Chapter 1

Teoria (il minimo sindacale)

Per esempi e esercizi seguire i link

1 Variabili aleatorie

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 rapresentano proprietà osservabili.

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

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

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**.

2 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 che chiameremo **successo** così definito

$$X^{-1}[1] \quad = \quad \{\omega : X(\omega) = 1\}$$

Chiameremo $\boldsymbol{p} = P(X = 1)$ la **propbabilità 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 $\boldsymbol{X \sim B(1, p)}$.

3 Distribuzione di probabilità discreta

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

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

La funzione $P(X = x)$ si chiama **distribuzione di probabilità discreta (probability mass function)**. La funzione $P(X \leq x)$ 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 $A \subseteq R$

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

Ovvero la probabilità è concentrata nei punti di R. Invece una variabile continua se $P(X = x) = 0$ per ogni $x \in R$. Per le variabili continue tutta l'informazione è contenuta nella funzione di ripartizione possiamo solo scrivere

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

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

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 Probabilità condizionata

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

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

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

Questo si legge **probabilità di A dato Φ**. Si verifica facilmente che $P(\,\cdot\,|\Phi)$ soddisfa a tutte le proprietà di $P(\,\cdot\,)$ 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

$$P(C) \quad = \quad \sum_{i=1}^{n} P(A_i) \cdot P\big(C|A_i\big).$$

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

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

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

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

5 Indipensenza stocastica

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

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

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

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

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

Nel caso di variabili aleatorie continue la condizione diventa

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

6 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 scriviamo A per $X^{-1}[\,1\,]$.

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 $P(A) \cdot P(B)$. Gli eventi $A \times \Omega$ e $\Omega \times B$ sono quindi indipendenti.

7 Variabili aleatorie binomiali

8 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

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

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

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

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

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

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.

9 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 $P(D)$ 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, $P(T_+|D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(T_-|\neg D)$
- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D|T_+)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(\neg D|T_-)$
- The **prevalence of the disease** is the marginal probability of disease, $P(D)$

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.

10 Campioni e statistiche

Un campione $\{X_1, \ldots, X_n\}$ è un insieme di n v.a. indipendenti e identicamente distribuite. Diremo che il campione ha **rango** (o **dimensione**) n.

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**

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

11 Test d'ipotesi

Esempi: Bernoulli test (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 $\mathbf{H_0}$ definisce l'insieme dei *sani* (qui denotato $\neg D$).
- L'**ipotesi alternativa** denotata con $\mathbf{H_A}$ descrive la *patologia*, ovvero definisce l'insieme dei *malati* (qui denotato D).
- $\mathrm{H_A}$ non è semplicemente la negazione di $\mathrm{H_0}$ nel caso in cui alcuni risultati vengono esclusi perché ritenuti impossibili.
- L'espressione: **$\mathbf{H_0}$ può essere rifiutata** è sinonima di *l'esito del test è positivo*. Noi denotiamo l'evento con T_+.
- L'espressione: **$\mathbf{H_0}$ 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 $P(T_+|\neg D)$ e $P(T_-|D)$.

N.B. È vacile progettare un test che minimizza solo una tra $P(T_+|\neg D)$ o $P(T_-|D)$. In un caso estremo: un test che a prescindere dai dati rifiuta sempre $\mathrm{H_0}$ avrà banalmente $P(T_-|D) = 0$; invece un test che non rifiuta mai $\mathrm{H_0}$ avrà $P(T_+|\neg D) = 0$.

La difficoltà nel progettare il test è trovare il giusto equilibrio tra $P(T_+|\neg D)$ e $P(T_-|D)$.

12 Test d'ipotesi (tavola riassuntiva)

Evidenziamo in questa tavola la terminologia usata nei test statistici (che differisce da quella usata per i test diagnostici). Molto comune sono i simboli $\boldsymbol{\alpha} = P(T_+|\neg D)$ e $\boldsymbol{\beta} = P(T_-|D)$.

$T_+ \cap \neg D$ falso pos. **errore del I tipo** $P(T_+\|\neg D) = \boldsymbol{\alpha}$ **significatività**	$T_+ \cap D$ corretto pos. $P(T_+\|D) = 1 - \boldsymbol{\beta}$ sensibilità **potenza**
$T_- \cap \neg D$ corretto neg. $P(T_-\|\neg D) = 1 - \boldsymbol{\alpha}$ **sepecificità**	$T_- \cap D$ falso neg. **errore del II tipo** $P(T_-\|D) = \boldsymbol{\beta}$

13 Il p-valore

Chapter 2

Esempi ed esercizi

1 Pobabilità totali: maschi e femmine

2 Regola di Bayes: fumatori e non fumatori

3 Bayes rule: 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
$P(R) = 5/365$	it rains 5 days out of the year
$P(\neg R) = 1 - P(R) = 360/365$	
$P(T_+\|R) = 0.9$	when it rains, 90% of the times rain is predicted
$P(T_+\|\neg R) = 0.1$	when it does not rain, 10% of the times rain is predicted

We want to know

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

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

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 $= 1$.

5 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 $P(D|T_+)$ given the sensitivity, $P(T_+|D) = .997$, the specificity, $P(T_-|\neg D) = .985$, and the prevalence $P(D) = .001$

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

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

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

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

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

6 Bernoulli test (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 scivendo

$\mathrm{H}_0 : p = 1/2$

$\mathrm{H}_\mathrm{A} : p > 1/2$

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. Stoamo descivendo una famiglia di test, uno per ogni scelta dei parametri n e k. Vogliamo vedere come variano gli errori del I e del II tipo al variare di questi parametri.

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

Condizionando a $\neg D$ otteniamo $X \sim \mathrm{B}(n, 1/2)$. Invece condizionando a D otteniamo $X \sim \mathrm{B}(n, p)$ con $p > 1/2$ ignota.

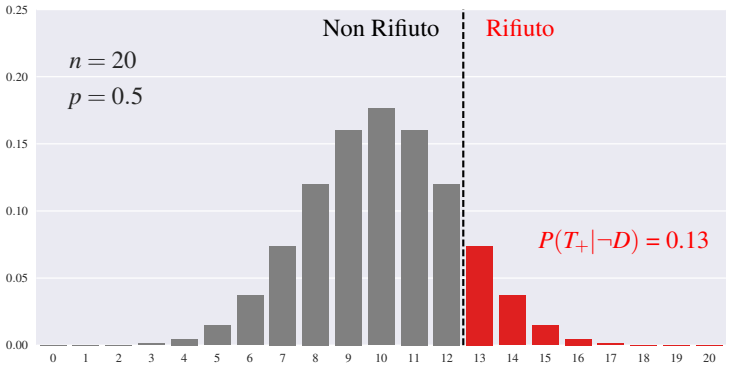
7 Bernoulli test (a una coda). Errore I tipo.

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

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

$$P(T_+|\neg D) \quad = \quad \sum_{i=k}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1-\frac{1}{2}\right)^{n-i} \quad = \quad \frac{1}{2^n} \sum_{i=k+1}^{n} \binom{n}{i}$$

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



8 Bernoulli test (a una coda). Errore II tipo.

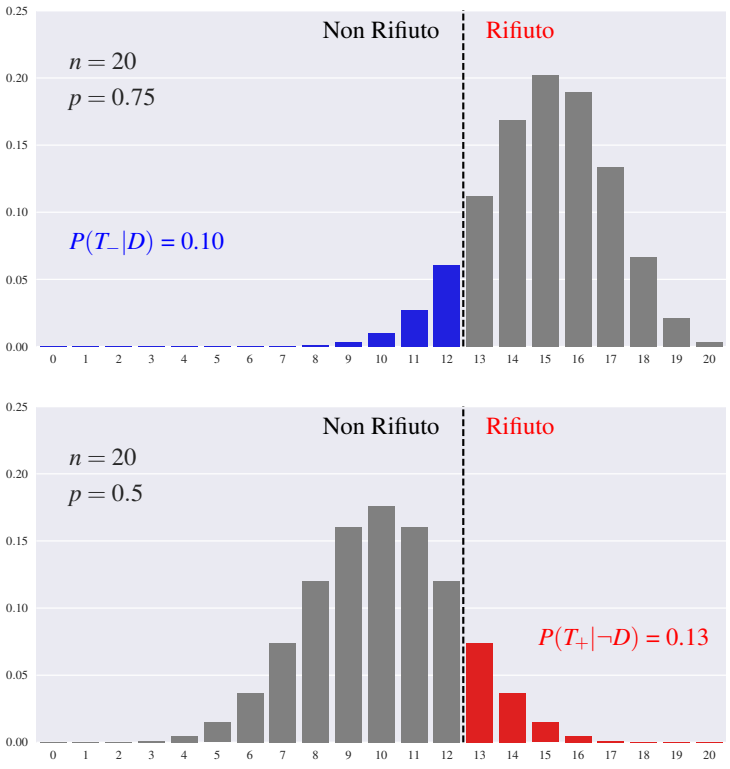
La probabilità dei falsi negativi può essere espressa in funzione di p (abbiamo solo assunto che $> 1/2$)

$$P(T_-|D) \quad = \quad \sum_{i=1}^{k-1} \binom{n}{i} p^i (1-p)^{n-i}$$

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

$\mathrm{H_A} : p = 3/4$

Rappresentiamo la distribuzione di X nel caso in cui l'ipotesi alterntiva è vera. Per confronto lo accostiamo al grafico del paragrafo precedente.



Cosa possiamo dire sul caso generale $\mathrm{H_A} : p > 1/2$?

Al crescere di p la distribuzione si sposta verso destra quindi gli errori del II tipo diminuiscono. Di converso, se p si avvicina a $1/2$ l'errore aumenta al limite quando $p \approx 1/2$ avremo $\alpha + \beta \approx 1$.

9 Bernoulli test (a due code).

Nell'esempio precedente avevamo un'informazione certa sul tipo di difetto delle monete: sapevamo che $p > 1/2$. Supponiamo questa manchi. Avremo quindi

$\mathrm{H}_0 : p = 1/2$

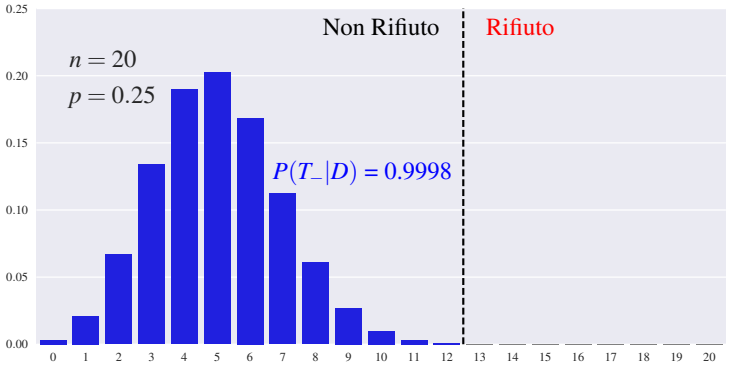
$\mathrm{H}_\mathrm{A} : p \neq 1/2$

Verifichiamo prima che il test discusso nei paragrafi precedenti **NON è adatto** alla nuova situazione. L'analisi di $P(T_+|\neg D)$ rimane invariata (inaftti l'insieme $\neg D$ non è cambiato).

Per semplificare la discussione dell'errore dell'secondo supponiamo per il momento che

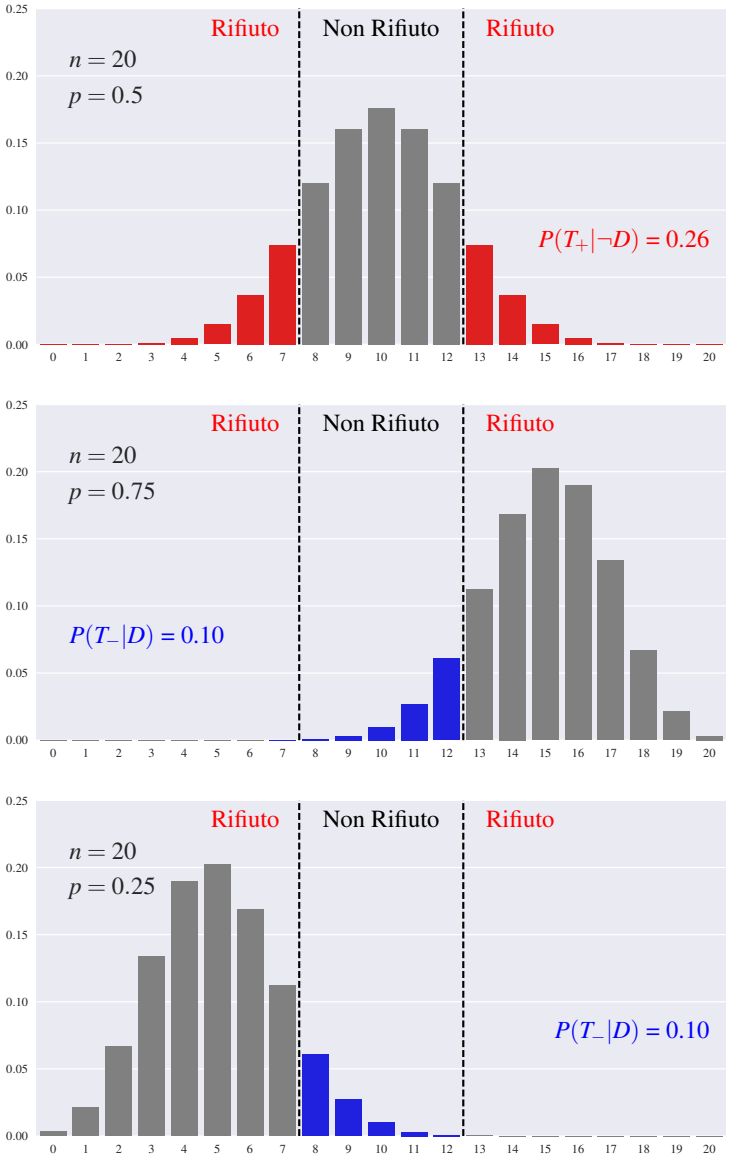
$\mathrm{H}_\mathrm{A} : p = 3/4 \quad \text{o} \quad p = 1/4$

Possiamo immaginare $D = D_{1/4} \cup D_{3/4}$. Il grafico del paragrafo precedente è valido ora se sostituiamo D con $D_{3/4}$. Ora però dobbiamo considerare il caso il cui la moneta appartenga all'insieme $D_{1/4}$



10 Bernoulli test (a due code). Errori I e II tipo.

Supponiamo di scegliere come zona di rifiuto $T_+ = \{0, \ldots, 7\} \cup \{13, \ldots, 20\}$



11 Bernoulli test (a due code) campione più ampio.

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

Entrambi gli errori diminuiscono.