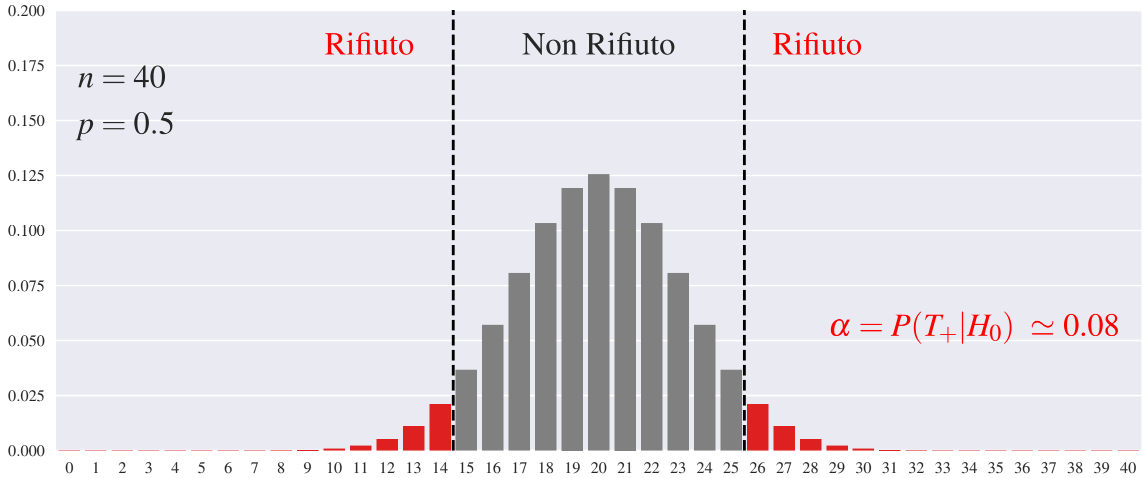
Per riparare il problema discusso al paragrafo precedente. prendiamo come zona di rifiuto $T_+ = \{0, \ldots, 7 = n - k\} \cup \{k = 13, \ldots, 20 = n\}$

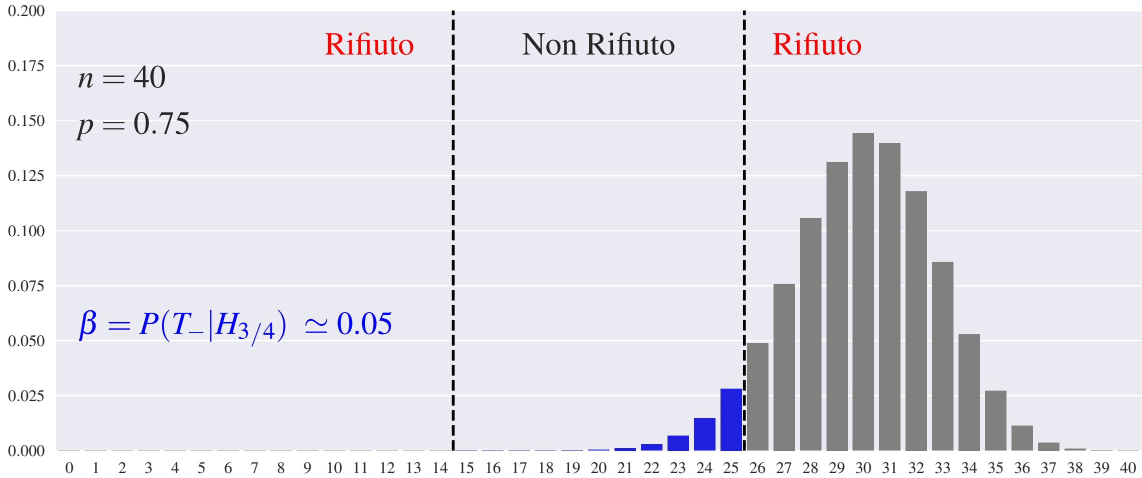


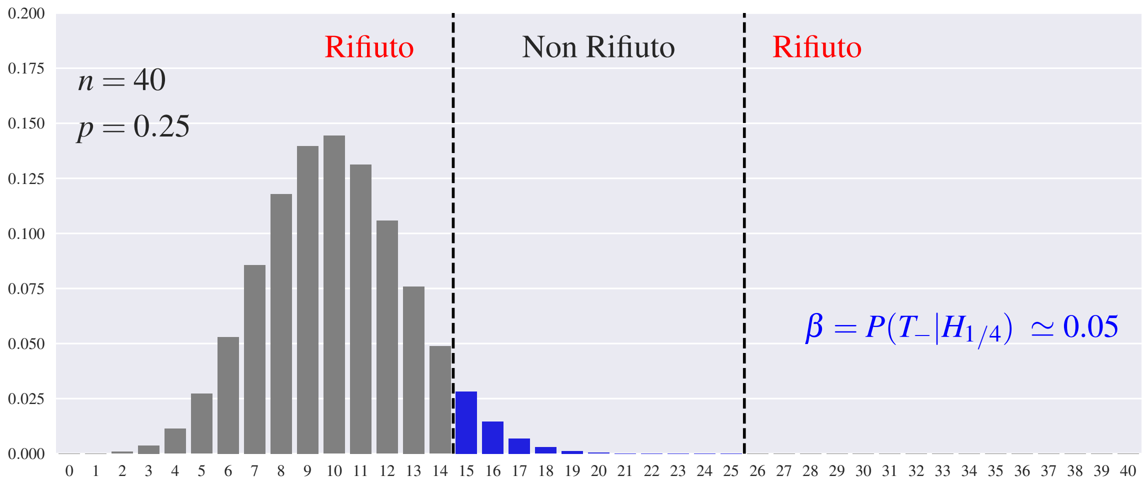
7.7 Test a due code, con campione più ampio

Supponiamo di raddoppiare la dimensione del campione ($n = 40$). Aggiustiamo la zona di rifiuto allo stesso modo ($k = 26$): $T_+ = \{0, \ldots, 14 = n - k\} \cup \{k = 26, \ldots, 40 = n\}$

Entrambi gli errori diminuiscono.







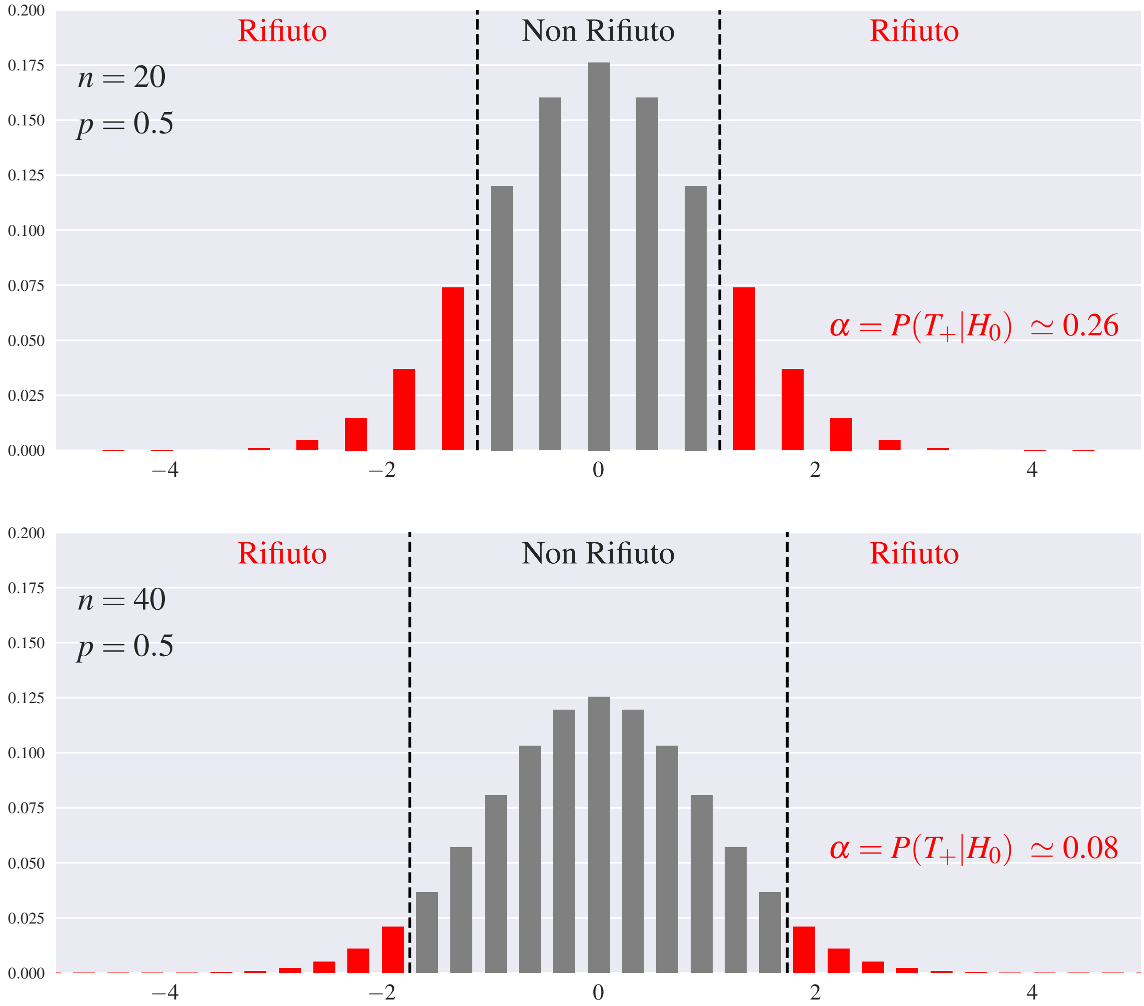
7.8 Standardizzazione (1)

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:

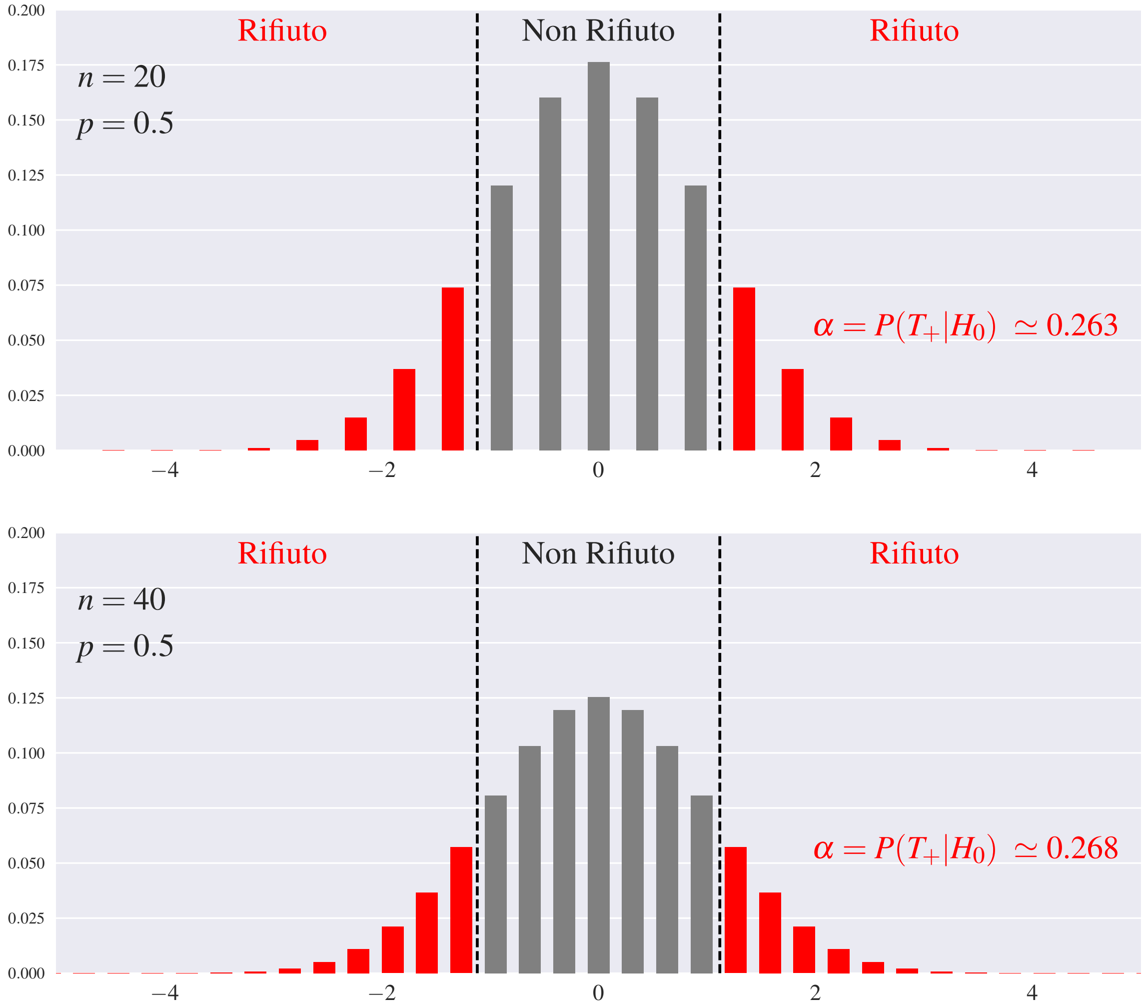


7.9 Standardizzazione (2)

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

$$\begin{aligned}\frac{k-np}{\sqrt{np(1-p)}} &= 1.34 \qquad \text{se } n=20,\ k=13,\ p=05 \\ &= 1.90 \qquad \text{se } n=40,\ k=26,\ p=05\end{aligned}$$

Se per entriambi i casi prendiamo la stessa regione di rifiuto misurata in punteggio Z, diciamo $(-\infty,-1.34]\cup[1.34,+\infty)]$ 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 Z-test

8.1 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

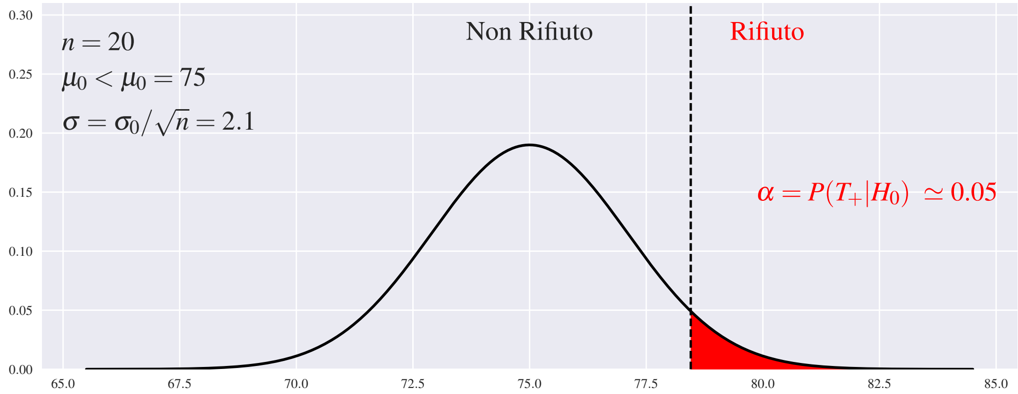
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

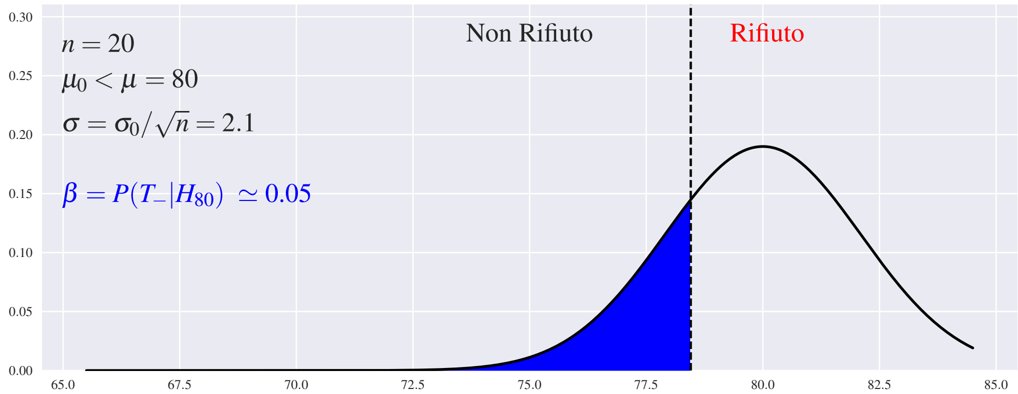
Dove X_i è la v.a. che dà la pressione dell'i-esimo pazione del campione. Rigetteremo H_0 se il valore ottenuto è suporiore ad un certo x_α che vogliamo fissare in modo che l'errore I tipo risulti uguale ad α. Quindi x_α dev'essere tale che x_α tale che $\Pr(\bar{X} > x_\alpha) = \alpha$.

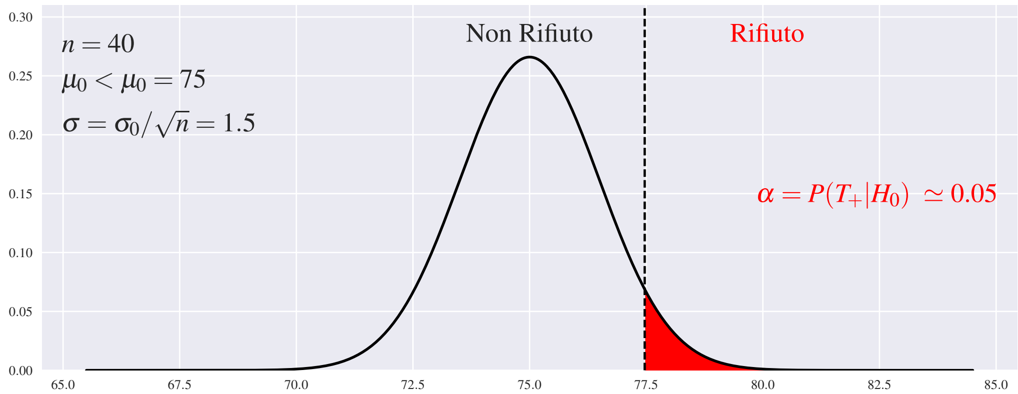
Se H_0 è vera, $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$.

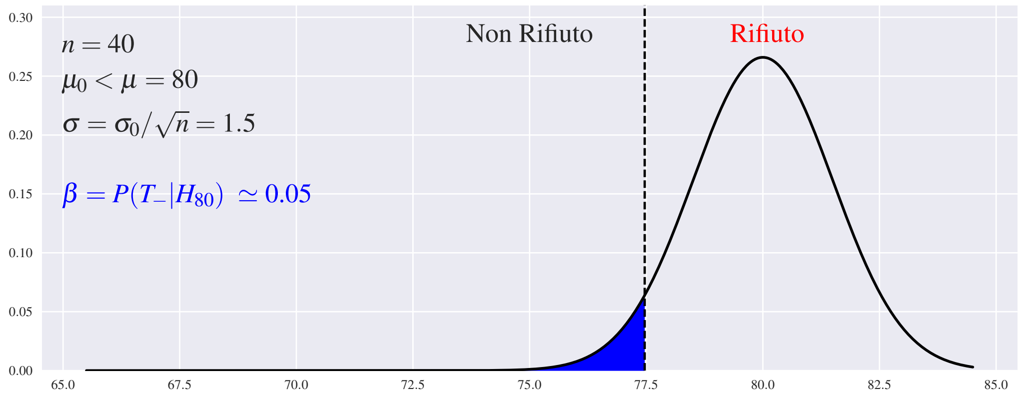
8.2 Una coda, errore I e II tipo

Qui rappresentiamo gli errori del I e II tipo per campioni di dimensione $n = 20, 40$. Per gli errori del II tipo prendiamo $\delta = 5$.









8.3 Una coda, p valore. **Z-test**

Supponiamo di ottenere $\bar{x} = 78.0$ da un campione di dimensione $n = 40$. Il p-valore di questa misura è $\Pr(\bar{X} > 78) = 1 - \Pr(\bar{X} \leq 78)$.

9 Esercizi vari

9.1 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: $1 - (0.98)^{50} = 64\%$

10 Inferenza Bayesiana

In un casinó si gioca a testa o croce (1 o 0) con tre tipi di monete che hanno probabilità di successo $1/2$, $2/3$, e $1/3$. Il casinò sceglie la moneta e continua a giocare con la stessa moneta per tutto il giorno. Il giocatore non sa quale moneta sia scelta.

Assumiamo per fissare le idee che il casinò scelga la monete a sorte e che le tre monete abbiano la stessa probabilità di essere scelte. Supponiamo osservare la sequenza 010 e di voler puntare sul prossimo lancio. Ci interessa sapere con che probabilità possiamo assumere quale moneta.

Siano $M_{1/2}$, $M_{2/3}$ ed $M_{1/3}$ gli eventi che corrispondono alla scelta delle tre monete $M_{1/2}$, $M_{2/3}$ ed $M_{1/3}$

$$\begin{aligned}
\Pr(010 \mid M_{1/2}) &= \frac{1}{2^3} = \frac{1}{8} \\
\Pr(010 \mid M_{2/3}) &= \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{2}{9} \\
\Pr(010 \mid M_{1/3}) &= \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}
\end{aligned}$$

Usando il teorema delle probabilità totali possiamo facilmente calcolare

$$\Pr(010) = \frac{1}{3}\left(\frac{1}{8} + \frac{2}{9} + \frac{4}{9}\right) = \frac{19}{72}.$$

Quindi possiamo calcolare la probabilità di $M_{1/2}$, $M_{2/3}$ ed $M_{1/3}$ data l'informazione 010.

$$\Pr(M_{1/2} \mid 010) = \frac{\Pr(010 \mid M_{1/2}) \cdot \Pr(M_{1/2})}{\Pr(010)} = \frac{1/8 \cdot 1/3}{19/72}$$

$$\Pr(M_{2/3} \mid 010) = \frac{\Pr(010 \mid M_{2/3}) \cdot \Pr(M_{2/3})}{\Pr(010)} = \frac{2/9 \cdot 1/3}{19/72}$$

$$\Pr(M_{1/3} \mid 010) = \frac{\Pr(010 \mid M_{1/3}) \cdot \Pr(M_{1/3})}{\Pr(010)} = \frac{4/9 \cdot 1/3}{19/72}$$