Metodi computazionali intensivi

Capitolo 19

Analisi statistica dei dati biologici

Vantaggi dell'uso dei PC in statistica

L'avvento del Personal Computer in statistica ha cambiato non solo il tempo necessario al calcolo, ma ha reso possibile nuovi metodi di analisi che prima non erano possibili:

- **Simulazione** (verifica di ipotesi)
- **Randomizzazione** (verifica di ipotesi)
- **Bootstrap** (precisione delle stime)

Metodi utili quando le assunzioni dei metodi standard non possono essere soddisfatte.

MCI: Verifica delle ipotesi con la Simulazione

- Questa tecnica viene impiegata quando il problema è determinare la distribuzione di probabilità della statistica test quando l'ipotesi nulla è vera.

La simulazione impiega il computer per riprodurre artificialmente (simula) il campionamento da una popolazione sotto l'ipotesi nulla.

Il risultato della simulazione (in genere ripetuta 1 000 o 10 000 volte) è usato per costruire la distribuzione di frequenza della statistica test approssimata sotto H0.

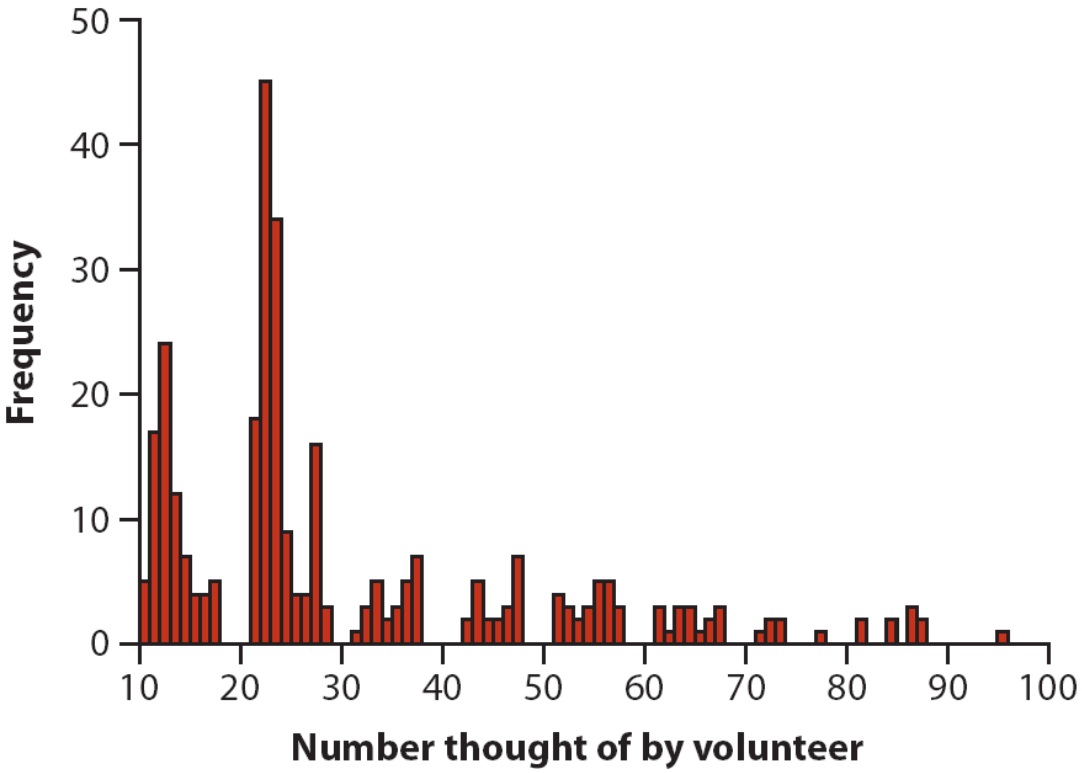
Questa distribuzione è in seguito utilizzata per calcolare il Pvalue.

MCI: Verifica delle ipotesi con la Simulazione – Esempio caso

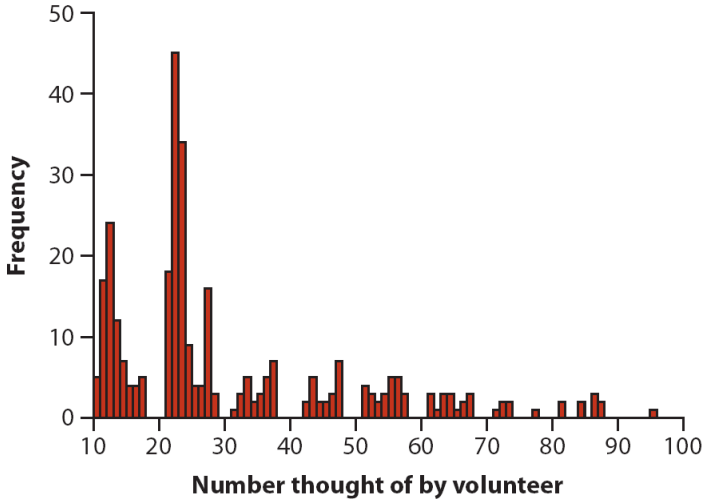
Esempio 19.1

Come ha fatto a saperlo?
La non casualità di una scelta a caso

I cosiddetti «mentalisti» sono persone che sostengono di essere dotate di poteri telepatici, cioè della capacità di leggere nella mente altrui. È però anche possibile fingere in maniera convincente di possedere tali poteri. In un tipo di performance, per esempio, il mentalista chiede ad ogni spettatore che assiste all'esibizione di pensare un numero a due cifre. Dopo avere affermato di essere capace di leggere nella mente delle persone che ha di fronte, il mentalista dice un numero che una frazione rilevante del pubblico stava effettivamente pensando. L'impresa sarebbe sorprendente se le persone pensassero a tutti i numeri possibili a due cifre con uguale probabilità, ma non lo sarebbe se tutte le persone tendessero a scegliere gli stessi pochi numeri. La Figura 19.1-1 riporta la distribuzione dei numeri scelti indipendentemente da 350 volontari (Marks, 2000). I numeri sono stati scelti con uguale probabilità? ■



MCI: Verifica delle ipotesi con la Simulazione – Esempio caso



Guardando l'istogramma sembra che i numeri non siano stati scelti con uguale probabilità

Vogliamo testare l'ipotesi nulla:
H0: I numeri a due cifre vengono scelti con uguale probabilità
HA: I numeri a due cifre non vengono scelti con uguale probabilità

Potremmo applicare il test Chi-quadro di bontà di adattamento su 90 categorie ma violeremmo l'assunzione del <20% di categorie con n°<5, quindi la statistica chi-quadro calcolato sui dati non segue una distribuzione chi-quadro

MCI: Verifica delle ipotesi con la Simulazione – Esempio caso

Con la simulazione possiamo trovare la distribuzione nulla della nostra statistica chi-quadro!

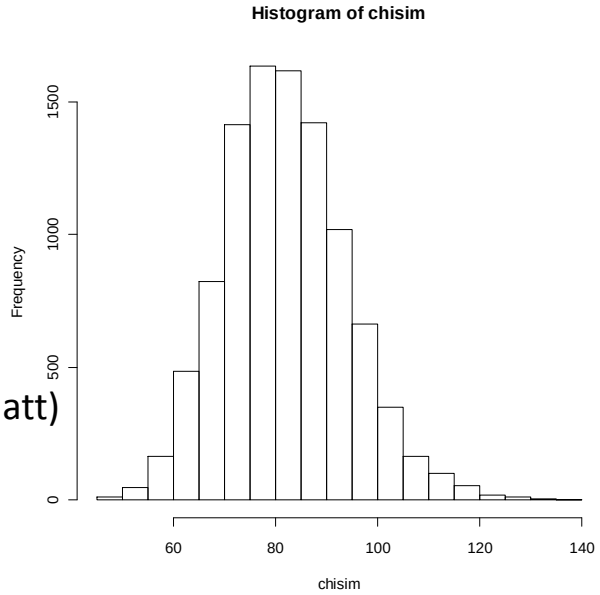
1) ***Usiamo un pc per creare una popolazione immaginaria i cui valori dei parametri siano quelli specificati da H0 e campioniamo n volte da questa popolazione***: ripetiamo n volte l'estrazione casuale di 350 numeri a due cifre
2) ***Calcoliamo la statistica test chi-quadro su ogni campione simulato***
3) ***Raccogliamo i valori della statistica test calcolata nelle n ripetizioni***: la distribuzione di frequenza degli n chi-quadro calcolati rappresenta la nostra distribuzione nulla simulata
4) **Confrontiamo la statistica test ottenuta nei dati con la distribuzione nulla**: nel caso del chi-quadro, il numero di simulazioni che hanno prodotto un chi-quadro simulato >= al valore osservato rappresenta il Pvalue.

MCI: Verifica delle ipotesi con la Simulazione – Esempio caso con R

```
#caricare il file «Metodi_computazionali_intensivi_Caso.txt»
x<-read.table(choose.files(),header=T)
#visualizzo l'istogramma
hist(x$numeri,breaks=90)
#H0: I numeri a due cifre vengono scelti con uguale probabilità
#HA: I numeri a due cifre non vengono scelti con uguale probabilità
#calcolo il valore chi-calcolato sui dati
numatt<-350/90
chioss<-sum((table(factor(x$numeri,levels=10:99))-numatt)^2/numatt)
#genero UNA simulazione secondo l'ipotesi nulla, visualizzo l'istogramma e calcolo il suo valore di chi-quardro
esx<-sample(10:99,350,replace=T)
hist(esx,breaks=90)
sum((table(factor(esx,levels=10:99))-numatt)^2/numatt)
```

MCI: Verifica delle ipotesi con la Simulazione – Esempio caso con R

```
#scelgo quante simulazioni fare
n<-10000
#inizializzo una variabile dove salvare i chi-quadro simulati
chisim<-c()
#ripeto n volte la simulazione e calcolo chi-quadro
for (i in 1:n){
        esx<-sample(10:99,350,replace=T)
        chisim[i]<-sum((table(factor(esx,levels=10:99))-numatt)^2/numatt)
}
hist(chisim)
```



Abbiamo ottenuto la distribuzione nulla della statistica test costruita sulla base di 10000 simulazioni.

Il pvalue per il test chi-quadro è la proporzione di simulazioni maggiori-uguali al chi-quadro osservato

```
#calcolo Pvalue
sum(chisim>=chioss)/n
#se il risultato è 0, allora significa che il pvalue è minore di 0,0001
#Rifiuto H0, i numeri non vengono scelti con uguale probabilità
```

MCI: Il test di randomizzazione

- Questa tecnica viene impiegata per verificare l'ipotesi di associazione tra due variabili:
  - tra due variabili categoriche (tabelle di contingenza)
  - tra una variabile categorica e una numerica (test t)
  - tra due variabili numeriche (correlazione)

Utilizzabile al posto dei metodi standard quando le loro assunzioni non sono soddisfatte o quando la distribuzione nulla è sconosciuta.

MCI: Il test di randomizzazione

Per effettuare un **test di randomizzazione**:

- Scegliere una statistica test (chi-quadro, $\Delta\bar{x}$, r)
- Assegnare A CASO agli individui UNA delle due variabili, e calcolare la statistica test sul nuovo campione (una variabile DEVE rimanere fissa)
- Ripetere la procedura n volte (n deve essere grande) per generare la distribuzione nulla della statistica test
- Confrontare la statistica test calcolata nel campione reale rispetto alla distribuzione nulla per ottenere un Pvalue

Un test di randomizzazione genera una distribuzione nulla dell'associazione tra due variabili riassegnando casualmente i valori di una delle due variabili ai dati e ripetendo questo processo molte volte.

MCI: Il test di randomizzazione – Esempio pseudoscorpioni

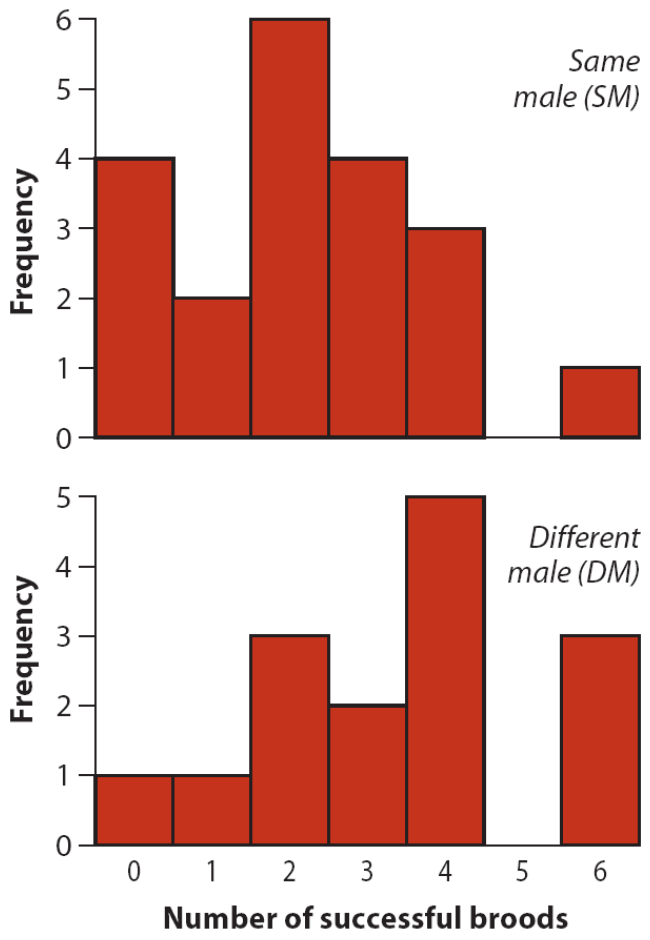
Esempio 19.2

Le ragazze vogliono solo... diversità genetica

Gli pseudoscorpioni della specie *Cordylochernes scorpioides* vivono nelle foreste tropicali, dove si fanno trasportare sul dorso dei coleotteri arlecchini per raggiungere i fichi marcescenti di cui si alimentano. Le femmine di questa specie sono promiscue e si accoppiano con più maschi durante la loro breve vita. Non è chiaro quali vantaggi comporti l'accoppiamento multiplo, dato che i maschi non si prendono cura della propria prole ed è sufficiente un singolo accoppiamento per fornire tutti gli spermatozoi di cui la femmina ha bisogno per fecondare le uova.

Ci potrebbe essere un vantaggio se gli spermatozoi di alcuni maschi fossero geneticamente incompatibili con una data femmina, e quindi, accoppiandosi più volte, una femmina aumenterebbe la probabilità di accoppiarsi con almeno un maschio i cui spermatozoi sono compatibili con essa. Per indagare questa ipotesi, Newcomer et al. (1999) hanno registrato il numero di covate «di successo» da parte di femmine di *C. scorpioides* assegnate casualmente a uno di due gruppi di trattamento. Ciascuna femmina del primo gruppo è stata fatta accoppiare con due maschi (DM), mentre le femmine dell'altro gruppo sono state fatte accoppiare due volte con lo stesso maschio (SM). Facendo accoppiare ciascuna femmina due volte in entrambi i trattamenti, la quantità totale di spermatozoi ricevuta era la stessa. Le femmine DM hanno però ricevuto spermatozoi più diversi geneticamente rispetto a quelli ricevuti dalle femmine SM. I ricercatori hanno confrontato il numero medio di covate di successo in ciascun gruppo. I dati[3] sono raccolti in Tabella 19.2-1 e sono visualizzati in Figura 19.2-1. ■

MCI: Il test di randomizzazione – Esempio pseudoscorpioni



- Distribuzione non normale (no test t)
- Distribuzioni con forme diverse (no Mann-Whitney)
- Test di randomizzazione (ok se n campionaria è grande)

MCI: Il test di randomizzazione – Esempio pseudoscorpioni

Tabella 19.2-1

Il numero di covate di successo in femmine di pseudoscorpione che sono state fatte accoppiare due volte con lo stesso maschio (SM) oppure una volta con due differenti maschi (DM). Sono presentati i dati per 20 femmine SM e 16 femmine DM. I dati relativi ai diversi trattamenti sono denotati con un colore diverso in modo da evidenziare più facilmente l'origine di ciascun valore nella successiva Tabella 19.2-2.

Tipo di accoppiamento	Numero di covate di successo	Tipo di accoppiamento	Numero di covate di successo
SM	4	DM	2
SM	0	DM	0
SM	3	DM	2
SM	1	DM	6
SM	2	DM	4
SM	3	DM	3
SM	4	DM	4
SM	2	DM	4
SM	4	DM	2
SM	2	DM	7
SM	0	DM	4
SM	2	DM	1
SM	0	DM	6
SM	1	DM	3
SM	2	DM	6
SM	6	DM	4
SM	0		
SM	2		
SM	3		
SM	3		

Due variabili:

1- Categorica: Tipo di accoppiamento

2- Numerica: Numero di covate di successo

Dobbiamo testare l'ipotesi:

H0: Non c'è differenza nel numero medio di covate di successo tra i due gruppi

HA: C'è una differenza ne numero medio di covate di successo tra i due gruppi

MCI: Il test di randomizzazione – Esempio pseudoscorpioni

La statistica test più semplice in questo caso è la differenza tra le medie dei due gruppi:

$$\bar{Y}_{SM} - \bar{Y}_{DM}$$

per i dati osservati questa differenza è:

$$2{,}2 - 3{,}625 = -1{,}425$$

a questo punto dobbiamo generare la distribuzione nulla della statistica test attraverso la randomizzazione

MCI: Il test di randomizzazione – Esempio pseudoscorpioni

Per generare la distribuzione nulla tramite la randomizzazione dobbiamo:

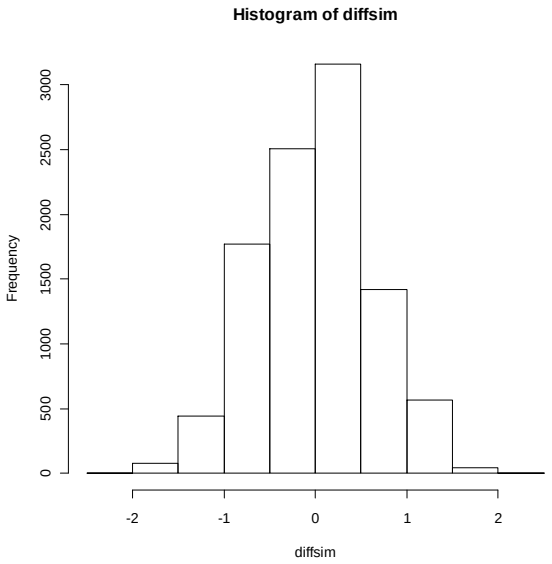
- Creare un insieme di dati randomizzato in cui i valori della variabile risposta (numero di covate) siano riassegnati casualmente: teniamo fissa la variabile di gruppo e riassegniamo a ogni individuo un valore casuale di covate tra quelli misurati.
- Calcolare la misura dell'associazione per il campione randomizzato: calcoliamo la statistica test nel campione randomizzato, cioè calcoliamo la differenza tra le medie nei due gruppi.
- Ripetere molte volte il processo di randomizzazione: replichiamo i primi due passaggi un numero elevato di volte (1000 o 10000 volte)
- Determinare il P value utilizzando la distribuzione nulla formata dalla distribuzione di frequenza della statistica test nei campioni randomizzati: nel caso dei pseudoscorpioni l'ipotesi alternativa è bidirezionale quindi il P value è uguale alla proporzione di campioni randomizzati con la statistica test più estrema rispetto al valore osservato nei dati reali, moltiplicato per due.

MCI: Il test di randomizzazione – Esempio pseudoscorpioni con R

```
#caricare il file «Metodi_computazionali_intensivi_Pseudoscorpioni.txt»
x<-read.table(choose.files(),header=T)
#visualizzo l'istogramma del numero di covate di successo nei due gruppi
barplot(table(x$broods[x$matingtype=="SM"]))
barplot(table(x$broods[x$matingtype=="DM"]))
#H0: Non c'è differenza nel numero medio di covate di successo tra i due gruppi
#HA: C'è una differenza ne numero medio di covate di successo tra i due gruppi
#calcolo la statistica test (differenza medie nei gruppi) nel campione osservato
do<-mean(x$broods[x$matingtype=="SM"])-mean(x$broods[x$matingtype=="DM"])
#genero un campione randomizzato
#per prima cosa creo un oggetto copia dei dati reali
ecr<-x
#poi effettuo la randomizzazione sostituendo la colonna «broods» con il risultato di un campionamento casuale CON RIMPIAZZO dai valori di «broods» osservati
ecr$broods<-sample(x$broods,length(x$broods),replace=F)
#calcolo la statistica test sul nuovo campione creato
dr<-mean(ecr$broods[ecr$matingtype=="SM"])-mean(ecr$broods[ecr$matingtype=="DM"])
```

MCI: Il test di randomizzazione – Esempio pseudoscorpioni con R

```
#scelgo quante randomizzazioni fare
n<-10000
#inizializzo una variabile dove salvare la statistica test in ogni randomizzazione
diffsim<-c()
#ripeto n volte la simulazione e calcolo chi-quadro
for (i in 1:n){
        ecr$broods<-sample(x$broods,length(x$broods),replace=F)
        diffsim[i]<-mean(ecr$broods[ecr$matingtype=="SM"])-
        mean(ecr$broods[ecr$matingtype=="DM"])
}
hist(diffsim)
```



Abbiamo ottenuto la distribuzione nulla della statistica test costruita sulla base di 10000 randomizzazioni.

Il pvalue in questo caso si calcola contando il numero di simulazioni minori-uguali alla statistica test osservata (-1,425)

```
#calcolo Pvalue a due code
sum(diffsim<=do)/n*2
#il P value ottenuto è circa 0.025 (NB: è normale che il valore sia leggermente diverso ogni volta che si ripete l'analisi)
#Rifiuto H0, le femmine degli pseudoscorpioni hanno una maggiore probabilità di avere una covata di successo quando si accoppiano con due maschi, rispetto al caso in cui si accoppiano con lo stesso maschio
```

MCI: Il bootstrap

- Questa tecnica viene impiegata per **approssimare la distribuzione campionaria di una stima**.

A differenza della simulazione e della randomizzazione, il **bootstrap permette di calcolare un errore standard o un intervallo di confidenza per una stima**.

Utilizzabile al posto dei metodi standard quando non è disponibile alcuna formula per calcolare l'errore standard o quando non si conosce la distribuzione campionaria della stima di interesse.

MCI: Il bootstrap

- Questa tecnica viene impiegata per **approssimare la distribuzione campionaria di una stima**.

A differenza della simulazione e della randomizzazione, il **bootstrap permette di calcolare un errore standard o un intervallo di confidenza per una stima**.

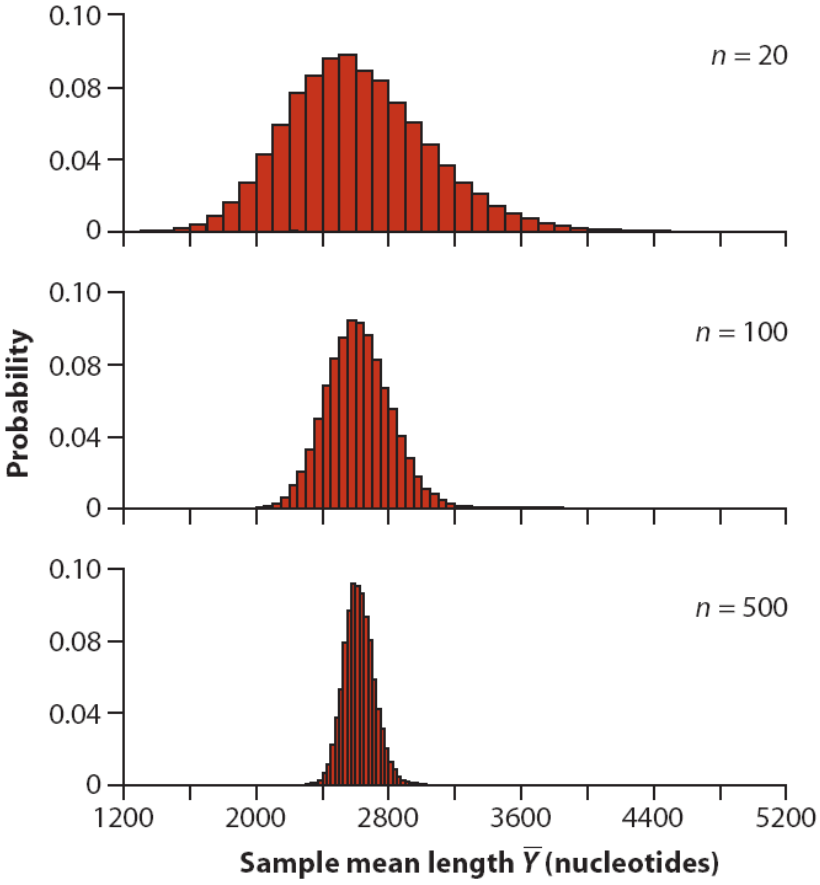
Utilizzabile al posto dei metodi standard quando non è disponibile alcuna formula per calcolare l'errore standard o quando non si conosce la distribuzione campionaria della stima di interesse.

MCI: Il bootstrap

La distribuzione campionaria è la distribuzione di probabilità delle stime campionarie quando una popolazione viene campionata ripetutamente e sempre nello stesso modo.

Ad esempio potremmo ottenere l'errore standard di una stima estraendo ripetutamente campioni dalla popolazione, calcolando ogni volta la stima campionaria e valutando infine la deviazione standard delle stime campionarie

Effettuare questo procedimento in realtà è spesso difficile se non impossibile o estremamente costoso



MCI: Il bootstrap

Se la dimensione del campione estratto dalla popolazione è sufficientemente grande, allora abbiamo accesso a una parte della popolazione.

Il **bootstrap** consiste in un campionamento ripetuto: **invece di campionare dalla popolazione campioniamo dai dati in nostro possesso**.

Se il campione è abbastanza grande, allora i campioni di bootstrap estratti avranno proprietà statistiche simili a quelli estratti dalla popolazione stessa.

MCI: Il bootstrap – Esempio scimpanzé

Esempio 19.3

Il centro del linguaggio nell'encefalo degli scimpanzé

Una delle caratteristiche che distinguono l'uomo dalle altre specie è la capacità di sviluppare un linguaggio complesso. Gli scimpanzé e i gorilla sono talvolta in grado di apprendere un linguaggio rudimentale, ma con una capacità assai inferiore alla nostra. Nell'uomo il linguaggio è associato a una regione dell'encefalo nota come «area 44 di Brodmann», che fa parte dell'«area di Broca». L'area 44 di Brodmann è più estesa nell'emisfero sinistro dell'encefalo rispetto all'emisfero destro, ed è noto che questa asimmetria è importante per lo sviluppo del linguaggio. Grazie all'avvento dell'imaging a risonanza magnetica (MRI), possiamo indagare se quest'area sia asimmetrica anche nell'encefalo di altre scimmie antropomorfe. Un campione di 20 scimpanzé è stato esaminato con l'MRI ed è stata registrata l'asimmetria della loro area 44 di Brodmann (Cantalupo & Hopkins, 2001). L'asimmetria è stata misurata valutando la differenza proporzionale tra l'estensione dell'area 44 di Brodmann nell'emisfero sinistro e nell'emisfero destro. (La differenza proporzionale è la differenza tra la misura sinistra e la misura destra divisa per la media dei due emisferi.) I dati grezzi sono raccolti nella Tabella 19.3-1. La mediana campionaria dei punteggi di asimmetria è risultata pari a 0,14. Vogliamo quantificare l'incertezza di questa stima della mediana nella popolazione. ■

Tabella 19.3-1

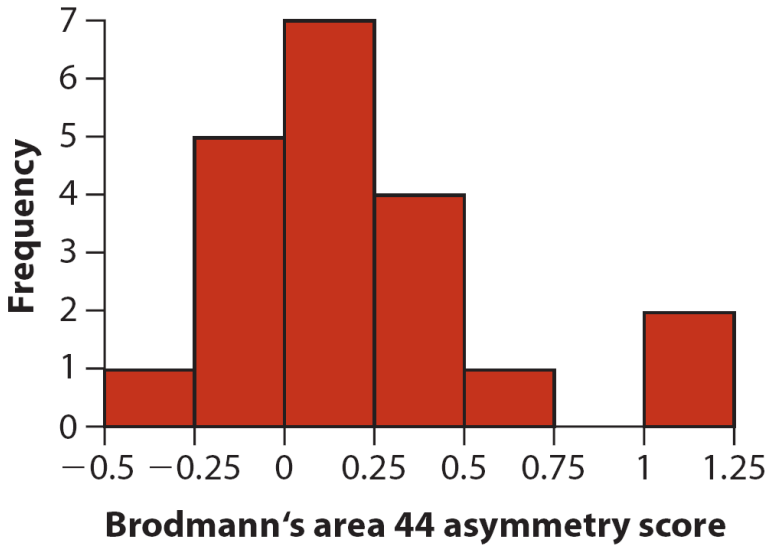
Valori di asimmetria per l'area 44 di Brodmann in 20 scimpanzé.

Nome dello scimpanzé	Valore di asimmetria
Austin	0,30
Carmichael	0,16
Chuck	–0,24
Dobbs	–0,25
Donald	0,36
Hoboh	0,17
Jimmy Carter	0,11
Lazarus	0,12
Merv	0,34
Storer	0,32
Ada	0,71
Anna	0,09
Atlanta	1,12
Cheri	–0,22
Jeannie	1,19
Kengee	0,01
Lana	–0,24
Lulu	0,24
Mary	–0,30
Panzee	–0,16

MCI: Il bootstrap – Esempio scimpanzé

Nel caso in cui la variabile abbia una distribuzione normale, l'errore standard della mediana può essere calcolato come:

$$ES_{mediana} = 1{,}253\frac{\sigma}{\sqrt{n}}$$



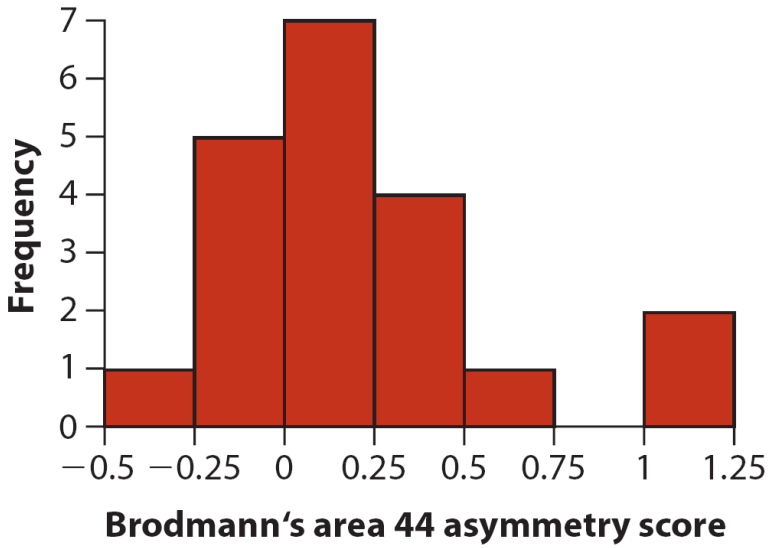
La distribuzione di frequenza dei punteggi di asimmetria è asimmetrica a destra e inoltre esiste il forte sospetto di bimodalità per cui sarebbe molto rischioso usare le formule conosciute per calcolare l'errore standard.

IL metodo più appropriato è il bootstrap.

MCI: Il bootstrap – Esempio scimpanzé

Nel caso in cui la variabile abbia una distribuzione normale, l'errore standard della mediana può essere calcolato come:

$$ES_{mediana} = 1{,}253\frac{\sigma}{\sqrt{n}}$$



La distribuzione di frequenza dei punteggi di asimmetria è asimmetrica a destra e inoltre esiste il forte sospetto di bimodalità per cui sarebbe molto rischioso usare le formule conosciute per calcolare l'errore standard.

IL metodo più appropriato è il bootstrap.

MCI: Il bootstrap – Esempio scimpanzé

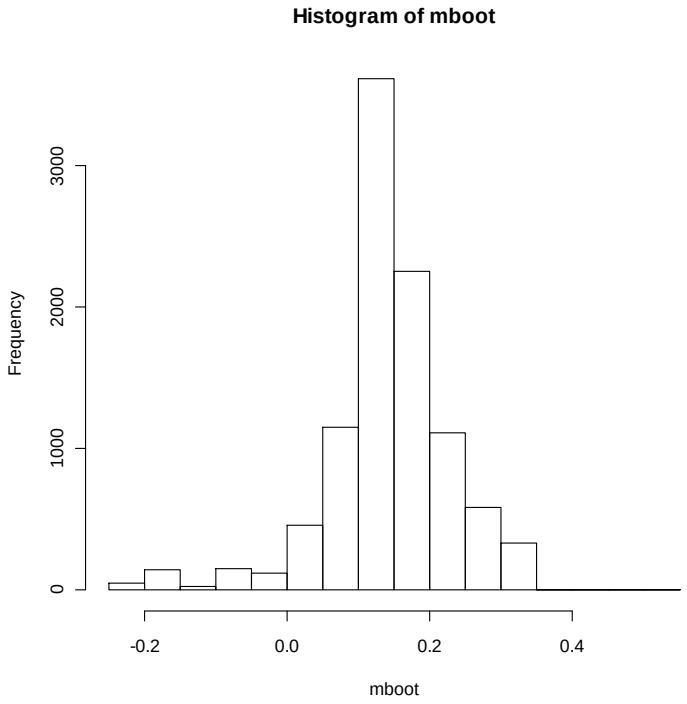
- Calcolo **dell'errore standard**:
  - Usare un computer per estrarre un campione casuale di individui dai dati originali (campione di bootstrap): il campione deve essere delle stesse dimensioni dei dati originali e ogni individuo ha la stessa probabilità di essere campionato e può essere preso più di una volta (campionamento con rimpiazzo)
  - Calcolare la stima usando i dati del campione di bootstrap: ad esempio la mediana del campione.
  - Ripetere il processo n volte (1 000 o 10 000 volte): la distribuzione di frequenza di tutte le stime di bootstrap approssima la distribuzione campionaria della stima.
  - Calcolare **la deviazione standard campionaria di tutte le stime di bootstrap** ottenute: la grandezza che si ottiene è **l'errore standard di bootstrap**.

MCI: Il bootstrap – Esempio scimpanzé con R (errore standard)

```
#caricare il file «Metodi_computazionali_intensivi_Scimpanzé.txt»
x<-read.table(choose.files(),header=T)
#calcolo la mediana campionaria osservata
moss<-median(x$Rounded_AQ)
#genero un campione di bootstrap
cb<-sample(x$Rounded_AQ,length(x$Rounded_AQ),replace=T)
#calcolo la mediana nel campione di bootstrap
mb<-median(cb)
```

MCI: Il bootstrap – Esempio scimpanzé con R

```
#scelgo quante repliche di bootstrap fare
n<-10000
#inizializzo una variabile dove salvare la mediana
#di ogni campione di bootstrap
mboot<-c()
#ripeto n volte la simulazione e calcolo chi-quadro
for (i in 1:n){
        cb<-sample(x$Rounded_AQ,length(x$Rounded_AQ),replace=T)
        mboot[i]<-median(cb)
}
#visualizzo la distribuzione campionaria della stima (mediana)
hist(mboot)
#calcolo l'errore standard di bootstrap per la mediana
ES<-sd(mboot)
```



Poiché i campioni di bootstrap sono stati ottenuti dai dati, l'errore standard di bootstrap tende ad essere lievemente inferiore all'errore standard vero.

Questo effetto è trascurabile quando la dimensione campionaria è grande.

MCI: Il bootstrap – Esempio scimpanzé

- Calcolo dell'**intervallo di confidenza**:
  - L'intervallo di confidenza di bootstrap con confidenza (1-α) si può ottenere dalla distribuzione campionaria di bootstrap trovando i punti che separano una frazione α/2 dell'area della distribuzione in ciascuna delle due code (sinistra e destra).

Per esempio, un intervallo di confidenza al 95% è compreso tra il quantile 0,025 e il quantile 0,975 della distribuzione campionari a di bootstrap.

MCI: Il bootstrap – Esempio scimpanzé con R

```
#una volta ottenute le repliche di bootstrap calcolo l'IC95% guardando gli estremi dei quantili 0.025 e 0.975
IC95<-quantile(mboot,c(0.025,0.975))
#l'intervallo al 99% si calcola modificando i quantili (0.005.0.995)
IC99<-quantile(mboot,c(0.005,0.995))
```

MCI: Il bootstrap

- Assunzioni e limitazioni:
  - Il campione deve essere estratto casualmente dalla popolazione
  - Il campione deve essere grande a sufficienza affinché la distribuzione di frequenza delle misure nel campione sia una buona rappresentazione di quella nella popolazione
  - Analisi di bootstrap su campioni piccoli produrranno errori standard troppo piccoli e intervalli di confidenza troppo stretti, determinando una sovrastima della precisione della stima.