Matematica e Statistica con R

Federico Comoglio e Maurizio Rinaldi

24 marzo 2016

Indice

Capitolo 1

Probabilità

Riprendiamo qui sinteticamente alcune definizioni.

1.1 Lo spazio campionario

Lo spazio campionario Ω (*sample space*) è l'insieme di tutte le uscite possibili di un esperimento. Per Esempioo

- Dado: lo spazio campionario consiste nelle uscite dei numeri da 1 a 6.
- Moneta: lo spazio campionario consiste di 2 possibili uscite "esce testa", "esce croce"
- Misure di lunghezza: lo spazio campionario è un intervallo del semiasse positivo della retta. Come si vede lo spazio campionario può essere discreto o continuo. Potrebbe essere anche una combinazione dei due.

1.2 Gli eventi

Un evento è un sottoinsieme dello spazio campionario. Ad esempio

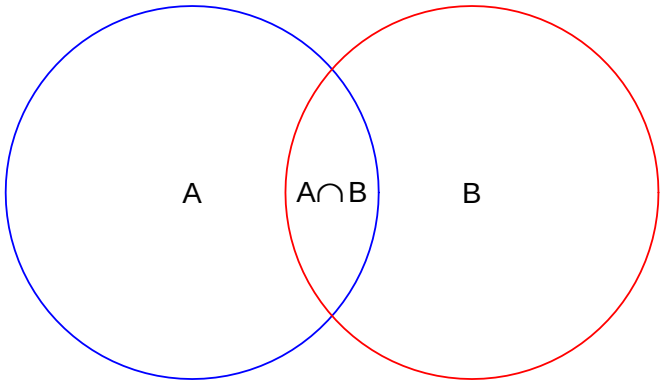
- Nel lancio di un dado l'uscita del 6 o l'uscita di un numero pari sono possibili eventi.
- Un evento semplice è un sottoinsieme di un elemento.

Costruzione di eventi

Siano A e B degli eventi che possono risultare da un esperimento. A partire da questi eventi possiamo costruire dei nuovi eventi

- $A \cup B$ (A oppure B, evento unione) indica il verificarsi di A o di B (o di ambedue).
- $A \cap B$ indica il verificarsi di A e di B.

- L'evento A^c (complemento di A, non A) indica il non verificarsi dell'evento A.



1.3 La probabilità

$P(A)$ indica la probabilità di un evento A. Questo è un numero nell'intervallo $[0, 1]$ che possiamo associare a ciascun evento che soddisfa certe regole (assiomi). Eventi certi corrispondono a probabilità uguale ad 1=100%, eventi impossibili corrispondono a fiducia uguale a 0=0%. La probabilità può essere definita per via assiomatica.

Assiomi

1. qualunque sia l'evento E, $P(E) \geq 0$
2. $P(\Omega) = 1$
3. $P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(E_1) + P(E_2) + \ldots + P(E_n)$ (anche per $n = \infty$) dove $E_i \cap E_j = \emptyset$

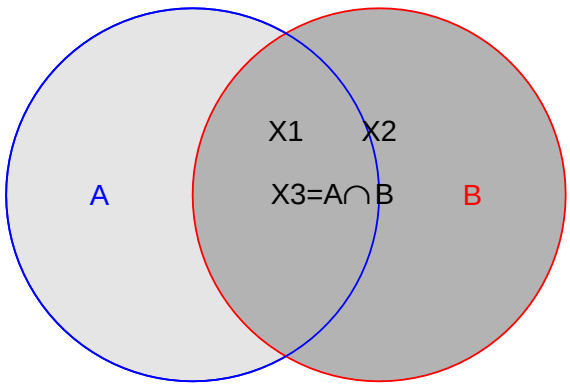
Consideriamo ad esempio il lancio di un dado e sia

- A = `esce pari`$=\{2, 4, 6\}$
- B = `esce un numero maggiore o uguale a 4`$=\{4, 5, 6\}$

Allora $A \cup B =$`esce 2 o 4 o 5 o 6`$=\{2, 4, 5, 6\}$
$A \cap B =$`esce 4 o 6`$= \{4, 6\}$
$A^C =$ `esce un numero dispari`$= \{1, 3, 5\}$
$B^C =$`esce 1 o 2 o 3`$= \{1, 2, 3\}$

1.3.1 Conseguenze

A e B sono *incompatibili* se $A \cap B = \emptyset$. In particolare per eventi incompatibili $P(A \cup B) =$



$P(A)+P(B)$.

Con riferimento alla figura abbiamo quindi

$$A = X_1 \cup X_3, \quad B = X_2 \cup X_3$$

da cui

$$\begin{aligned} P(A) =& P(X_1) + P(X_3), \quad P(B) = P(X_3) + P(X_2) \\ P(X_1) =& P(A) - P(X_3), \quad P(X_2) = P(B) - P(X_3) \end{aligned}$$

$$\begin{aligned} P(A \cup B) &= P(X_1 \cup X_2 \cup X_3) = P(X_1) + P(X_2) + P(X_3) = \\ &= P(A) + P(B) - P(X_3) = P(A) + P(B) - P(A \cap B) \end{aligned}$$

e quindi

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

ESEMPIO: Il 60% degli studenti in questa classe è geniale mentre il 70% ama lo sport; il 40%, oltre ad essere geniale, ama lo sport. Determinare la probabilità che uno studente scelto a caso non sia geniale e non ami lo sport.

1.3.2 Sistema completo di eventi

Diciamo sistema completo

$$A_1, A_2, \ldots, A_N$$

un insieme di eventi relativi ad un certo esperimento tali che in ogni realizzazione dell'esperimento si verifichi uno e uno solo di essi.

In generale sono un sistema completo di eventi si ha

$$P(A_1) + P(A_2) + \ldots + P(A_N) = 1$$

- Lanciando una moneta possiamo prendere come sistema completo di eventi gli eventi A_1="esce testa" e A_2="esce croce"
- lanciando un dado possiamo prendere come sistema completo di eventi le uscita A_i dei numeri i da 1 a 6, ma anche gli eventi "esce pari", "esce dispari"

Se tutti gli eventi in un sistema completo di eventi sono equiprobabili possiamo facilmente ricavare la probabilità di ciascun evento.

$$\begin{gathered} P(A_1) + P(A_2) + \ldots + P(A_N) = 1 = P(A_i) + P(A_i) + \ldots + P(A_i) \\ P(A_i) = 1/N \end{gathered}$$

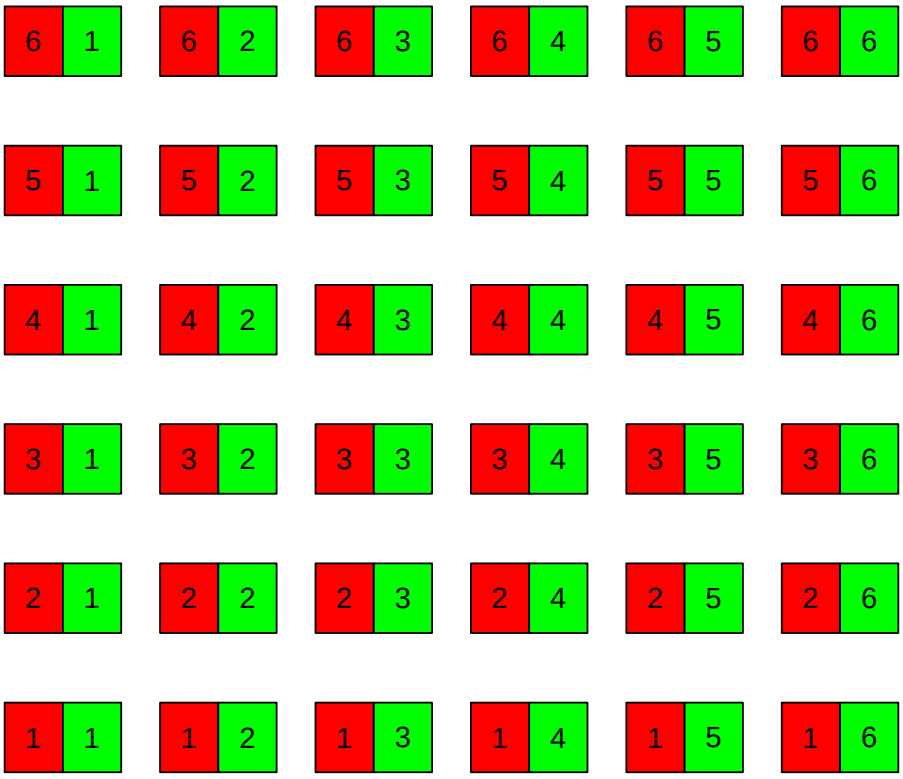
Nel caso del dado equo

$$P(\text{esce i}) = 1/6$$

1.4 Lancio di 2 dadi

Immaginimamo di lanciare simultaneamente 2 dadi a 6 facce. Per semplicità assumiamo che uno dei due dadi sia rosso e l'altro sia verde. Un sistema completo di eventi i 36 eventi é

$$A_{i,j} = \text{esce i sul dado verde e j sul dado rosso}$$



Se i dadi sono equi ogni evento $A_{i,j}$ ha probabilità 1/36. Per esempio

$$P(\text{la somma fa } 8) = 5/36$$

1.5 Probabilità condizionata

La notazione

$$P(A \mid B)$$

indica la probabilità dell'evento A condizionata al verificarsi dell'evento B. In altre parole quale é la probabilità che si verifichi A quando anche B é verificato.

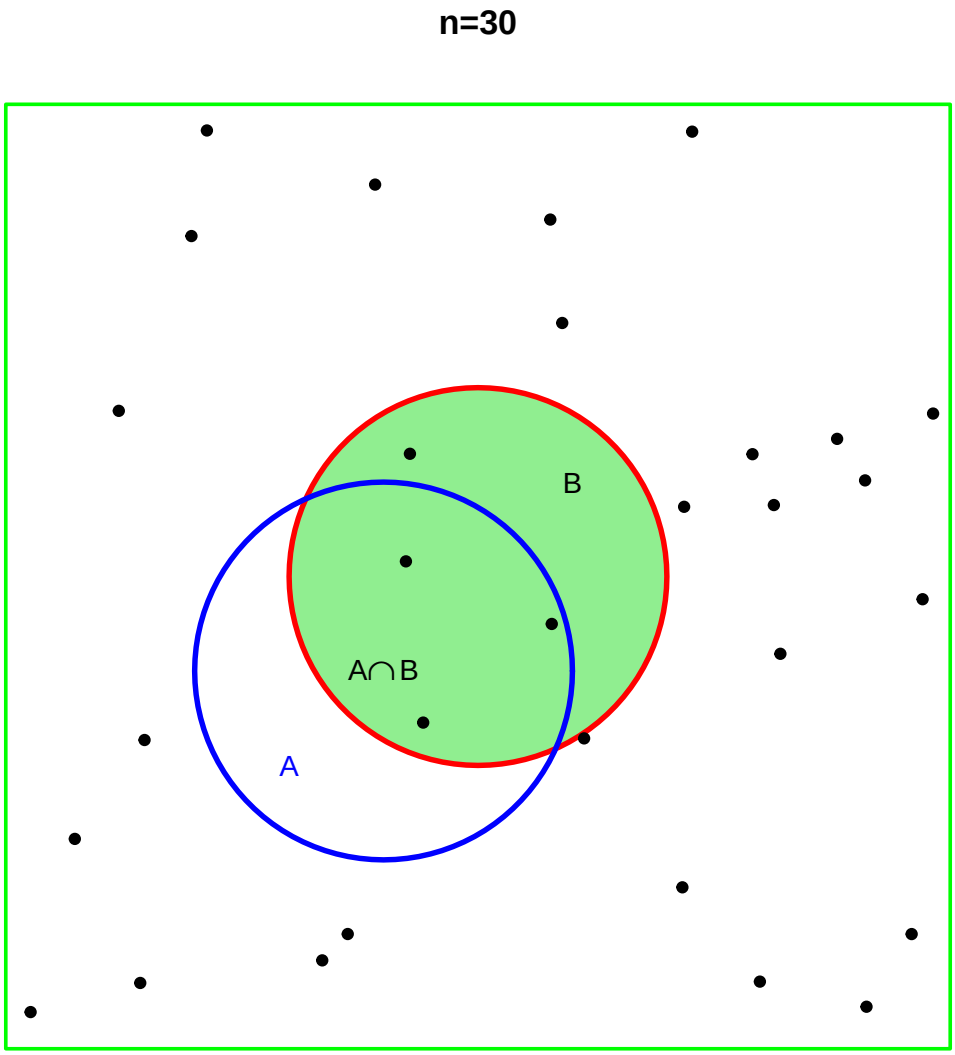
```
##
## Attaching package:  'MASS'
## The following object is masked from 'package:EsamiR':
##
##    crabs
```

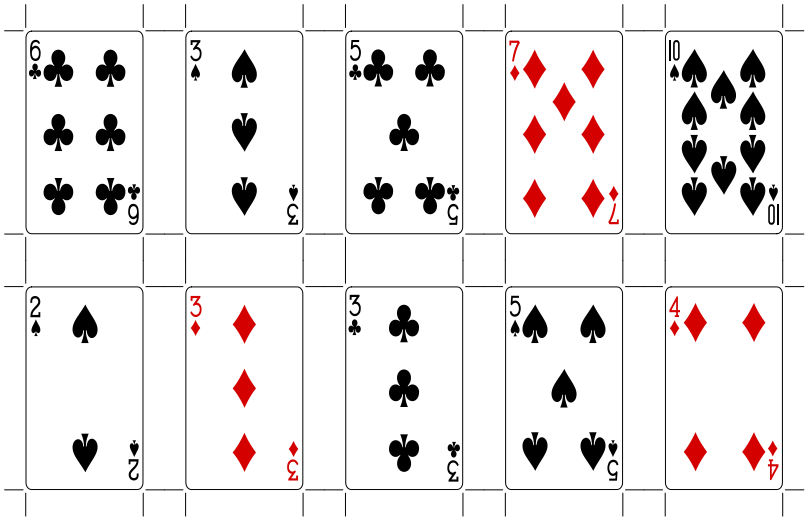
$$P(A|B) = \frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}/n}{n_B/n} = \frac{P(A \cap B)}{P(B)} = 3/4$$

Quindi

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (\text{se } B \neq \emptyset)$$

1.6 Esempio 1


n=30
B
A∩B
A



$P(\text{Rosso}) = 3/10$

$P(\text{Pari}) = 4/10$

$P(\text{Rosso} \cap \text{Pari})$

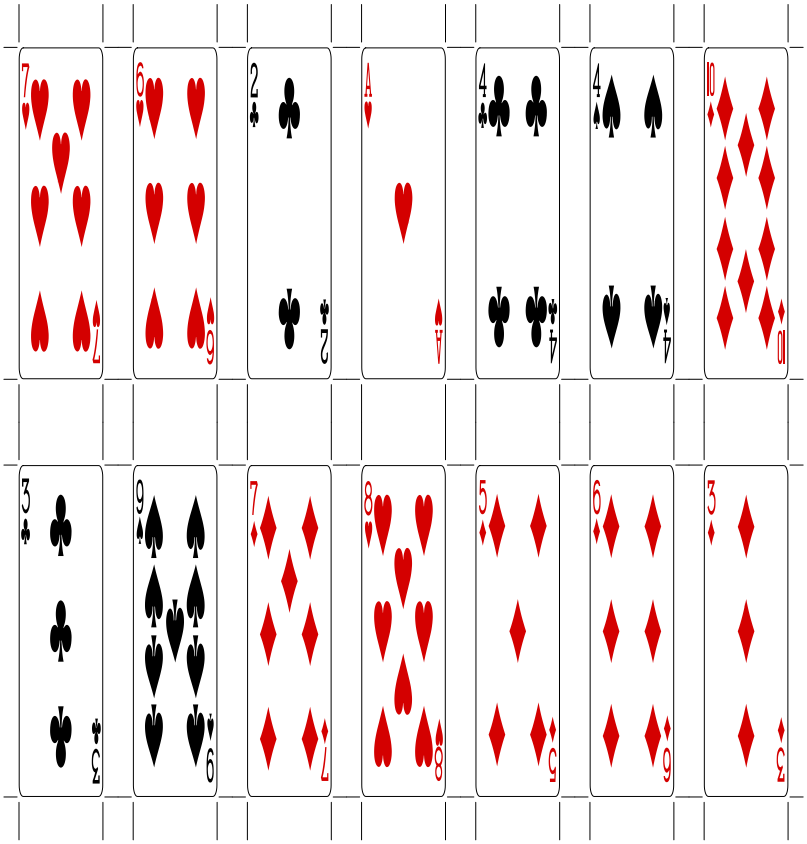
$>$ - $1/10$

$P(\text{Rosso} \mid \text{Pari}) = 1/4$

$P(\text{Pari} \mid \text{Rosso}) = 1/3$

$P(\text{Pari} \cup \text{Rossa}) = 6/10$

Esempio 2



$P(\text{Rosso})=9/14$
$P(\text{Pari}) = 7/14$
$P(\text{Rosso} \cap \text{Pari})= 4/14$
$P(\text{Rosso} \mid \text{Pari})=4/7$
$P(\text{Pari} \mid \text{Rosso}) =4/9$
$P(\text{Pari} \cup \text{Rossa}) = 12/14$

Tabelle di contingenza

Consideriamo una popolazione P di $N = 10^6$ individui uomini o donne (M/W) che possono essere mancini o destrorsi (L/R). Sia

$$WL = \text{Donne Mancine} = 60000$$

$$L = \text{Mancini totali} = 110000$$

$$W = \text{Donne} = 650000$$

	L	R	Totali
W	60000	-	650000
M	-	-	-
Totali	110000	-	1000000

Completando la tabella con i dati a nostra disposizione

	L	R	Totali
W	60000	590000	650000
M	50000	300000	350000
Totali	110000	890000	1000000

Per calcolare $P(L \mid W)$ occorre determinare la probabilità di essere mancini se si é donne. In altre parole la popolazione é quella delle donne e all'interno di quello determiniamo la probabilità di essere mancini:

$$P(L) = 110000/1000000 = 0.11$$

$$P(W) = 650000/1000000 = 0.65$$

$$P(L \cap W) = 60000/1000000$$

$$P(L \mid W) = (\#\text{Donne mancine})/(\#\text{Donne}) = 60000/650000 = 0.0923077$$

Così come

$$P(L \cap W)/P(W) = \frac{60000/1000000}{650000/1000000}$$

$$P(L \mid M) = 50000/350000 = 0.1428571$$

$$P(M \mid L) = 50000/110000 = 0.4545455$$

$$P(W \mid L) = 1 - 0.1428571 = 0.5454545$$

1.7 Regola di Bayes

Dalla definizione di probabilità condizionata

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

segue

$$P(A \cap B) = P(A \mid B)P(B)$$

e scambiando A con B

$$P(B \cap A) = P(A \mid B)P(B)$$

si ha quindi

$$P(A \mid B)P(B) = P(B|A)P(A)$$

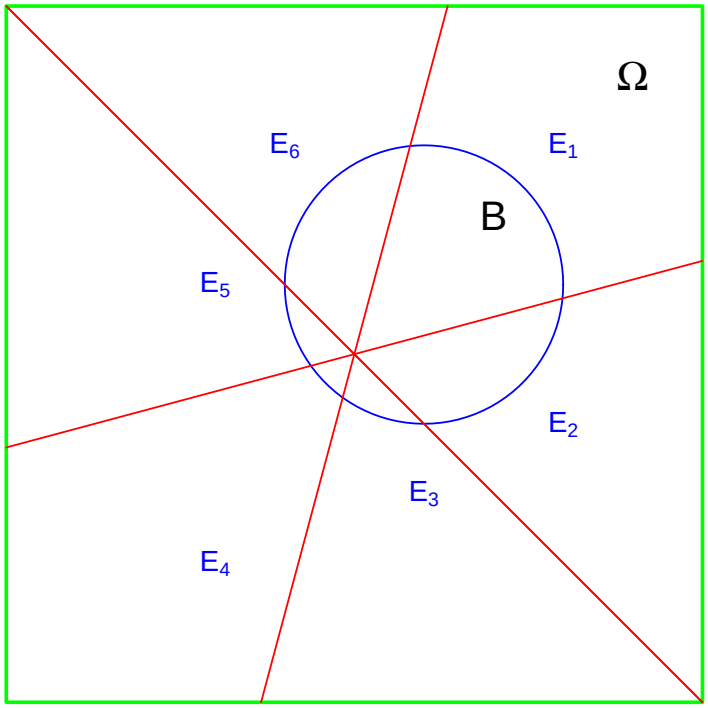
e
<ixmg src=”bayesrule.png” height=”200px” width=”300px” />

1.8 Legge della probabilità totale

Sia dato un sistema completo di esempi

$$E_1, \ldots, E_n$$

$$\Omega = E_1 \cup E_2 \cup \ldots \cup E_n \quad E_i \cap E_j = \emptyset$$



Ovviamente

$$B = (B \cap E_1) \cup (B \cap E_2) \cup \ldots \cup (B \cap E_N)$$

E quindi

$$P(B) = P(B \cap E_1) + P(B \cap E_2) + \ldots + P(B \cap E_N) =$$
$$P(B \mid E_1)P(E_1) + \ldots + P(B \mid E_N)P(E_N)$$

1.8.1 Esempio

$$B = \text{piove}$$

$$E_1 = \text{vado a lezione}$$

$$E_2 = \text{non vado a lezione}$$

$$P(\text{piove}) = P(\text{piove e vado a lezione}) + P(\text{piove e non vado a lezione})$$

1.9 Teorema di Bayes [forma estesa]

Combinando con la regola di Bayes:

$$P(E_i \mid B) = \frac{P(B \mid E_i)P(E_i)}{P(B)} = \frac{P(B \mid E_i)P(E_i)}{\sum_{i=1}^{N} P(B \mid E_i)P(E_i)}$$

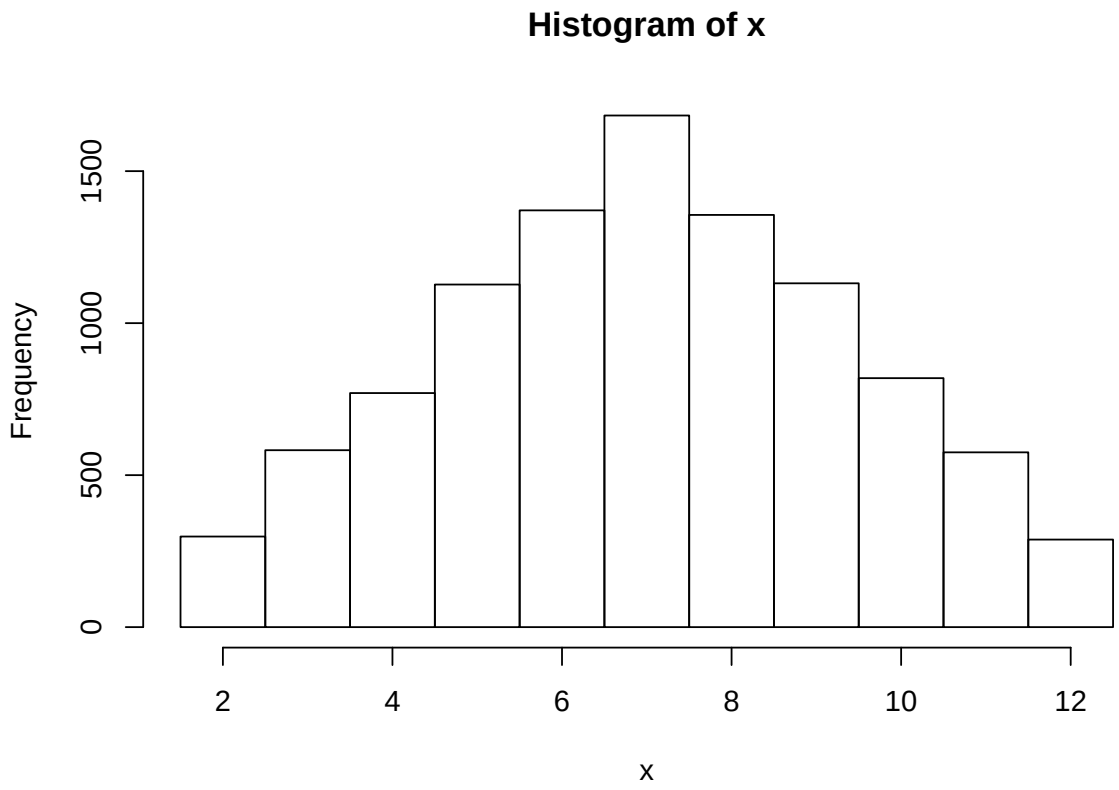
Nel caso particolare in cui gli eventi possibili E_1 e E_2 siano A e il suo complemento A^C si ha

$$P(A|B) = \frac{P[B \mid A]P(A)}{P(B \mid A)P(A) + P(B \mid A^C)P(A^C)}$$

1.10 La concezione frequentista

1.10.1 Lancio di una coppia di dadi

```
## x
##    2    3    4    5    6    7    8    9   10   11   12
##  298  582  770 1127 1371 1683 1356 1131  819  575  288
```




```
y

## x
##    2    3    4    5    6    7    8    9   10   11   12
##  298  582  770 1127 1371 1683 1356 1131  819  575  288
```

Il numero 8 è uscito in questo caso k=1356 volte su N= 10000 lanci. Possiamo dire che il risultato esce di norma k volte su 10000 e scrivere

$$\textit{Probabilità che esca } 8 = \lim_{N\to\infty} \frac{k}{N}$$

Il valore stimato con 10000 lanci è

$$\text{probabilità che esca } 8 \approx 1356/10000 = 0.1356$$

Vedremo in seguito che

$$\text{Probabilità che esca } 8 = 5/36 = 0.1389$$

e quindi che la differenza relativa della nostra stima dal valor vero è di

$$\frac{0.1356 - 0.1389}{0.1389} = -0.0237581$$

ovvero la discrepanza relativa è circa del 2.4

Stimiamo la probabilità di un evento A come la frequenza relativa dell'evento

$$P(A) \approx \frac{\text{numero di volte in cui l'evento si realizza}}{\text{numero di esperimenti eseguiti}}$$

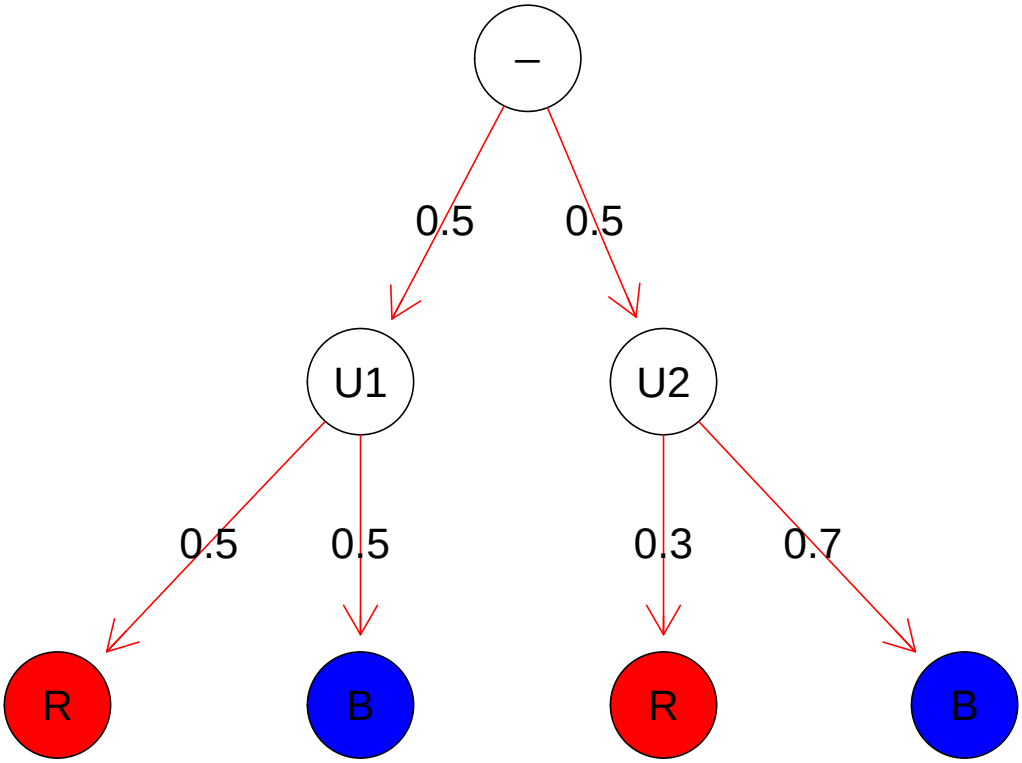
Questa concezione "frequentista" della probabilità è sufficiente in tutti i casi in cui l'esperimento può essere ripetuto (almeno a livello concettuale) quante volte si vuole. Non è però sufficiente per i casi in cui l'esperimento non può essere ripetuto o forse nemmeno eseguito.

In tali casi si ricorre alla concezione *soggettiva* della probabilità, in cui la probabilità viene stimata da un soggetto sulla base dell'esperienza personale.

1.10.2 Esercizio 1

Due urne contengono palle colorate. La prima urna contiene 50 palle rosse e 50 palle blu. La seconda contiene 30 palle rosse e 70 blu. Si sceglie una delle due urne a caso e si estrae da questa una palla a caso. La palla estratta è rossa. Quale è la probabilità che la palla provenga della prima urna?

1.10.3 Albero di probabilità



Motliplicando i valori delle frecce che puntano ad una foglia si trova

$$P(R \mid U1)P(U_1) = 0.5 \times 0.5 = P(R \cap U1)$$

$$P(R \mid U2)P(U_2) = 0.5 \times 0.3 = P(R \cap U2)$$

e quindi per esempio

$$P(R) = P(R \cap U1) + P(R \cap U2) = 0.4$$

Usando poi la regola di Bayes

$$P(U_1 \mid R) = \frac{P(R \mid U_1)P(U_1)}{P(R)} = 0.25/0.4 = 5/8$$

1.10.4 Esempio 2

Consideriamo dei container per la raccolta di materiale tossico. I container potrebbero non avere una tenuta perfetta ed avere delle perdite. Un programma di monitoraggio controlla regolarmente se ci sono state perdite. Gli strumenti dedicati a tali controlli non sono perfetti e talora danno luogo a falsi positivi o falsi negativi. In altre parole a volte viene

emesso un segnale d'allarme quando non si ha perdita (falso positivo) a volte la perdita non viene segnalata quando c'è (falso negativo).

Assumiamo di avere 3 informazioni

Informazioni sui container

1. P (perdita)= 0.1 P (nessuna perdita)=0.9

Per quanto riguarda il test

2. P(test positivo | perdita)=P (analisi corretta se si ha perdita)=0.95

P(test negativo | perdita)=P(test negativo dato che si ha perdita) = 0.05 (falsi negativi)

3. P(test positivo | non perdita)=P(test positivo quando non si ha perdita) = 0.1 (falso positivo)

P(test negativo| non perdita)=P(non rilevamento quando non si ha perdita)=0.9

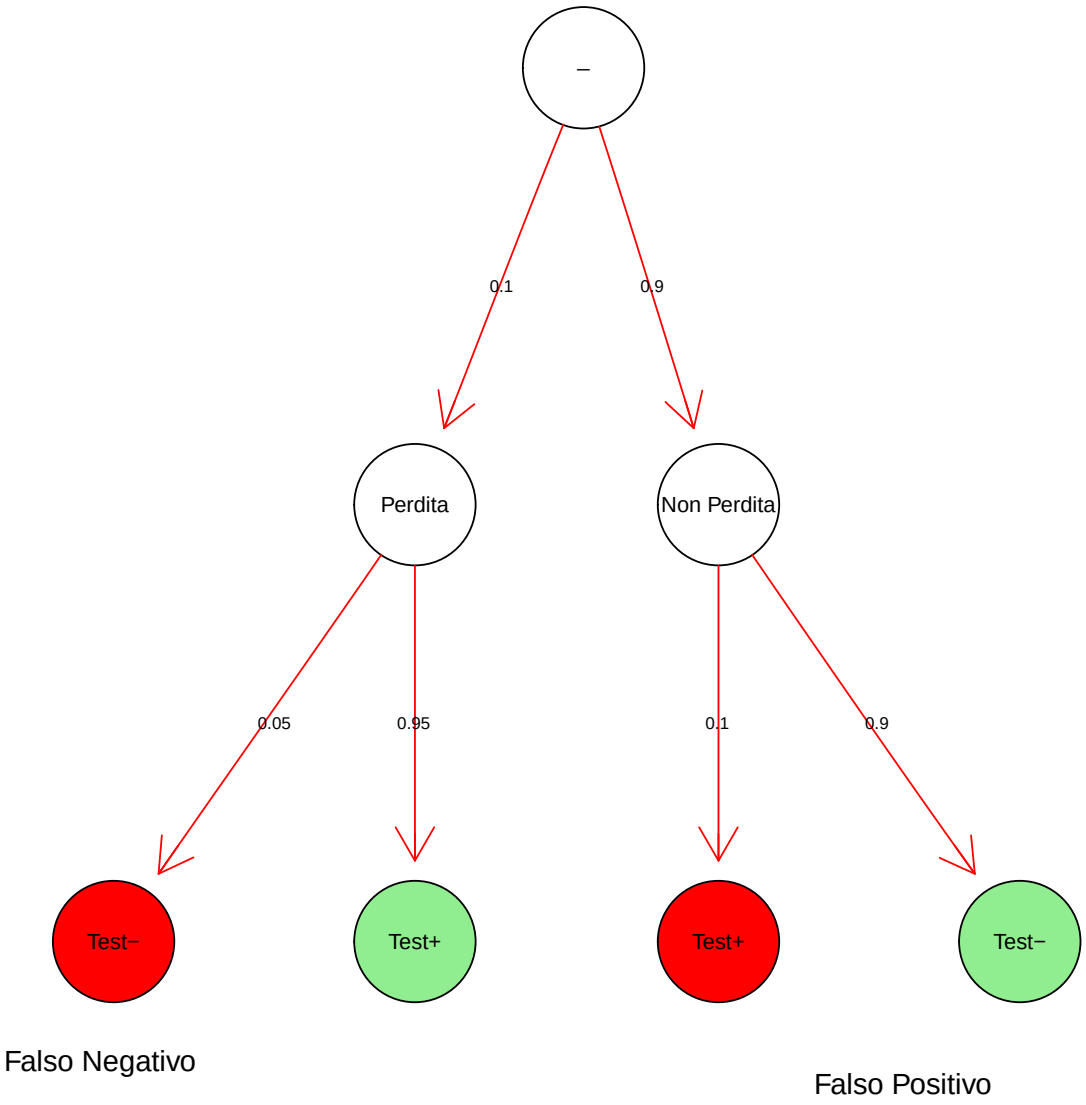
Vorremmo rispondere alle seguenti 2 domande.

1. Se l'allarme scatta quale è la probabilità che ci sia stata effettivamente una perdita?

2. Se l'allarme non scatta quale è la probabilità che il sito sia invece contaminato?

Costruiamo un diagramma ad albero come prima

—



$$P(\text{positivo}) = 0.1 \times 0.95 + 0.1 \times 0.9 = 0.185$$

e utilizzando la regola di Bayes

$$P(\text{perdita} \mid \text{positivo}) = \frac{P(\text{positivo} \mid \text{perdita})P(\text{perdita})}{P(\text{positivo})} = 0.095/0.185 = 0.5135135$$

$$P(\text{perdita} \mid \text{negativo}) = \frac{P(\text{negativo} \mid \text{perdita})P(\text{perdita})}{P(\text{negativo})} = 0.005/0.815 = 0.006135$$

1.10.5 Esempio 3: specificità e sensitività di un test diagnostico

Sensitività

"proporzione dei positivi identificati tra i malati: indica la capacità di individuare malati"

$$P(+ \mid D) = P(\text{test positivo} \mid \text{malato})$$

Supponiamo sia

$$P(\text{test positivo} \mid \text{malato}) = 0.97$$

$$P(\text{test negativo} \mid \text{malato}) = 0.03 \text{ (Falsi Negativi)}$$

Specificità

"proporzione dei negativi identificati: indica la capacità di individuare i sani"

$$P(- \mid D^c) = P(\text{test negativo} \mid \text{sano})$$

Supponiamo sia

$$P(\text{test negativo} \mid \text{sano}) = 0.95$$

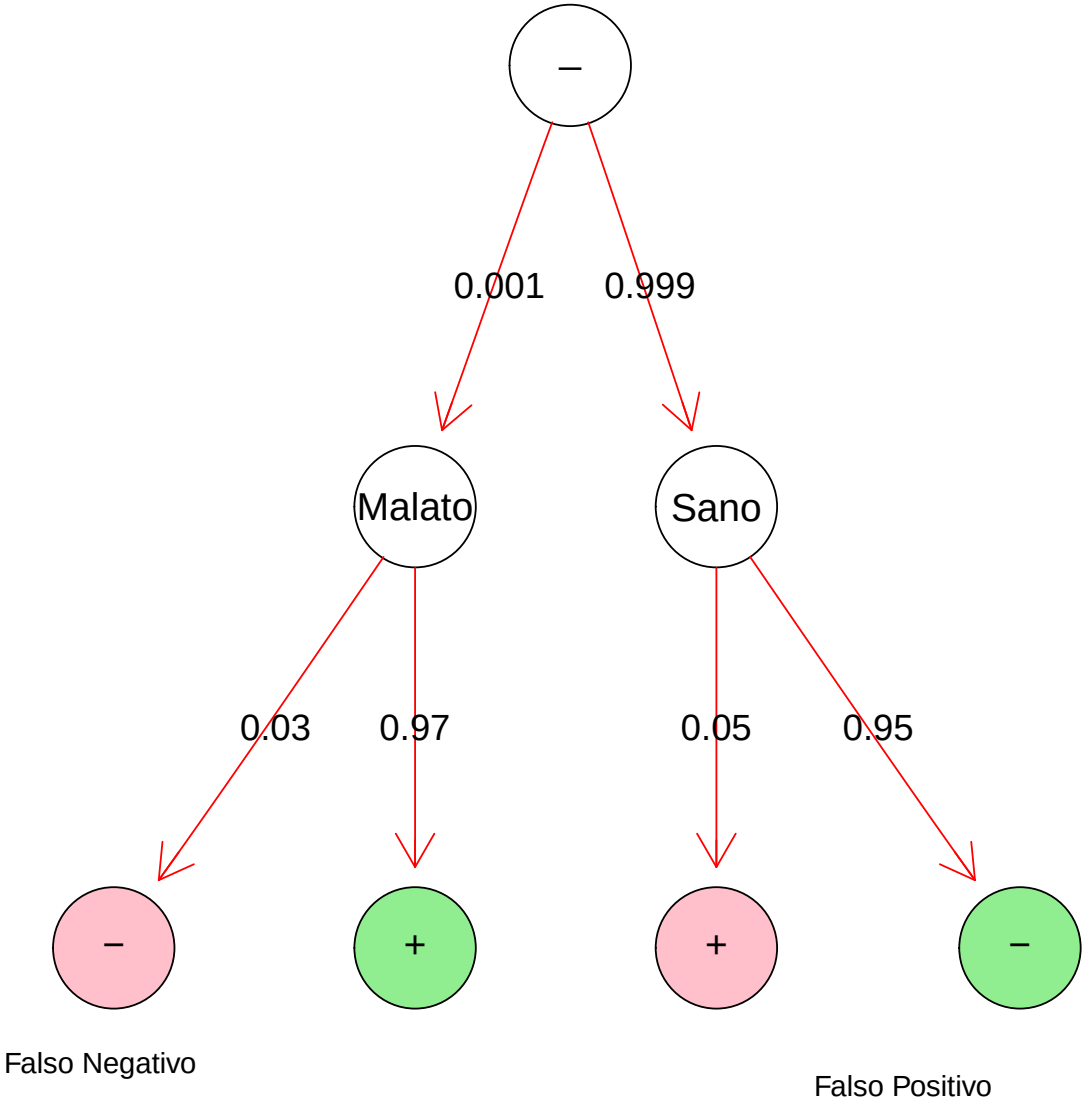
$$P(\text{test positivo} \mid \text{sano}) = 0.05 \quad \text{(Falsi positivi)}$$

* Supponiamo di essere testati per una malattia che colpisce lo 0.1% della popolazione.
* Determinare probabilità di non avere la malattia avendo avuto un esito positivo del test.
In formule

$$P(\text{sani} \mid \text{test positivo}) = ?$$

$$P(\text{malati} \mid \text{test negativo}) = ?$$

1.10.6 Albero sensitività/specificità

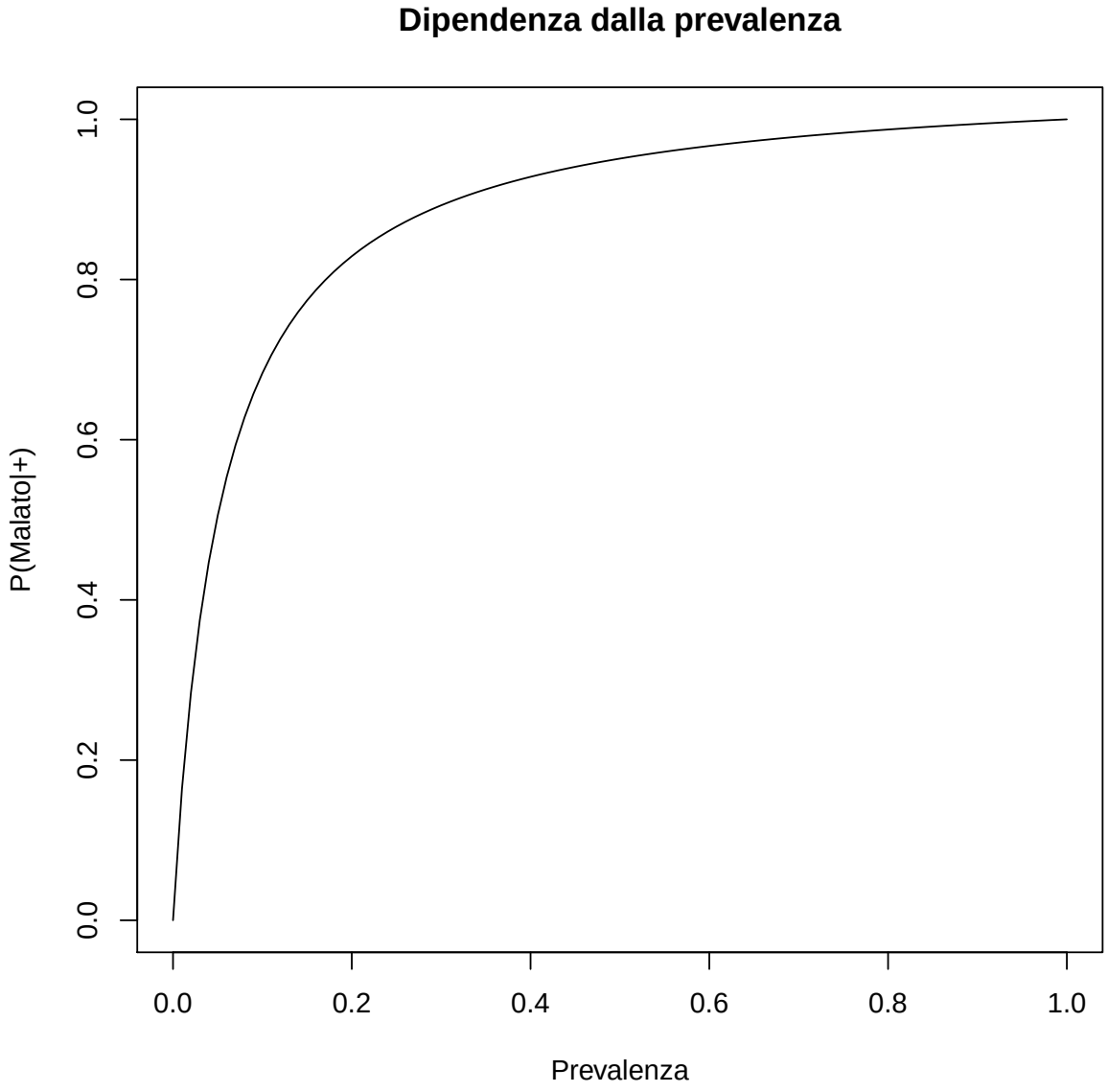


$$P(+) = P(+ \mid \text{sano})P(\text{sano}) + P(+ \mid \text{malato})P(\text{malato})$$
$$= 0.05 \times 0.999 + 0.97 \times 0.001 = 0.05092$$

$$P(\text{sano} \mid +) = \frac{P(+ \mid \text{sano})P(\text{sano})}{P(+)} = \frac{0.05 * 0.999}{0.05092} = 0.9809505$$

In genere

$$P(+) = \text{(1-specificità)(1-prevalenza)} + \text{sensitività prevalenza}$$



1.10.7 Approccio frequentista

Immaginando 10^6 soggetti

```
##         Test - Test +  Totali
## Affetti             -
## Sani         -      -       -
## Totali              - 1000000
## Usando la prevalenza
##         Test - Test +  Totali
```

```
## Affetti      -      -    1000
## Sani         -      -  999000
## Totali       -      - 1000000
## Usando la sensitivita'
##         Test - Test +  Totali
## Affetti     30    970    1000
## Sani                   999000
## Totali                1000000
## Usando la specificita'
##         Test - Test +  Totali
## Affetti     30    970    1000
## Sani    949050  49950  999000
## Totali                1000000
## Calcolando i totali
##         Test - Test +  Totali
## Affetti     30    970    1000
## Sani    949050  49950  999000
## Totali  949080  50920 1000000
```

$$P(\text{affetti} \mid \text{test negativo}) = 30/949080 = 0.0000316$$

Terminologia

Il *valore predittivo positivo* è (D sta per disease)

$$P(D \mid +)$$

Il *valore predittivo negativo* è

$$P(D^c \mid -)$$

probabilità di non avere la malattia assunto un esito negativo del test.
La *prevalenza* di una malattia è

$$P(D)$$

Il *rapporto di verosimiglianza diagnostico di un test positivo* DLR_+ è

$$\frac{P(+ \mid D)}{P(+ \mid D^c)} = \frac{\text{sensitivita}}{1 - \text{specificità}}$$

Il *rapporto di verosimiglianza diagnostico di un test negativo* DLR_- è

$$\frac{P(- \mid D)}{P(- \mid D^c)} = \frac{1 - \text{sensitivita}}{\text{specificita}}$$

1.10.8 HIV

Uno studio sull'efficacia dei test HIV riporta un test con sensitività $P(+|\text{malato}) = 99.7\%$ e specificità $P(-|\text{sano}) = 98.5\%$. Supponiamo che la prevalenza dell'HIV sia dello 0.1%. Determinare la probabilità $P(D|+)$, che un soggetto (positivo al test) abbia l'HIV.

Calcolare inoltre $P(D^c|-)$ e $P(D)$

$$P(+) = P(+|D)P(D) + P(+|S)P(S) = 0.997 \times 0.001 + 0.015 \times 0.999 = 0.015982$$

$$P(D|+) = P(+|D)P(D)/P(+) = 0.997 \times 0.001/0.015982 = 0.0623827$$

Test di gravidanza

Un sito web `http://www.medicine.ox.ac.uk/bandolier/band64/b64-7.html` che esamina studi medici considera un test di gravidanza che afferma:

"Quando i soggetti che hanno effettuato i test erano le donne che hanno raccolto e analizzato i loro campioni la sensitività era del 75%. La specificità era anche bassa, nel range dal 52% al 75%."

Consideriamo il valore inferiore della specificità. Supponiamo che il risultato sia negativo e che il 30% delle donne che fanno il test siano effettivamente pregnanti. Determinare la probabilità di essere incinte dato il risultato negativo del test?

$$P(\text{Incinta } | -) = \frac{P(- \mid \text{Incinta})P(\text{Incinta})}{P(-)} =$$

$$\frac{P(- \mid \text{Incinta})P(\text{incinta})}{(P(- \mid \text{Incinta})P(\text{Incinta}) + P(- \mid \text{Non Incinta})P(\text{Non Incinta})} =$$

$$0.250 * 0.3/(0.250 * 0.3 + 0.52 * 0.7) = 0.1708428$$

Fallacia del Pubblico Ministero

Supponiamo che in una comunità di 10 abitanti sia stato commesso un delitto. Il colpevole è un abitante della comunità (immediatamente isolata).

Immaginiamo che da un test del sangue si trovi che un abitante individuato in modo casuale (il sospettato) e il colpevole condividano una caratteristica comune allo 5

Il pubblico ministero afferma che la probabilità di essere innocente del sospettato è dello 5

In realtà (indicando con G la colpevolezza e I l'innocenza)

$$P(\text{G} \mid \text{test}) = P(\text{test} \mid \text{ G})P(\text{G})/P(\text{test})$$

$$P(\text{test} \mid \text{G}) = 1 \quad P(\text{G}) = 1/10$$

$$P(\text{test}) = P(\text{test} \mid \text{G})P(\text{G}) + P(\text{test} \mid \text{I})P(\text{I}) = 1 * 1/10 + 0.05 * 9/10 = 0.145$$

Quindi

$$P(\text{G} \mid \text{test}) = 0.1/0.145 = 0.6896552$$

Odds

Si definiscono le odds di un evento

$$\text{Odds}(E) = \frac{P(E)}{P(\text{not E})} = \frac{P(E)}{1 - P(E)}$$

Il logaritmo naturale delle odds è detto logit. Il rapporto tra due odds è detto odds ratio.

Il problema delle 3 porte (Monty Hall problem)

Ci sono 3 porte. Dietro a 2 delle 3 porte una capra. Dietro alla restante una Ferrari. Il giocatore sceglie una porta (ma non la fa aprire). Il conduttore ne apre una delle altre 2 mostrando una capra e chiede al giocatore se vuole cambiare la sua scelta originale. Conviene attenersi alla scelta originale o conviene cambiare?

Supponiamo per concretezza che che si sia deciso di aprire la porta 1 e che il presentatore apra la 3. Poniamo B="il presentatore apre la 3". Conviene scegliere al porta 2 (e cambiare la nostra scelta iniziale) o restare ostinati e scegliere la porta 1?

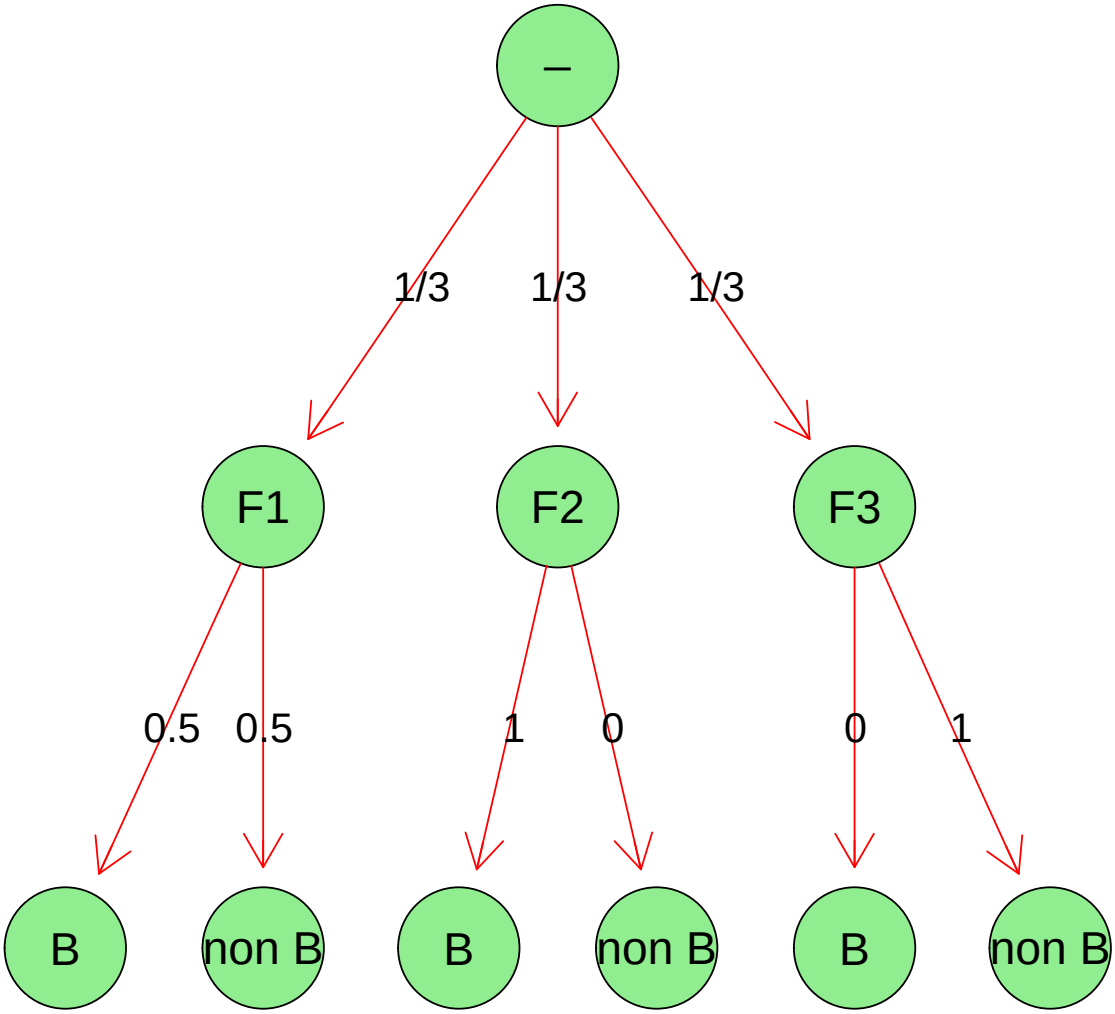
F_1 è l'evento "la Ferrari è dietro alla porta 1", F_2 e F_3 sono definiti in modo uguale con le porte 2 e 3 rispettivamente. Per esempio, abbiamo scelto la porta 1

$$P(F_1) = P(F_2) = P(F_3) = 1/3$$
$$P(B) = 1/2$$
$$P(B \mid F_1) = 1/2$$
$$P(B \mid F_2) = 1$$
$$P(B \mid F_3) = 0$$

Possiamo quindi calcolare

$$P(F_1 \mid B) = \frac{P(B \mid F_1)P(F_1)}{P(B)} = \frac{1/2 * 1/3}{1/2} = 1/3$$
$$P(F_2 \mid B) = \frac{P(B \mid F_2)P(F_2)}{P(B)} = \frac{1 * 1/3}{1/2} = 2/3$$
$$P(F_3 \mid B) = \frac{P(B \mid F_3)P(F_3)}{P(B)} = 0$$

Grafico



Recessione

Una agenzia economica ha creato un modello che predice recessione. Il modello predice recessione con probabilità del 80% quando la recessione sta effettivamente arrivando e con probabilità del 10% quando la recessione non sta arrivando. La probabilità che la recessione sia in arrivo è del 20%. Se il modello predice recessione quale è la probabilità che la recessione stia effettivamente arrivando?

P(Prevista) = P(Prevista | Recessione)P(Recessione) + P(Prevista | No recessione) P(No Recessione) =$0.8 \times 0.2 + 0.1 \times 0.8$=0.24

P(Recessione | Prevista) = P(Prevista | Recessione)*P(Recessione)/P(Prevista)=$0.8 \times 0.2/0.24$=0.6666667

1.10.9 Esercizio

Alice ha 2 monete nella borsa. La prima con 2 facce diverse e la seconda con 2 teste. Alice preleva a caso una moneta dalla borsa. La lancia e vede testa. Quale è la probabilità che abbia lanciato la moneta equa?

1.11 Variabili aleatorie

Una variabile aleatoria (*random variabile*) è una variabile i cui valori sono soggetti a variazioni casuali. Quando i valori possibili di una variabile aleatoria possono essere elencati parliamo di variabile aleatoria discreta. Quando i valori non possono essere elencati parliamo di variabile aleatoria continua.

1.12 Variabili aleatorie discrete

Le variabili aleatorie discrete che assumono un numero limitato di valori si dicono anche *finite*. I valori di una variabile aleatoria discreta possono essere numerici o nominali. Supponiamo di avere una variabile aleatoria che possa assumere un insieme di valori in un *alfabeto* assegnato costituito da lettere, parole o numeri. Per esempio un alfabeto può essere del tipo che segue

- (Femmina, Maschio)
- (A,C,T,G)
- (0,1)
- (Ottimo, Buono, Discreto, Sufficiente, Insufficiente)
- (Testa, Croce).
- I numeri interi

Per caratterizzare completamente una variabile aleatoria discreta oltre ai valori che questa può assumere occorre conoscere la probabilità di questi valori. Per semplicità considereremo variabili aleatorie finite.

La probabilità $P(A)$ di un evento A è il grado di fiducia che lo sperimentatore pone nella realizzazione dell'evento (Jacob Bernoulli, 1654-1705)

1.13 Stima dei parametri statistici

1.13.1 Covarianza

In genere ae X e Y sono variabili aleatorie allora

$$\mu_{aX+bY} = E[aX + bY] = aE[X] + bE[Y]$$

mentre la relazione che segue vale se X e Y sono scorrelate ossia $\mathrm{Cov}(X, Y) = 0$ (o se sono indipendenti)

$$\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y]$$

dove la covarianza di una coppia di variabili aleatorie è definita come

$$\mathrm{Cov}[X, Y] = E\big[(X - E[X])(Y - E[Y])\big]$$

e nel caso discreto si può scrivere

$$\mathrm{Cov}[X, Y] = \sum_{i,j}(x_i - \mu_X)(y_j - \mu_Y)p_{X=x_i,Y=y_j} \quad \text{caso discreto}$$

mentre nel caso continuo servirebbe un integrale doppio. per esempio la formula precedente per $a = 1$ e $b = 1$ diviene

$$\mathrm{Var}[X+Y] = E[(X+Y-E[XbY])^2] = E[(X-E[X]+Y-E[Y])^2] = \mathrm{Var}[X]-2\mathrm{Cov}[X, Y]+\mathrm{Var}[Y]$$

1.13.2 Stima di mu_X

Come stimare il valore di μ_X eseguendo n misure di X?

Il problema più frequente nelle applicazioni statistiche è di dare una stima dei valori di parametri statistici di una popolazione una volta che siano noti i valori dei parametri di un campione.

Stima di Media e Varianze

Consideriamo una serie di n esperimenti in cui si misura una grandezza X che ha valore atteso $\mu_X = E[X]$ e varianza $\sigma_X^2 = \mathrm{Var}[X]$. Il risultato atteso nell'esperimento numero i è indicato con X_i (anche se gli esperimenti sono tutti uguali). Gli n esperimenti che eseguiremo costituiscono un **campione aleatorio** di dimensione n.

Possiamo inoltre considerare le due espressioni

$$M_n[X] = \frac{\sum_{i=1}^n X_i}{n} \quad \text{media campionaria aleatoria}$$

$$\mathrm{S}_n^2[X] = \frac{\sum_{i=1}^n (X_i - M_n[X])^2}{n-1} \quad \text{varianza campionaria aleatoria}$$

che dipendono dal campione aleatorio X_i e come tali sono esempi di 'consuntivi aleatori' (*statistiche campionarie*).

Per osservare **un** valore di M_n e di S_n^2 occorre eseguire n misure di X.

Indichiamo con le corrispondenti lettere minuscole i numeri in cui si traducono le variabili aleatorie ad esperimento eseguito. Il risultato dell'esperimento i-esimo è x_i mentre media e varianza del campione sono

$$m_n = \frac{\sum_{i=1}^n x_i}{n} \quad \text{media osservata}$$

$$s_n^2 = \frac{\sum_{i=1}^n (x_i - m_n)^2}{n-1} \text{varianza osservata}$$

1.13.3 Stimatori

Il valore osservato m_n della media campionaria viene usato per stimare μ. La media campionaria M_n è dunque uno **stimatore** della media vera μ.

Se uno stimatore arbitrario T_n viene usato per stimare il parametro incognito τ, sono significative le due proprietà che seguono:

1. Lo stimatore T_n è corretto ('unbiased') se il suo valore atteso coincide con il valore del parametro che si vuole stimare:

$$E[T_n] = \tau$$

2. Lo stimatore corretto T_n è anche coerente, o consistente ('consistent'), se la sua varianza tende a zero quando il numero degli esperimenti cresce:

$$\lim_{n\to\infty} \mathrm{Var}[T_n] = 0.$$

Media campionaria e varianza campionaria (effettuando gli esperimenti) ci forniscono buone stime per la media e la varianza della popolazione (μ_X e $\sigma^2[X]$). Infatti si può mostrare che

* Proprietà 1

M_n e S_N^2 sono variabili aleatorie e hanno quindi valori attesi e varianze. Si ha

$$E[M_N[X]] = \mu_X$$

$$E[S_n^2[X]] = \sigma_X^2$$

per esempio

$$
\begin{aligned}
E[S_n^2[X]] = \frac{1}{N-1}E[(X_i - M_N)^2] = \frac{1}{N-1}E[(X_i - \mu + \mu - M_N)^2] \\
=&\frac{1}{N-1}\sum_i \left(E[(X_i-\mu)^2] + E[(\mu - M_N)^2] + 2E[(X_i-\mu)(\mu - M_n)]\right) = \\
=&\frac{1}{N-1}\left(\sum_i E[(X_i-\mu)^2] + \sum_i E[(\mu - M_N)^2] + 2\sum_i E[(X_i-\mu)(\mu - M_n)]\right) = \\
=&\frac{1}{N-1}\left(NVar[X] + NE[(M_N-\mu)^2] + 2NE[(M_n-\mu)(\mu-M_n)]\right) = \\
=&\frac{1}{N-1}\left(NVar[X] - NE[(M_n-\mu)^2]\right) = \\
=&\frac{1}{N-1}\left(NVar[X] - NVar[M_N]\right) = \frac{1}{N-1}\left(NVar[X] - Var[X]\right) = Var[X]
\end{aligned}
$$

* Proprietà 2
Se
$n \to \infty$

$$\mathrm{Var}[M_n[X]] \to 0$$
$$\mathrm{Var}[S_n^2[X]] \to 0$$

In altre parole se eseguiamo un numero molto elevato di esperimenti la media campionaria approssima μ_X e la varianza campionaria σ_X^2.

Quindi eseguire un numero elevato di esperimenti è un modo per stimare μ_X e σ_X.

Mostreremo le proprietà per la media campionaria

Valore atteso di M_n Visto che

$$E[aX + bY] = aE[X] + bE[Y]$$

$$E[M_n] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \frac{1}{n}\sum_{i=1}^{n} E[X] = E[X]$$

1.14 Varianza di M_n

Ricordiamo che se $\mathrm{Cov}(X,Y) = 0$

$$Var[aX + bY] = a^2Var[X] + b^2Var[Y]$$

In particolare se $X_1, \ldots, X_n$ sono repliche **indipendenti** dello stesso esperimento X

$$\mathrm{Var}[X_1 + \ldots + X_n] = n\mathrm{Var}[X]$$

e

$$\mathrm{Var}(M_n) = \sum_{i=1}^{n} \frac{\mathrm{Var}[X_i]}{n^2} = \frac{\mathrm{Var[X]}}{n} \Rightarrow \sigma^2_{M_n} = \sigma^2_X/n$$

Esercizio

Nello studio di un fenomeno aleatorio X ci chiediamo come stimare μ e σ: media e deviazione standard della popolazione. Per stimare il valore atteso μ ricorriamo alla media campionaria:

Consideriamo i seguenti dati ottenuti misurando le quantità di farmaco (in mg) in 5 fialette

$$(100, 100, 97, 99, 101$$

Stimare il valore atteso (media) della popolazione.

Soluzione

Per valutare la media della popolazione X consideriamo la variabile aleatoria M_N

$$M_n(X) = \frac{\sum_{i=1}^{n} X_i}{n}$$

La media del campione in esame è

$$\bar{x} = `rmean(fiale)` \text{ mg}$$

e quindi possiamo stimare il valore atteso per il contenuto delle fialette come

$$\mu \approx `rmean(fiale)` \text{ mg}$$

1.14.1 I criteri di attendibilità

Iniziamo con un esempio che ci convinca del fatto che le stime puntuali non sono le migliori possibili.

Esempio

Due extraterrestri sono appena sbarcati sulla Terra.

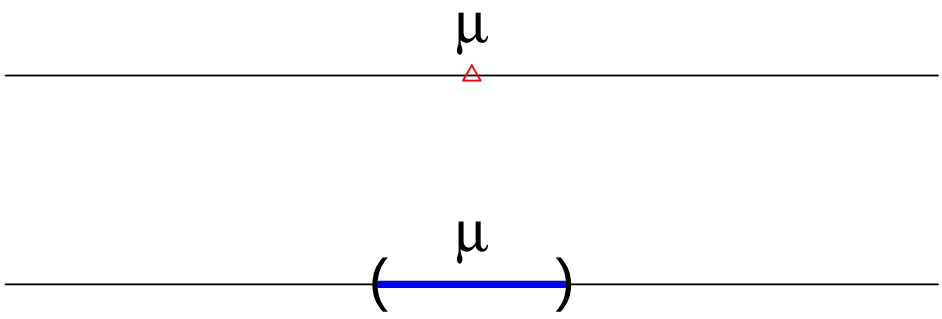
Non sapendo che altro fare lo statistico decide di misurarne l'altezza e di stimare il valore atteso dell'altezza X degli extraterrestri. Rileva una misura di 40 cm per il primo extraterrestre e di 200 cm per il secondo.

Calcola la media (purtroppo ha solo due extraterrestri a disposizione) e stima

$$\mu \approx \frac{40 + 200}{2} \text{ cm} = 120 \text{ cm}$$

La stima fatto dello statistico è *ottimale* (meglio di così non poteva fare) ma non attendibile (affidabile).

La stima puntuale (un punto sulla retta, un numero) risulta qui certamente priva di ogni significato. Introdurremo ora un altro tipo di stime: le stime intervallari in cui ha rilevanza l'attendibilità di un risultato.



1.14.2 Livello di fiducia

Per determinare un criterio di attendibilità occorre affidarsi ad un criterio soggettivo (lasciato allo sperimentatore). Se vogliamo stimare una grandezza occorre prima precisare cosa significa per noi essere pressoché certi di una conclusione.

Dobbiamo rinunciare alle certezze e affidarci alle probabilità che una cosa sia vera. Si fissa un numero ϵ tra 0 e 1

1. Si escludono (si ritengono impossibili) eventi con probabilità inferiore a ϵ (ϵ = *significatività *)

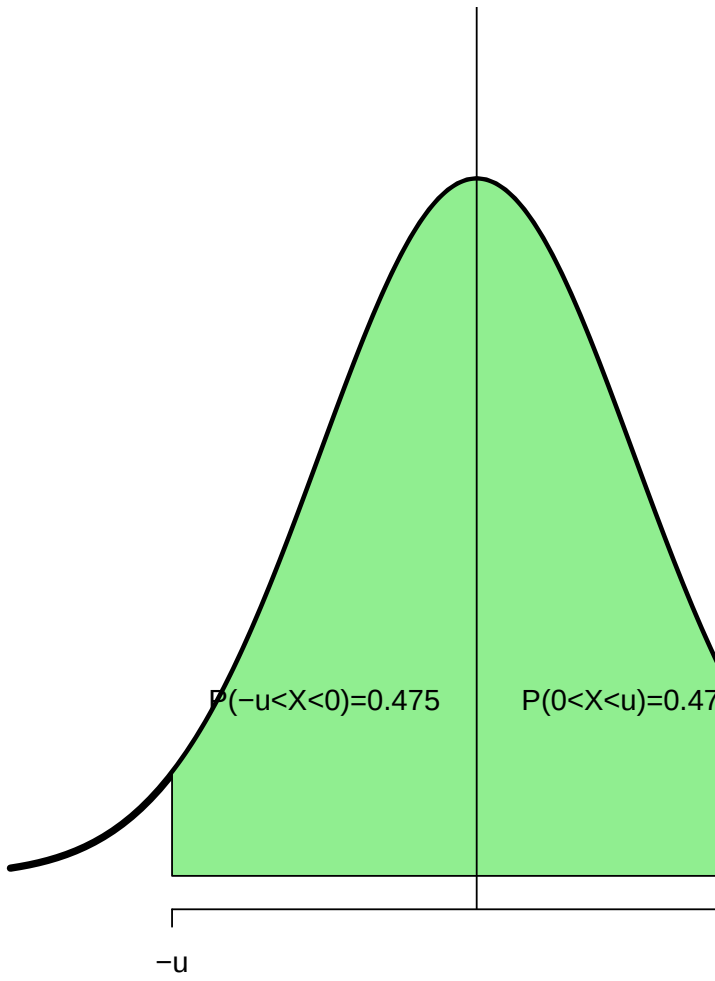
2. Si accettano (si ritengono certi) gli eventi con probabilità $\geq 1 - \epsilon$ ($1 - \epsilon$, *livello di fiducia* o confidence level).

Molto spesso nelle applicazioni sceglieremo $\epsilon = 0.05$ o 0.01. Fissato quindi il nostro criterio di attendibilità vedremo ad esempio come dare una stima intervallare della media vera di una **popolazione normale**, ossia come assegnare un intervallo entro il quale con la media della popolazione cada con una probabilità pari al livello di fiducia scelto.

1.15 La tabella inversa

Scegliamo un livello di fiducia, per esempio il 95

Possiamo allora chiederci per quale valore di u l'intervallo $[-u, u]$ racchiude una probabi-



lità pari al livello di fiducia scelto $(1-\epsilon)$.

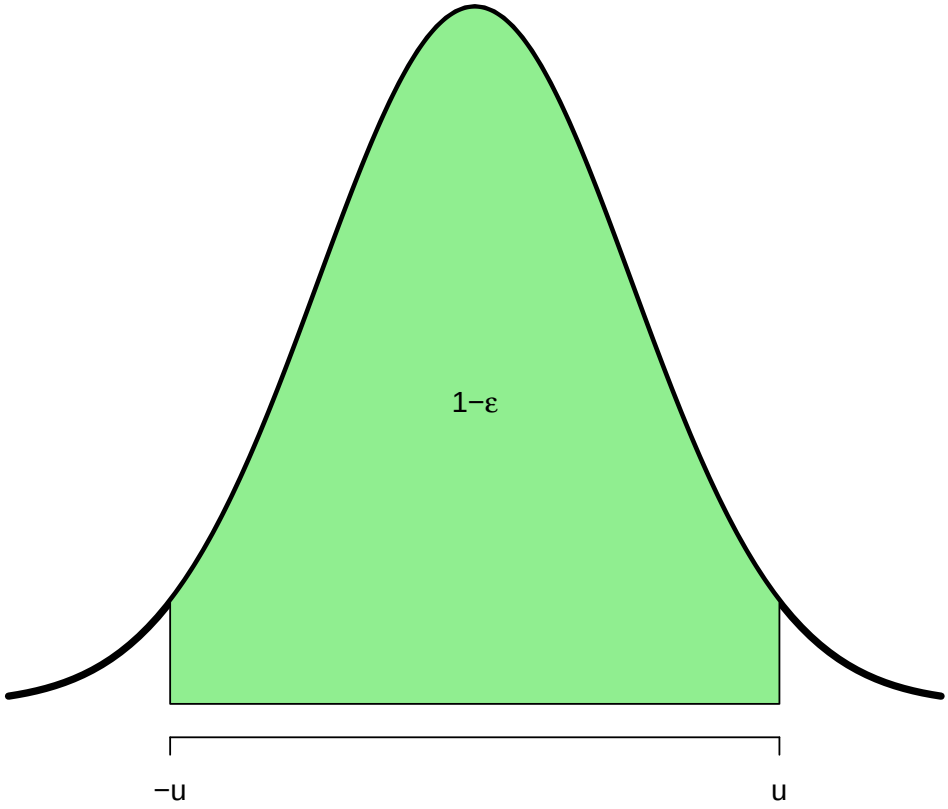
x Area x Area x Area — —— — —— — —— 0.1 0.0398 1.1 0.3643 2.1 0.4821
0.2 0.0793 1.2 0.3849 2.2 0.4861
0.3 0.1179 1.3 0.4032 2.3 0.4893
0.4 0.1554 1.4 0.4192 2.4 0.4918
0.5 0.1915 1.5 0.4332 2.5 0.4938
0.6 0.2257 1.6 0.4452 2.6 0.4953
0.7 0.258 1.7 0.4554 2.7 0.4965
0.8 0.2881 1.8 0.4641 2.8 0.4974
0.9 0.3159 1.9 0.4713 2.9 0.4981
1 0.3413 2 0.4772 3 0.4987

Guardando la tabella la soluzione non è immediata. Se per esempio scegliamo un livello di fiducia del 95E' quindi utile una tabella in cui, assegnato il livello di fiducia $1-\epsilon$ (l'area) si vuole conoscere il numero x tale che il grafico della normale da $-u$ ad u sottenda un'area pari al livello di fiducia $1-\epsilon$. In altri termini, nota l'area, si deve determinare u:

Tabella inversa

Considerando solo valori della probabilità prossimi a 1

1-epsilon	u	1-epsilon	u	1-epsilon	u
0.800	1.2816	0.870	1.5141	0.940	1.8808
0.805	1.2959	0.875	1.5341	0.945	1.9189
0.810	1.3106	0.880	1.5548	0.950	1.9600
0.815	1.3255	0.885	1.5761	0.955	2.0047
0.820	1.3408	0.890	1.5982	0.960	2.0537
0.825	1.3563	0.895	1.6211	0.965	2.1084
0.830	1.3722	0.900	1.6449	0.970	2.1701
0.835	1.3885	0.905	1.6696	0.975	2.2414
0.840	1.4051	0.910	1.6954	0.980	2.3263
0.845	1.4221	0.915	1.7224	0.985	2.4324
0.850	1.4395	0.920	1.7507	0.990	2.5758
0.855	1.4574	0.925	1.7805	0.995	2.8070
0.860	1.4758	0.930	1.8119	0.999	3.2905
0.865	1.4947	0.935	1.8453	1.000	Inf



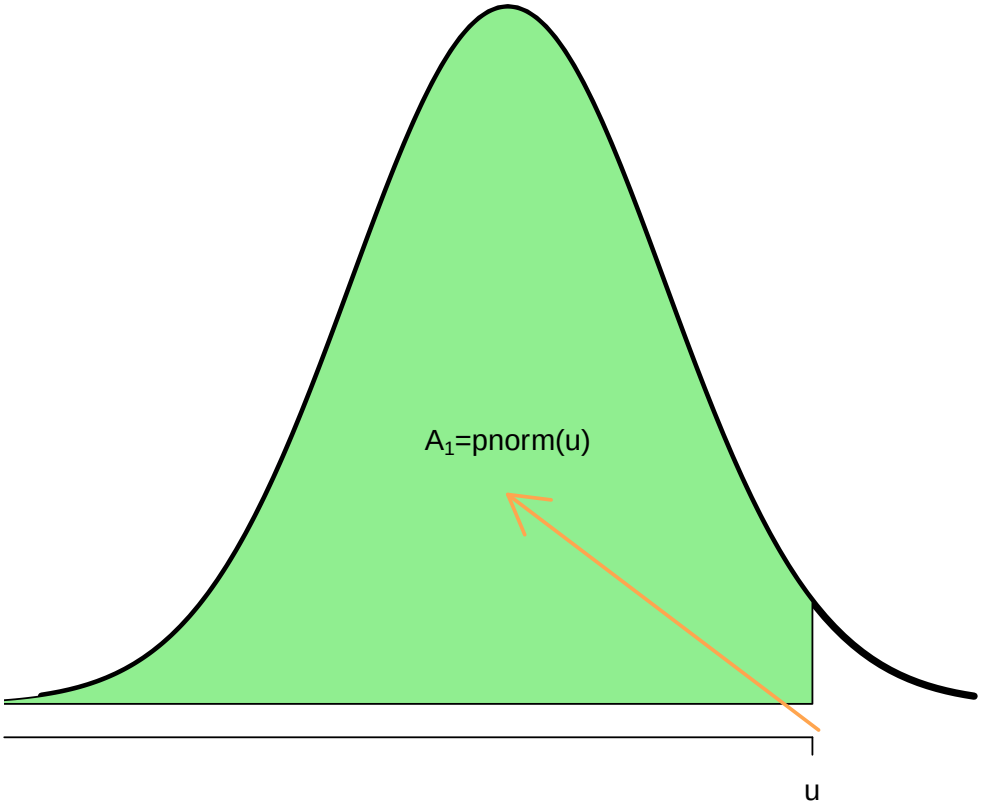
Con R

pnorm(u, μ, σ) fornisce l'area da $-\infty$ a u sotto alla normale generica. Se non specifichiamo μ e σ ci riferiamo alla normale standard.

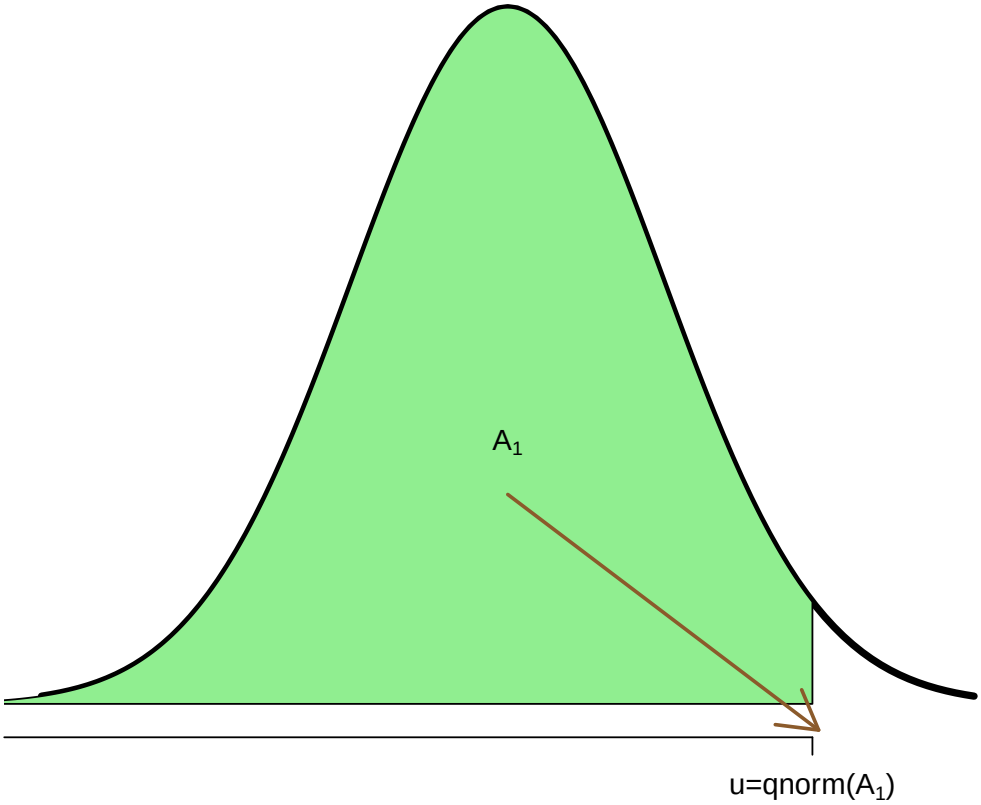
$$\text{pnorm}(u) = \int_{-\infty}^{u} \text{dnorm}(x)\text{d}x$$

pnorm è quindi una primitiva di dnorm. Per esempio

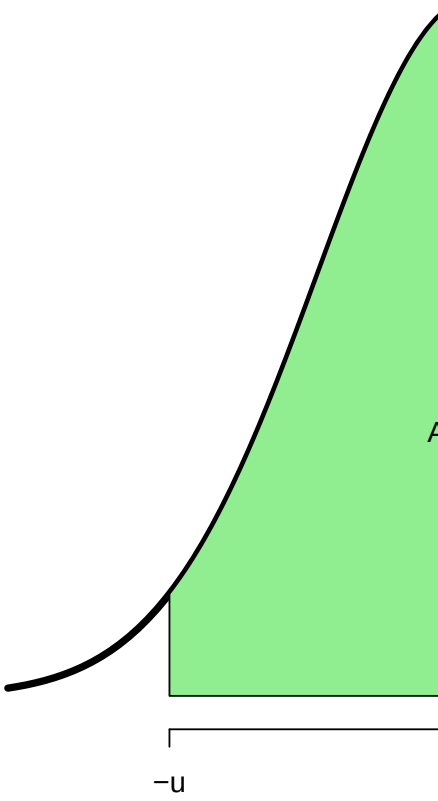
$$P(a < X < b) = \int_{a}^{b} \text{dnorm}(x, \mu, \sigma) = \text{pnorm}(x, \mu, \sigma)]_{a}^{b} = \text{pnorm}(b, \mu, \sigma) - \text{pnorm}(a, \mu, \sigma)$$



La funzione qnorm(x, μ, σ) è la funzione inversa di pnorm(x, μ, σ).

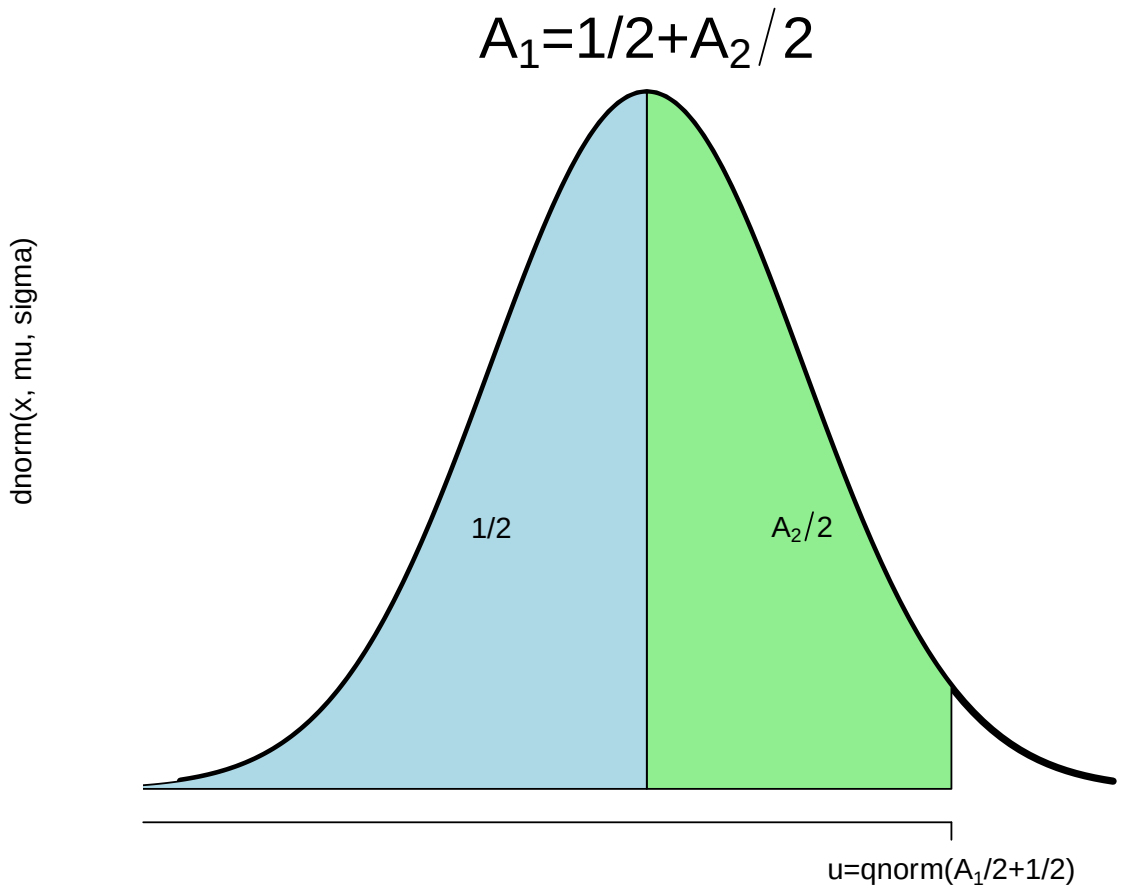


Vogliamo una funzione, diciamo 'U' che assegnato un valore di area $A_2 = 1 - \epsilon$ fornisca



l'ascissa u come in figura in modo che tra - u e u l'area sia A_2.

Il legame con il grafico precedente è



Allora la funzione che riproduce la tabella è

```
U <-function (a) qnorm (1/2 + a/2)
```

E ad esempio

```
U(0.95)

## [1] 1.959964
```

1.16 Simulazioni

Come possiamo simulare variabili aventi valore nell'alfabeto assegnato? In effetti qualunque comando di generazione su un computer non è perfettamente casuale; infatti la generazione avviene in effetti in modo pseudo-casuale e secondo un meccanismo che dipende dallo stato interno del computer codificato in una variabile indicata con `.Random.seed`. Se il *seme* iniziale è lo stesso i numeri generati saranno uguali. Spesso conviene che i calcoli (ad esempio a fine didattico) siano riproducibili. Ad esempio mettendo in una variabile `seme` il valore corrente di `.Random.seed` e richiamandolo o generandolo all'occorrenza. Un altro modo di procedere consiste nell'impostare il valore di `.Random.seed` attraverso il comando `set.seed` la cui sintassi è `set.seed`(n) dove n è un numero intero.

```
set.seed(3)
```

A questo punto possiamo simulare le variabili richieste usando la struttura

$$\texttt{sample}(\textit{alfabeto}, n) \tag{1.1}$$

Se l'alfabeto consiste di tutte le lettere minuscole dell'alfabeto ordinario e ne vogliamo selezionare $n = 8$ (in modo che ciascun uscita abbia la stessa probabilità) basta scrivere

```
sample(letters,8)

## [1] "e" "u" "j" "h" "n" "m" "c" "f"
```

Se invece l'alfabeto consiste delle basi del DNA

```
alfabeto=c("A","C","G","T")
sample(alfabeto,2)

## [1] "G" "C"
```

Notiamo che

```
 sample(alfabeto)

## [1] "G" "C" "T" "A"
```

restituisce una permutazione dell'alfabeto, mentre chiedendo un campione di lunghezza superiore alla lunghezza dell'alfabeto otteniamo un messaggio di errore. Possiamo però immaginare di re-immettere la lettera estratta nell'urna dopo ogni estrazione. In questo caso non c'è limite alla sequenza generata. Per esempio

```
alfabeto=c("testa","croce")
sample(alfabeto,5,replace=T)

## [1] "croce" "croce" "testa" "croce" "croce"
```

Il precursore del dado era chiamato *astragalo* ed era giocato nell'antica Grecia e nell'antica Roma [**?**]. Gli astragali sono dei piccoli ossicini di forma irregolare ed hanno 6 facce ma atterranno in modo stabile solo su 4 di esse numerate 1, 3, 4 e 6 con probabilità all'incirca 0.4 per il 3 e il 4 e di 0.1 per l'1 e il 6. In altre parole l'astragalo è descritto dalla tabella

valore	probabilità
1	0.1
3	0.4
4	0.4
6	0.1

Il tiro più gettonato all'epoca era l'uscita di 4 facce diverse nel lancio di 4 astragali e si chiamava *Venus*. Il lancio considerato peggiore sul singolo lancio era l'1 chiamato cane o avvoltoio. Per simulare un astragalo su un computer

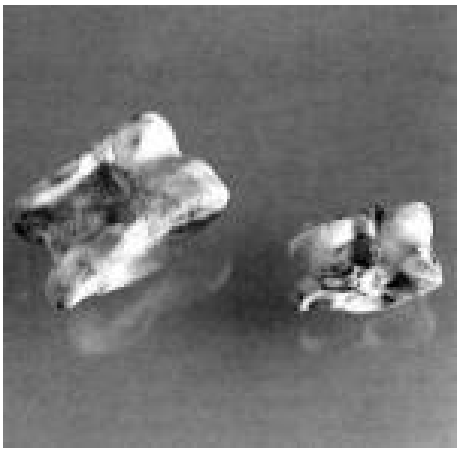


Figura 1.1: Astragalo.

```
sample(c(1,3,4,6),4,replace=T,prob=c(0.1,0.4,0.4,0.1))

## [1] 3 3 3 3
```

Torniamo ora ai classici dadi a 6 facce. Supponiamo di lanciare 100 volte un dado equo a 6 facce e di registrare in `x` le uscite rilevate

```
set.seed(3)
dadi100<-sample(1:6,100,replace=T)
dadi100

##   [1] 2 5 3 2 4 4 1 2 4 4 4 4 4 4 6 5 1 5 6 2 2 1 1 1 2 5 4 6
##  [29] 4 5 3 3 2 3 2 3 6 2 4 2 2 5 2 4 3 2 1 1 2 5 2 2 6 6 6 6
##  [57] 3 2 1 2 5 1 5 1 5 2 5 4 3 1 5 5 6 6 4 4 1 1 5 5 5 4 3 1
##  [85] 6 6 2 3 4 6 1 2 3 5 6 2 2 2 2 5
```

Volendo invece simulare una combinazione da giocare al SuperEnalotto possiamo scrivere

```
(x<-sample(1:90,6,replace=T))

## [1] 69 62 19 65 55 31
```

I numeri usciti sono stati salvati in una variabile `x`, per poter effettuare la ricerca di indicatori statistici. Il comando che consente di ordinare una lista o un vettore è `sort`, esso può essere usato in associazione al nome di una variabile o di una lista, ossia:

$$\texttt{sort}(\textit{variabile/lista}) \tag{1.2}$$

Volendo ordinare i numeri precedentemente ricavati scriveremo

```
sort(x)

## [1] 19 31 55 62 65 69
```