C Richiami di inferenza statistica

SOMMARIO

Obiettivi d'apprendimento

Lo studio di questo capitolo dovrebbe insegnarvi come:

1. Discutere la differenza fra una popolazione e un campione e spiegare perché usiamo i dati campionari per fare inferenza sui parametri della popolazione.
2. Collegare i concetti di popolazione e variabile casuale, indicando in che modo la funzione di densità o di probabilità, il valore atteso e la varianza di una variabile casuale ci forniscono informazioni sulla popolazione.
3. Spiegare la differenza fra media della popolazione e media campionaria.
4. Spiegare la differenza fra una stima e uno stimatore e le ragioni per le quali quest'ultimo è una variabile casuale.
5. Spiegare il significato dei termini variabilità campionaria e distribuzione campionaria.
6. Spiegare il concetto di correttezza e usare le proprietà dei valori attesi per mostrare che la media campionaria è corretta.
7. Spiegare perché all'interno del gruppo degli stimatori corretti preferiamo quelli con varianze più piccole a quelli con varianze più elevate.
8. Descrivere il teorema del limite centrale e le sue implicazioni sull'inferenza statistica.
9. Spiegare la relazione fra lo "scarto quadratico medio" della popolazione e lo standard error della media campionaria.
10. Spiegare la differenza fra stima puntuale e intervallare, nonché costruire e interpretare stime intervallari della media di una popolazione a partire da un campione di osservazioni.
11. Spiegare in termini semplici che cosa significhi e che cosa non significhi il termine "livello di confidenza del 95%" nell'ambito della stima intervallare.
12. Spiegare l'obiettivo di una verifica d'ipotesi e illustrare gli elementi necessari all'effettuazione di un test.
13. Discutere le implicazioni delle diverse ipotesi alternative nel verificare l'ipotesi nulla $H_0 : \mu = 7$. Fornire un esempio economico in cui un test verifica la validità di questa ipotesi rispetto a una delle alternative.
14. Descrivere il livello di significatività di un test e spiegare la differenza fra livello di significatività e p-value di un test.
15. Definire l'errore di prima specie e la sua relazione con il livello di significatività di un test.
16. Spiegare la differenza fra i test a una coda e quelli a due code, descrivendo quando debba essere preferito l'uno o l'altro tipo.
17. Spiegare la differenza fra le frasi "accetto l'ipotesi nulla" e "non rifiuto l'ipotesi nulla" e discuterne le implicazioni.
18. Fornire una spiegazione intuitiva della stima di massima verosimiglianza e descrivere le proprietà dello stimatore di massima verosimiglianza.
19. Elencare i tre tipi di test associati alla stima di massima verosimiglianza e discuterne somiglianze e differenze.
20. Distinguere fra stima parametrica e non parametrica.
21. Capire in che modo uno stimatore kernel della densità si adatta a una distribuzione empirica.

Parole chiave

BLUE
campione casuale
distribuzione asintotica
distribuzione campionaria
errore di prima specie
errore di seconda specie
funzione di logverosimiglianza
funzione di verosimiglianza
inferenza statistica
ipotesi alternativa
legge dei grandi numeri
livello di significatività
media campionaria
misura di informazione
momenti centrati
parametro della popolazione
p-value
regione di rifiuto
standard error
standard error della media
standard error della stima
statistica test
stima
stima di massima verosimiglianza
stima intervallare
stima non parametrica
stima puntuale
stimatore
stimatore kernel della densità
stimatore lineare
stimatori corretti
teorema del limite centrale
teoria dei campioni
test a due code
test del moltiplicatore di Lagrange
test del rapporto di verosimiglianza
test di Wald
variabilità campionaria
varianza campionaria

Gli economisti sono interessati a studiare relazioni fra variabili economiche. Per esempio, quale aumento ci possiamo aspettare per le vendite del gelato Gelida Bontà se il suo prezzo viene ridotto del 5%? Di quanto crescerà la spesa familiare in beni alimentari se il reddito aumenta di 100 dollari al mese? Domande come queste sono alla base di questo volume.

Talvolta tuttavia l'analisi si concentra su un'unica variabile economica. Per esempio, per prevedere uno spazio adeguato per ogni persona un progettista di sedili per aeroplani deve tenere conto della larghezza media del bacino di un passeggero, ma allo stesso tempo deve strutturare l'aeromobile in modo da raggiungere il numero di passeggeri che massimizza i profitti. Qual è la larghezza media del bacino dei passeggeri aerei negli Stati Uniti? Se si decide di considerare una larghezza di 18 pollici (45 centimetri circa), quale sarà la percentuale di clienti che non riusciranno a sedersi? Moltissime aziende, con prodotti che vanno dalle vetture elettriche per spostarsi sui campi da golf ai jeans da donna, sono costantemente alle prese con quesiti di questo tipo. Com'è possibile rispondere? Certamente non prendendo le misure di ogni singolo uomo, donna o bambino nella popolazione statunitense. Questa è una delle situazioni in cui viene usata l'inferenza statistica. Inferire significa "trarre delle conclusioni ragionando a partire da qualcosa di noto o assunto vero". L'**inferenza statistica** trae conclusioni a proposito di una popolazione a partire da un campione di osservazioni.

C.1. Un campione di osservazioni

Per fare inferenza statistica abbiamo bisogno di dati estratti dalla popolazione cui siamo interessati. Nel caso del progettista di sedili per aeroplani, questa popolazione coincide con quella dei residenti negli Stati Uniti di età superiore a due anni, dato che bambini di età inferiore possono volare "gratuitamente" sulle ginocchia dei loro poveri genitori. Un ramo specifico della statistica, chiamato **teoria dei campioni**, studia il meccanismo di campionamento migliore per raccogliere un insieme di dati rappresentativo della popolazione. Come pensereste di procedere se vi fosse chiesto di selezionare 50 misure di larghezza del bacino rappresentative dell'intera popolazione? Questo compito non è semplice. In linea di principio i 50 individui dovrebbero essere scelti casualmente, in modo da evitare distorsioni sistematiche. Supponiamo di concentrarci esclusivamente sulla popolazione composta

da adulti che prendono l'aereo, dato che di solito sui voli il numero di bambini è abbastanza ridotto. Il nostro consulente specialista in teoria dei campioni seleziona le osservazioni riportate nella tabella C.1 e raccolte nel file *hip.dat*.

Tabella C.1
Campione di osservazioni di larghezze del bacino

14,96	14,76	15,97	15,71	17,77
17,34	17,89	17,19	13,53	17,81
16,40	18,36	16,87	17,89	16,90
19,33	17,59	15,26	17,31	19,26
17,69	16,64	13,90	13,71	16,03
17,50	20,23	16,40	17,92	15,86
15,84	16,98	20,40	14,91	16,56
18,69	16,23	15,94	20,00	16,71
18,63	14,21	19,08	19,22	20,23
18,55	20,33	19,40	16,48	15,54

Un primo modo per analizzare un campione di osservazioni consiste nell'esaminarlo visivamente. La figura C.1 illustra un istogramma delle 50 osservazioni. Sulla base di questa figura la larghezza "media" del bacino in questo campione sembra essere compresa fra 16 e 18 pollici (rispettivamente 40 e 45 centimetri circa). Per il nostro progettista di sedili alla ricerca della configurazione di massimo profitto questa misura approssimata della media non è abbastanza accurata. Nel prossimo paragrafo costruiremo un modello econometrico per questo problema che sarà usato come punto di partenza per l'inferenza.

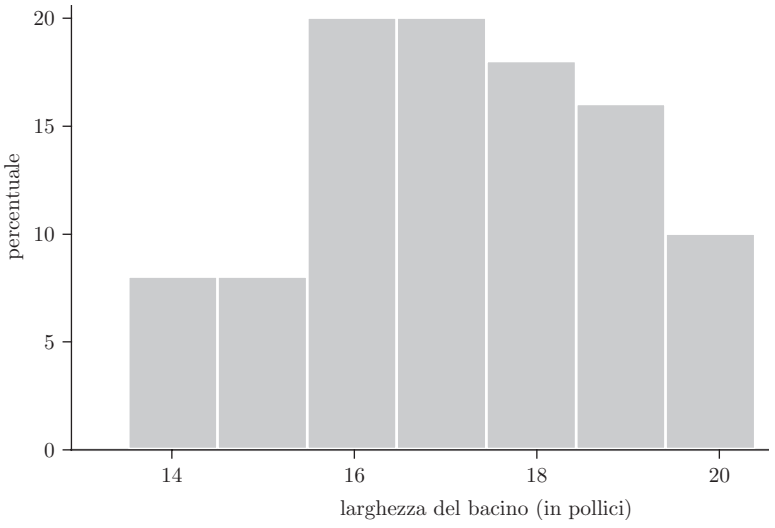

Figura C.1
Istogramma di larghezze del bacino.

C.2. Un modello econometrico

I dati riportati nella tabella C.1 sono stati ottenuti con un campionamento. L'estrazione di un campione da una popolazione costituisce un esperimento. La variabile di interesse in questo esperimento è la larghezza del bacino di un individuo. Prima di effettuare l'esperimento non conosciamo i valori della variabile di interesse; di conseguenza, la misura relativa a una persona scelta in maniera casuale sarà una variabile casuale. Indichiamo questa variabile casuale con Y. Il campione selezionato contiene $N = 50$ misure di larghezza del bacino per altrettanti individui,

$Y_1, Y_2, \ldots, Y_N$, dove ciascun Y_i rappresenta la larghezza del bacino di una persona diversa. I valori contenuti nella tabella C.1 rappresentano valori specifici di queste variabili, indicati con $y_1, y_2, \ldots, y_N$. Assumeremo che la popolazione considerata sia caratterizzata da un valore centrale, descritto dal valore atteso della variabile casuale Y:

$$E(Y) = \mu \tag{C.1}$$

Usiamo la lettera greca μ ("mu") per indicare la media della variabile casuale Y, che è anche la media della popolazione che stiamo studiando. Se conoscessimo μ, dunque, avremmo la risposta alla domanda: qual è la larghezza media del bacino di un adulto negli Stati Uniti? Data la sua importanza nel descrivere una caratteristica della popolazione, definiremo μ come un **parametro della popolazione** o, più brevemente, come un parametro. Il nostro obiettivo consiste nell'usare il campione di osservazioni riportate nella tabella C.1 per fare inferenza, o formulare valutazioni, sul parametro ignoto della popolazione μ.

L'altra caratteristica interessante di una variabile casuale è la sua variabilità, misurata dalla sua varianza:

$$\text{Var}(Y) = E[(Y - E(Y))^2] = E[(Y - \mu)^2] = \sigma^2 \tag{C.2}$$

Anche la varianza σ^2 è un parametro ignoto della popolazione. Nel Piccolo manuale di probabilità abbiamo spiegato che la varianza misura il grado di "dispersione" di una distribuzione di probabilità attorno alla media della popolazione; un valore elevato della varianza implica che la dispersione è maggiore, come illustrato dalla figura P.3. Nel contesto delle misure di larghezza del bacino, la varianza ci dice di quanto può variare la larghezza da una persona all'altra, entrambe scelte casualmente. Per semplificare la notazione, indicheremo la media e la varianza di una variabile casuale con $Y \sim (\mu, \sigma^2)$, dove $\sim$ significa "è distribuita come". Il primo elemento fra parentesi rappresenta la media della popolazione e il secondo la varianza. Si noti che fino a questo punto non abbiamo detto nulla sulla distribuzione di probabilità che pensiamo possa avere Y.

Il modello econometrico non è completo. Se il nostro è un campione casuale, possiamo assumere che $Y_1, Y_2, \ldots, Y_N$ siano statisticamente indipendenti. La larghezza del bacino di un individuo scelto in maniera casuale è indipendente da quella di qualsiasi altro individuo scelto casualmente. Assumiamo inoltre che ognuna delle osservazioni raccolte provenga dalla popolazione di interesse e che di conseguenza tutte le variabili casuali Y_i abbiano la stessa media e varianza: $Y_i \sim (\mu, \sigma^2)$. Le Y_i costituiscono un **campione casuale** in senso statistico, perché $Y_1, Y_2, \ldots, Y_N$ sono statisticamente indipendenti con la stessa distribuzione di probabilità. Talvolta è ragionevole assumere che nella popolazione i diversi valori abbiano distribuzione *normale*; in questo caso useremo la notazione $Y \sim \mathcal{N}(\mu, \sigma^2)$.

C.3. Stima della media di una popolazione

Come possiamo stimare la media della popolazione μ usando il campione di osservazioni descritto dalla tabella C.1? La media della popolazione è rappresentata dal valore atteso $E(Y) = \mu$ e il valore atteso di una variabile casuale è la media di tutti i suoi valori nella popolazione. Per analogia, sembra ragionevole stimarlo con la media dei valori nel campione, la **media campionaria**. Se indichiamo con

$y_1, y_2, \ldots, y_N$ il campione di N osservazioni, la media campionaria è definita da:

$$\bar{y} = \sum y_i / N \tag{C.3}$$

Il simbolo $\bar{y}$ (che si pronuncia "y barrato") è molto usato per indicare la media campionaria, come avete probabilmente già osservato nel corso dei vostri studi di statistica. La media delle misure di larghezza del bacino riportate nella tabella C.1 è data da $\bar{y} = 17{,}1582$. Possiamo dunque affermare che la stima della larghezza media del bacino nella popolazione è pari a 17,1582.

Data $\bar{y} = 17{,}1582$, chiediamoci quanto è buona questa stima di μ. In pratica ci stiamo chiedendo quanto sia vicino 17,1582 alla vera media della popolazione, μ. Sfortunatamente questa domanda è mal posta, nel senso che non può avere una risposta. Per averla dovremmo conoscere μ, ma in questo caso non ci sarebbe stato nessun bisogno di una stima!

Invece di chiederci se la *stima* è buona, chiediamoci se lo è la *procedura di stima*, detta **stimatore**. Quali sono le proprietà della media campionaria come stimatore della media di una popolazione? A questa domanda è possibile fornire una risposta. Per distinguere fra la stima e lo stimatore della media della popolazione μ indicheremo lo stimatore con:

$$\overline{Y} = \sum_{i=1}^{N} Y_i / N \tag{C.4}$$

Nella (C.4) abbiamo usato Y_i al posto di y_i per indicare il fatto che questa formula generale può essere usata qualunque sia il campione. In questo contesto le Y_i sono variabili casuali e di conseguenza lo è anche lo stimatore $\overline{Y}$.

Il valore dello stimatore $\overline{Y}$ è ignoto fino al momento in cui viene raccolto il campione, e campioni diversi forniscono valori di $\overline{Y}$ diversi. Per fare un esempio, raccogliamo altri 10 campioni di numerosità $N = 50$ e per ciascuno di essi calcoliamo la larghezza media del bacino; i risultati così ottenuti sono riportati nella tabella C.2. Le stime sono diverse da campione a campione perché $\overline{Y}$ è una variabile casuale. Questa variabilità, dovuta al fatto che sono stati raccolti campioni casuali diversi, è detta **variabilità campionaria**. Una caratteristica cruciale delle analisi statistiche è che lo stimatore $\overline{Y}$ – più in generale, qualsiasi procedura di stima statistica – presenta un certo grado di variabilità campionaria. Alla luce di questa osservazione, la funzione di densità o di probabilità di uno stimatore è definita come la sua **distribuzione campionaria**.

Per valutare le proprietà dello stimatore $\overline{Y}$ possiamo esaminarne il valore atteso, la varianza e la distribuzione campionaria.

Tabella C.2
Cambiamento di variabile: il caso discreto

Campione	$\bar{y}$
1	17,3544
2	16,8220
3	17,4114
4	17,1654
5	16,9004
6	16,9956
7	16,8368
8	16,7534
9	17,0974
10	16,8770

C.3.1. Valore atteso di $\overline{Y}$

Riscriviamo la formula (C.4) come:

$$\overline{Y} = \sum_{i=1}^{N} \frac{1}{N} Y_i = \frac{1}{N} Y_1 + \frac{1}{N} Y_2 + \ldots + \frac{1}{N} Y_N \tag{C.5}$$

Dalla (P.16) sappiamo che il valore atteso di questa somma è pari alla somma dei valori attesi:

$$\begin{aligned} E(\overline{Y}) &= E\left(\frac{1}{N}\,Y_1\right) + E\left(\frac{1}{N}\,Y_2\right) + \ldots + E\left(\frac{1}{N}\,Y_N\right) \\ &= \frac{1}{N}\,E(Y_1) + \frac{1}{N}\,E(Y_2) + \ldots + \frac{1}{N}\,E(Y_N) \\ &= \frac{1}{N}\,\mu + \frac{1}{N}\,\mu + \ldots + \frac{1}{N}\,\mu \\ &= \mu \end{aligned}$$

Il valore atteso dello stimatore $\overline{Y}$ coincide con la media della popolazione μ che stiamo cercando di stimare. Quali sono le implicazioni di questo risultato? Il valore atteso di una variabile casuale è la media dei suoi valori calcolata su un gran numero di replicazioni di un esperimento, che in questo caso corrisponde a raccogliere un gran numero di campioni casuali dalla popolazione. Se disponessimo di molti campioni di numerosità N e calcolassimo la media per ciascuno di essi, proprio come abbiamo fatto per la tabella C.2, la media di tutti *questi* valori sarebbe pari alla vera media nella popolazione, μ. Per uno stimatore questa è una proprietà importante. Gli stimatori con questa proprietà sono definiti **stimatori corretti**. La media campionaria $\overline{Y}$ è uno stimatore corretto della media nella popolazione μ.

Sfortunatamente, anche se per uno stimatore è una buona cosa essere corretto, la correttezza in sé non ci dice nulla sul fatto che la stima $\bar{y} = 17{,}1582$, basata su un unico campione di osservazioni, sia o meno vicina alla vera media nella popolazione μ. Per misurare quanto la stima possa essere lontana da μ calcoleremo la varianza dello stimatore.

C.3.2. Varianza di $\overline{Y}$

Per calcolare la varianza di $\overline{Y}$ useremo la formula (P.23) della varianza di una somma di variabili casuali incorrelate (con covarianza nulla). Se i dati sono stati ottenuti mediante un campionamento casuale questa proprietà è plausibile, perché in questo caso le osservazioni sono statisticamente indipendenti e dunque anche incorrelate. Un'altra ipotesi che abbiamo formulato è che $\mathrm{Var}(Y_i) = \sigma^2$ per tutte le osservazioni. Osservate attentamente il modo in cui sono utilizzate queste ipotesi nella derivazione della varianza di $\overline{Y}$, che indichiamo con $\mathrm{Var}(\overline{Y})$:

$$\begin{aligned} \mathrm{Var}(\overline{Y}) &= \mathrm{Var}\left(\frac{1}{N}\,Y_1 + \frac{1}{N}\,Y_2 + \ldots + \frac{1}{N}\,Y_N\right) \\ &= \frac{1}{N^2}\,\mathrm{Var}(Y_1) + \frac{1}{N^2}\,\mathrm{Var}(Y_2) + \ldots + \frac{1}{N^2}\,\mathrm{Var}(Y_N) \\ &= \frac{1}{N^2}\,\sigma^2 + \frac{1}{N^2}\,\sigma^2 + \ldots + \frac{1}{N^2}\,\sigma^2 \\ &= \frac{\sigma^2}{N} \end{aligned} \tag{C.6}$$

Questo risultato ci dice che (i) la varianza di $\overline{Y}$ è *minore* della varianza della popolazione, dato che la numerosità campionaria N è maggiore o uguale a 2, e (ii) quanto maggiore è N, tanto minore è la variabilità campionaria di $\overline{Y}$ misurata dalla sua varianza.

C.3.3. Distribuzione campionaria di $\overline{Y}$

Se nella popolazione i dati sono distribuiti secondo una distribuzione normale diremo che la variabile casuale Y_i segue una distribuzione normale. In questo caso anche lo stimatore $\overline{Y}$ ha la stessa distribuzione. Nella (P.30) abbiamo osservato che medie ponderate di variabili casuali normali sono anch'esse normali. Dalla (C.5) sappiamo che $\overline{Y}$ è una media ponderata delle Y_i. Se $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, allora anche $\overline{Y}$ ha distribuzione normale: $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/N)$.

Per comprendere meglio il significato e l'utilità del risultato $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/N)$ esaminiamo la figura C.2.

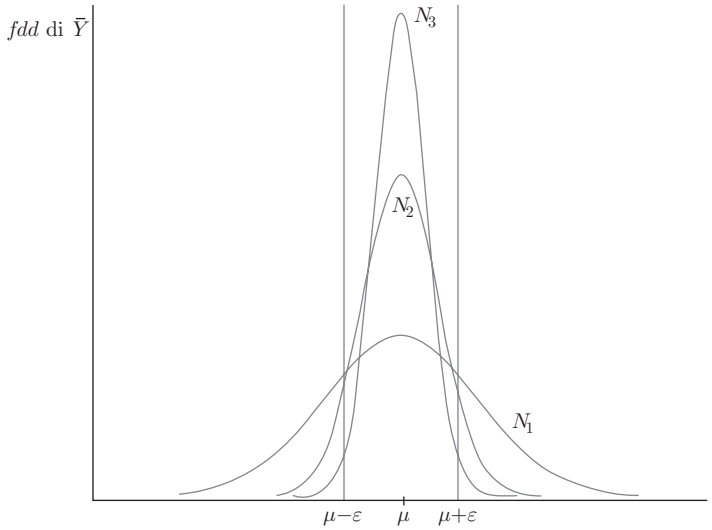

Figura C.2
Aumento della numerosità campionaria e distribuzioni campionarie di $\overline{Y}$.

In questa figura ciascuna delle distribuzioni normali è una distribuzione campionaria di $\overline{Y}$; ciò che le differenzia è la numerosità campionaria usata nella stima. Nella figura, $N_3 > N_2 > N_1$: aumentando la numerosità campionaria diminuisce la varianza dello stimatore $\overline{Y}$, $\text{Var}(\overline{Y}) = \sigma^2/N$, e questo fa crescere la probabilità che la media campionaria sia "vicina" al vero valore del parametro della popolazione μ. Nell'esaminare la figura C.2 tenete conto che l'area al di sotto di una funzione di densità (*fdd*) misura la probabilità che si verifichi l'evento corrispondente. Se ε è un numero positivo, la probabilità che $\overline{Y}$ cada nell'intervallo compreso fra $\mu - \varepsilon$ e $\mu + \varepsilon$ è maggiore per campioni più numerosi. Il messaggio di questa figura è che avere più dati è meglio che averne meno, perché un campione più numeroso aumenta la probabilità di ottenere una stima "vicina a", o "a una distanza minore di ε da", il vero parametro della popolazione μ.

Nel nostro esempio numerico, supponiamo di volere ottenere una stima di μ che si trovi a meno di un pollice (2,5 centimetri circa) dal vero valore. Calcoliamo la probabilità di ottenere una stima nel raggio di $\varepsilon = 1$ pollici da μ, o, in altre parole, all'interno dell'intervallo $[\mu - 1; \mu + 1]$. A titolo illustrativo, assumiamo che la popolazione sia normale, che $\sigma^2 = 10$ e che $N = 40$. In questo caso, $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/N = 10/40 = 0{,}25)$. Per misurare la probabilità che $\overline{Y}$ sia a meno di un pollice da μ possiamo calcolare $P(\mu - 1 \leq \overline{Y} \leq \mu + 1)$. Per farlo, standardizziamo $\overline{Y}$ sottraendo la media μ e dividendo la differenza per lo scarto quadratico medio $\sigma/\sqrt{N}$, e usiamo la proprietà della distribuzione normale standardizzata e

la tabella 1 dell'appendice D:

$$\begin{aligned} P(\mu - 1 \leq \overline{Y} \leq \mu + 1) &= P\left(-\frac{1}{\sigma/\sqrt{N}} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}} \leq \frac{1}{\sigma/\sqrt{N}}\right) \\ &= P\left(-\frac{1}{\sqrt{0{,}25}} \leq Z \leq \frac{1}{\sqrt{0{,}25}}\right) \\ &= P(-2 \leq Z \leq 2) = 0{,}9544 \end{aligned}$$

Con un campione casuale di numerosità $N = 40$ osservazioni tratte da una popolazione normale con varianza 10, dunque, lo stimatore della media campionaria fornirà una stima a meno di un pollice di distanza dal vero valore nel 95% circa delle volte. Se $N = 80$, la probabilità che $\overline{Y}$ si trovi a meno di un pollice da μ aumenta a 0,995.

C.3.4. Teorema del limite centrale

Per svolgere l'analisi alla fine del paragrafo precedente abbiamo dovuto assumere che la popolazione che stiamo esaminando, l'insieme delle misure del bacino degli adulti statunitensi, abbia distribuzione normale. Data la rilevanza di questa ipotesi è importante chiedersi quale sia la distribuzione campionaria della media $\overline{Y}$ se la popolazione non è normale. La risposta a questa domanda è fornita dal **teorema del limite centrale**.

TEOREMA DEL LIMITE CENTRALE: Se $Y_1, Y_2, \ldots, Y_N$ sono variabili casuali indipendenti e identicamente distribuite di media μ e varianza σ^2, e se $\overline{Y} = \sum Y_i/N$:

$$Z_N = \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}}$$

ha una distribuzione di probabilità che converge alla normale standardizzata $\mathcal{N}(0,1)$ per $N \to \infty$.

Questo teorema afferma che la media campionaria di N variabili casuali indipendenti, *qualunque* sia la loro distribuzione, ha approssimativamente distribuzione $\mathcal{N}(0,1)$ dopo essere stata standardizzata (in altre parole, dopo averle sottratto la media e diviso la differenza per lo scarto quadratico medio), a condizione che il campione sia sufficientemente numeroso. Una notazione sintetica per indicare questa proprietà è $\overline{Y} \overset{a}{\sim} \mathcal{N}(\mu, \sigma^2)$, dove il simbolo $\overset{a}{\sim}$ significa *asintoticamente distribuito*. Il termine **asintotico** implica che la normalità approssimata di $\overline{Y}$ dipende dalla disponibilità di un campione numeroso. Anche se la popolazione non è normale, dunque, è comunque possibile effettuare calcoli come quelli alla fine del paragrafo precedente a condizione che la numerosità campionaria sia sufficientemente elevata. Quanto deve essere numeroso il campione affinché il risultato del teorema sia utilizzabile? In generale la risposta dipende dalla complessità del problema, ma nel caso particolarmente semplice della stima della media di una popolazione la condizione $N \geq 30$ è sufficiente per ritenere con un certo grado di sicurezza che la media campionaria abbia effettivamente la distribuzione normale, $\overline{Y} \overset{a}{\sim} \mathcal{N}(\mu, \sigma^2)$, indicata dal teorema del limite centrale.

Per illustrare in che modo opera effettivamente il teorema del limite centrale effettuiamo un esperimento di simulazione. Supponiamo che la variabile casuale Y

abbia distribuzione triangolare, con funzione di densità:

$$f(y) = \begin{cases} 2y & 0 < y < 1 \\ 0 & \text{altrimenti} \end{cases}$$

Per capire il motivo del nome basta tracciare il grafico della *fdd* triangolare. Il valore atteso di Y è $\mu = E(Y) = 2/3$ e la sua varianza è $\sigma^2 = \text{Var}(Y) = 1/18$. Il teorema del limite centrale afferma che se $Y_1, \ldots, Y_N$ sono variabili casuali indipendenti e identicamente distribuite di densità $f(y)$:

$$Z_N = \frac{\overline{Y} - 2/3}{\sqrt{\frac{1/18}{N}}}$$

ha distribuzione di probabilità che tende a quella della normale standardizzata al crescere di N all'infinito.

Per generare valori casuali dalla *fdd* triangolare usiamo un generatore di numeri casuali. La figura C.3a illustra l'istogramma di 10 000 di questi valori. Creiamo 10 000 campioni di numerosità $N = 3$, 10 e 30, e calcoliamo le medie campionarie e i valori di Z_N per ciascun campione. Gli istogrammi di questi valori sono rappresentati nelle figure C.3b-d. Questi grafici illustrano bene la rapida convergenza della distribuzione della media campionaria standardizzata a una distribuzione a campana, centrata sullo zero, simmetrica e con quasi tutti i valori compresi fra -3 e 3, proprio come una distribuzione normale standardizzata, persino con una numerosità campionaria modesta come $N = 10$.

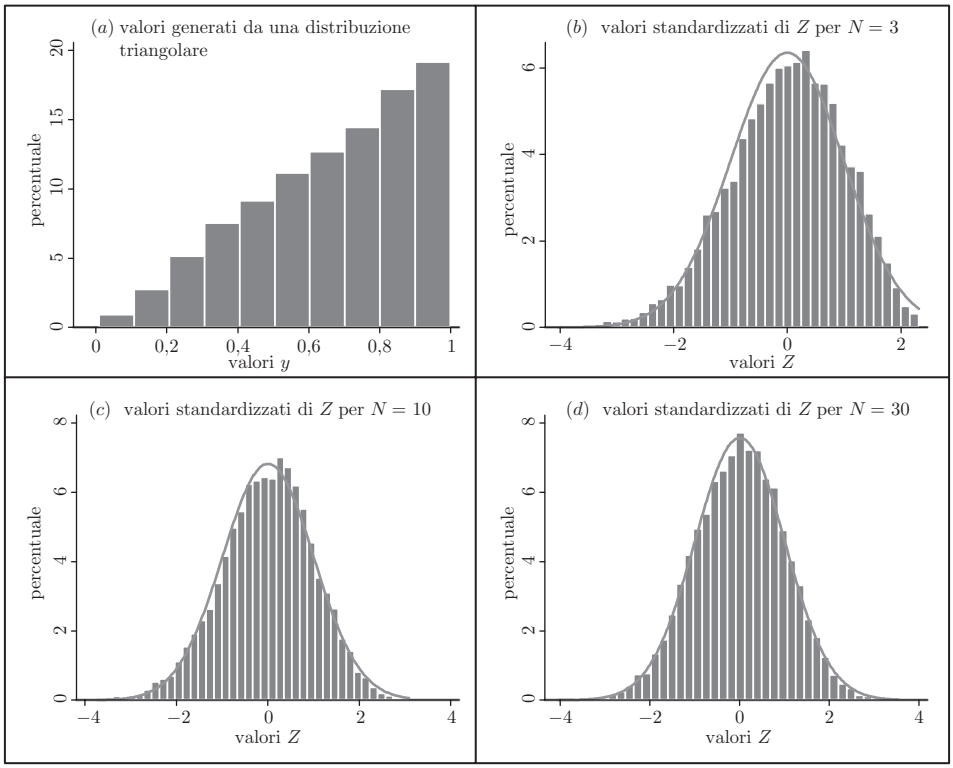

Figura C.3
Teorema del limite centrale.

C.3.5. Miglior stimatore lineare corretto

Un altro risultato interessante a proposito dello stimatore $\overline{Y}$ della media nella popolazione è che si tratta del migliore fra tutti i possibili stimatori che sono sia *lineari* sia *corretti*. Uno **stimatore** è **lineare** se è definito come una media ponderata delle Y_i, come per esempio $\widetilde{Y} = \sum a_i Y_i$, in cui i termini a_i sono delle costanti. La media campionaria $\overline{Y}$, definita dalla (C.4), è uno stimatore lineare con $a_i = 1/N$. Il fatto che $\overline{Y}$ sia il "migliore" stimatore lineare e corretto (BLUE, da *Best Linear Unbiased Estimator*) spiega perché sia così utilizzato. Per "migliore" intendiamo che si tratta dello stimatore lineare e corretto con varianza minima. Nel paragrafo precedente abbiamo dimostrato che gli stimatori con varianza bassa sono preferibili a quelli con varianza elevata perché questa caratteristica aumenta le possibilità di ottenere una stima vicina al vero valore della media della popolazione μ. Questa proprietà così importante dello stimatore $\overline{Y}$ è valida *se* i valori campionari $Y_i \sim (\mu, \sigma^2)$ sono identicamente distribuiti e fra loro incorrelati, ma non dipende dal fatto che la popolazione abbia distribuzione normale. Il paragrafo C.9.2 propone una dimostrazione di questo risultato.

C.4. Stima della varianza e di altri momenti della popolazione

La media campionaria $\overline{Y}$ è una stima della media nella popolazione ed è spesso chiamata "momento primo" perché si tratta del valore atteso di Y elevato a potenza 1. Possiamo definire momenti di ordine più elevato considerando il valore atteso di potenze più elevate della variabile casuale; il momento secondo di Y è dunque dato da $E(Y^2)$, il momento terzo da $E(Y^3)$ e così via. Quando alla variabile casuale viene sottratta la sua media nella popolazione, la variabile è detta *centrata* e i valori attesi di potenze di variabili casuali centrate sono chiamati **momenti centrati**, spesso indicati con μ_r. L'r-esimo momento centrato di Y è dunque dato da:

$$\mu_r = E[(Y - \mu)^r]$$

Il valore del primo momento centrato è 0, dato che $\mu_1 = E[(Y-\mu)^1] = E(Y)-\mu = 0$. I momenti centrati di Y più interessanti sono quelli di ordine superiore a 1:

$$\begin{aligned} \mu_2 &= E[(Y - \mu)^2] = \sigma^2 \\ \mu_3 &= E[(Y - \mu)^3] \\ \mu_4 &= E[(Y - \mu)^4] \end{aligned}$$

Come potete osservare, il momento secondo centrato di Y è la sua varianza e i momenti di ordine 3 e 4 compaiono nelle definizioni di asimmetria e curtosi introdotte nell'appendice B.1.2. La domanda che ci poniamo in questo paragrafo è la seguente: ora che abbiamo un eccellente stimatore della media di una popolazione, come possiamo stimare questi momenti di ordine più elevato? Inizieremo considerando la stima della varianza della popolazione e passeremo poi al problema di stimare i momenti di ordine terzo e quarto.

C.4.1. Stima della varianza della popolazione

La varianza nella popolazione è data da $\mathrm{Var}(Y) = \sigma^2 = E[(Y - \mu)^2]$. Dato che un valore atteso è una specie di "media", se conoscessimo μ potremmo stimare la

varianza usando il corrispondente momento campionario $\tilde{\sigma}^2 = \sum(Y_i - \mu)^2/N$. Non conoscendo μ, useremo al suo posto lo stimatore $\overline{Y}$:

$$\tilde{\sigma}^2 = \frac{\sum(Y_i - \overline{Y})^2}{N}$$

Questo stimatore non è male: ha un fondamento logico ed è possibile dimostrare che al crescere della numerosità campionaria all'infinito, $N \to \infty$, esso converge al vero valore di σ^2; purtroppo però è distorto, cioè non corretto. Per renderlo corretto dobbiamo dividere per $N - 1$, anziché per N. Il motivo di questa correzione è che prima di stimare la varianza è necessario stimare la media nella popolazione μ. In campioni di almeno 30 osservazioni questa modifica non ha un effetto rilevante, ma in campioni più piccoli fa una certa differenza. Lo stimatore corretto della varianza della popolazione σ^2 è dunque dato da:

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \overline{Y})^2}{N - 1} \tag{C.7}$$

Forse avete già incontrato questo stimatore in un primo corso di statistica con il nome di "varianza campionaria". Usando la varianza campionaria possiamo stimare la varianza dello stimatore $\overline{Y}$ come:

$$\widehat{\mathrm{Var}(\overline{Y})} = \hat{\sigma}^2/N \tag{C.8}$$

Si noti che in (C.8) abbiamo messo un "cappello" ($\widehat{\ }$) sopra la varianza per indicare che si tratta di una sua stima. La radice quadrata di questa stima è chiamata **standard error** di $\overline{Y}$, **standard error della media** o **standard error della stima**:

$$\mathrm{se}(\overline{Y}) = \sqrt{\widehat{\mathrm{Var}(\overline{Y})}} = \hat{\sigma}/\sqrt{N} \tag{C.9}$$

C.4.2. Stima dei momenti di ordine superiore

Si ricordi che i momenti centrati sono valori attesi, $\mu_r = E[(Y - \mu)^r]$ e che di conseguenza si tratta di medie relative alla popolazione. In statistica la **legge dei grandi numeri** afferma che al divergere della numerosità campionaria ($N \to \infty$) le medie campionarie convergono alle medie (valori attesi) nella popolazione. Possiamo perciò stimare i momenti di ordine superiore utilizzando il corrispondente momento campionario e sostituendo la media nella popolazione μ con la sua stima $\overline{Y}$:

$$\begin{aligned}
\tilde{\mu}_2 &= \sum(Y_i - \overline{Y})^2/N = \tilde{\sigma}^2 \\
\tilde{\mu}_3 &= \sum(Y_i - \overline{Y})^3/N \\
\tilde{\mu}_4 &= \sum(Y_i - \overline{Y})^4/N
\end{aligned}$$

Si noti che in questi calcoli dividiamo per N e non per $N - 1$, dato che per giustificare questi stimatori stiamo usando la legge dei grandi numeri (in altre parole, stiamo assumendo che la numerosità campionaria sia elevata) e che in grandi campioni la correzione del denominatore avrebbe un effetto irrisorio. Usando le stime campionarie dei momenti di ordine superiore possiamo calcolare stime del

coefficiente di asimmetria (A) e di curtosi (C):

$$\widehat{\text{asimmetria}} = A = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3}$$
$$\widehat{\text{curtosi}} = C = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$$

C.4.3. Un esempio: i dati sulla larghezza del bacino

Per queste osservazioni la varianza campionaria è data da:

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \overline{Y})^2}{N-1} = \frac{\sum(Y_i - 17{,}1582)^2}{49} = \frac{159{,}9995}{49} = 3{,}2653$$

Ciò significa che la stima della varianza della media campionaria è:

$$\widehat{\text{Var}(\overline{Y})} = \frac{\hat{\sigma}^2}{N} = \frac{3{,}2653}{50} = 0{,}0653$$

e che il suo standard error è:

$$\text{se}(\overline{Y}) = \hat{\sigma}/\sqrt{N} = 2{,}2556$$

La stima dell'indice di asimmetria è $A = -0{,}0138$ e quella dell'indice di curtosi è $C = 2{,}3315$; questi valori sono stati calcolati usando:

$$\begin{aligned}\tilde{\sigma} &= \sqrt{\sum(Y_i - \overline{Y})^2/N} = \sqrt{159{,}9995/50} = 1{,}7889\\ \tilde{\mu}_3 &= \sum(Y_i - \overline{Y})^3/N = -0{,}0791\\ \tilde{\mu}_4 &= \sum(Y_i - \overline{Y})^4/N = 23{,}8748\end{aligned}$$

I dati sulla larghezza del bacino sono dunque caratterizzati da una leggera asimmetria negativa e le code della loro distribuzione sono meno spesse di quelle di una distribuzione normale. Ciò nonostante, come vedremo nel paragrafo C.7.4, non possiamo per questo concludere che i dati siano stati generati da una distribuzione diversa dalla normale.

C.4.4. Uso delle stime

Come possiamo riepilogare ciò che abbiamo appreso fino a questo punto? Le nostre stime suggeriscono che la larghezza del bacino di uno adulto statunitense ha distribuzione normale di media 17,1582 e varianza 3,265 pollici: $Y \sim \mathcal{N}(17{,}1582; 3{,}265)$. Sulla base di questa informazione, qual è la percentuale di clienti che non riuscirà a sedersi se i sedili di un aeroplano sono larghi 18 pollici? Possiamo riformulare la domanda chiedendoci quale sia la probabilità che il bacino di un individuo scelto a caso sia più largo di 18 pollici:

$$P(Y > 18) = P\left(\frac{Y-\mu}{\sigma} > \frac{18-\mu}{\sigma}\right)$$

Possiamo dare una risposta approssimativa a questa domanda sostituendo i parametri ignoti con le loro stime:

$$\widehat{P(Y > 18)} \approx P\left(\frac{Y-\bar{y}}{\hat{\sigma}} > \frac{18 - 17{,}1582}{1{,}8070}\right) = P(Z > 0{,}4659) = 0{,}3207$$

Sulla base delle nostre stime, il 32% della popolazione non sarà in grado di sedersi in un sedile largo 18 pollici.

Quanto dovrebbero essere larghi i sedili per riuscire a contenere il 95% della popolazione? Se indichiamo con y^* l'ampiezza di seduta che cerchiamo:

$$\widehat{P(Y \leq y^*)} \approx P\left(\frac{Y - \bar{y}}{\hat{\sigma}} \leq \frac{y^* - 17{,}1582}{1{,}8070}\right) = P\left(Z \leq \frac{y^* - 17{,}1582}{1{,}8070}\right) = 0{,}95$$

Usando un software econometrico o consultando le tavole della distribuzione normale, otteniamo che $P(Z \leq z^*) = 0{,}95$ se $z^* = 1{,}645$. Di conseguenza:

$$\frac{y^* - 17{,}1582}{1{,}8070} = 1{,}645 \;\Rightarrow\; y^* = 20{,}1305$$

Per riuscire a far sedere il 95% dei passeggeri statunitensi adulti, dunque, le nostre stime suggeriscono che l'ampiezza dei sedili dovrebbe essere leggermente maggiore di 20 pollici (50 centimetri circa).

C.5. Stima intervallare

Diversamente dalla stima puntuale $\bar{y} = 17{,}1582$ della media nella popolazione μ, un intervallo di confidenza, o stima intervallare, è un intervallo di valori che può contenere la vera media della popolazione. Un intervallo di confidenza non riflette solo l'informazione sulla posizione di μ, ma anche sulla precisione con la quale possiamo stimare questo parametro.

C.5.1. Stima intervallare: σ^2 nota

Sia Y una variabile casuale con distribuzione normale, $Y \sim \mathcal{N}(\mu, \sigma^2)$. Assumiamo di disporre di un campione casuale di numerosità N tratto da questa popolazione, $Y_1, Y_2, \ldots, Y_N$. Lo stimatore della media nella popolazione è $\overline{Y} = \sum_{i=1}^{N} Y_i/N$. Dato che per ipotesi Y ha distribuzione normale, lo stesso è vero anche per la media campionaria: $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/N)$.

Per il momento assumiamo inoltre che la varianza della popolazione σ^2 sia nota. Questa ipotesi difficilmente potrà essere vera, ma formulandola possiamo introdurre il concetto di intervallo di confidenza in un contesto relativamente semplice. Nel prossimo paragrafo introdurremo i metodi da utilizzare nel caso in cui σ^2 sia ignota. Iniziamo dalla variabile casuale normale standardizzata:

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/N}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1) \tag{C.10}$$

Per calcolare probabilità relative alla normale standardizzata possiamo utilizzare la sua funzione di ripartizione (si veda il paragrafo P.6 del Piccolo manuale di probabilità):

$$P(Z \leq z) = \Phi(z)$$

Questi valori sono riportati nella tabella 1 dell'appendice D. Indichiamo con z_c un "valore critico" per la distribuzione normale standardizzata tale che $\alpha = 0{,}05$ sia la probabilità nelle code della distribuzione, con probabilità $\alpha/2 = 0{,}025$ nella coda a destra di z_c e probabilità $\alpha/2 = 0{,}025$ nella coda a sinistra di $-z_c$. Il valore critico è il percentile di livello 97,5% della distribuzione normale standardizzata, $z_c = 1{,}96$, con $\Phi(1{,}96) = 0{,}975$, ed è illustrato nella figura C.4. Abbiamo dunque che $P(Z \geq 1{,}96) = P(Z \leq -1{,}96) = 0{,}025$ e:

$$P(-1{,}96 \leq Z \leq 1{,}96) = 1 - 0{,}05 = 0{,}95 \tag{C.11}$$

Sostituiamo (C.10) nella (C.11) e, con qualche passaggio algebrico, otteniamo:

$$P\left(\overline{Y} - 1{,}96\,\sigma/\sqrt{N} \leq \mu \leq \overline{Y} + 1{,}96\,\sigma/\sqrt{N}\right) = 0{,}95$$

In generale:

$$P\left(\overline{Y} - z_c\,\frac{\sigma}{\sqrt{N}} \leq \mu \leq \overline{Y} + z_c\,\frac{\sigma}{\sqrt{N}}\right) = 1 - \alpha \tag{C.12}$$

dove z_c è il valore critico corrispondente a un livello di probabilità α tale che $\Phi(z_c) = 1 - \alpha/2$. Nella (C.12) abbiamo definito lo stimatore intervallare:

$$\overline{Y} \pm z_c\,\frac{\sigma}{\sqrt{N}} \tag{C.13}$$

La (C.13) rappresenta uno *stimatore* intervallare perché in campioni ripetuti estratti dalla stessa popolazione gli intervalli costruiti in questo modo conterranno la vera media della popolazione μ il $100\alpha\%$ delle volte.

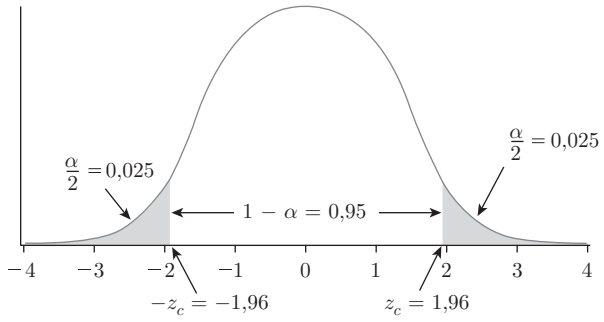

Figura C.4
Valori critici della distribuzione $\mathcal{N}(0, 1)$ per un livello di significatività $\alpha = 0{,}05$.

C.5.2. Una simulazione

Per usare lo stimatore intervallare (C.13) sono necessarie osservazioni tratte da una distribuzione normale di varianza nota. Per illustrare il calcolo e il significato di stima intervallare genereremo un campione di osservazioni usando una simulazione al computer. Tutti i software statistici contengono generatori di numeri casuali, cioè comandi in grado di generare valori in accordo con una data distribuzione di probabilità. La tabella C.3 (*table_c3.dat*) contiene 30 valori casuali estratti da una popolazione normale di media $\mu = 10$ e varianza $\sigma^2 = 10$.

La media campionaria di questi valori è $\bar{y} = 10{,}206$ e la corrispondente stima intervallare di μ, ottenuta applicando ai dati lo stimatore intervallare (C.13) con un

Tabella C.3
30 valori casuali generati dalla distribuzione $\mathcal{N}(10, 10)$

11,939	11,407	13,809
10,706	12,157	7,443
6,644	10,829	8,855
13,187	12,368	9,461
8,433	10,052	2,439
9,210	5,036	5,527
7,961	14,799	9,921
14,921	10,478	11,814
6,223	13,859	13,403
10,123	12,355	10,819

livello di probabilità 0,95, è data da $10{,}206 \pm 1{,}96 \times \sqrt{10/30} = (9{,}074;\ 11{,}338)$. Per capire in che cosa consista la variabilità campionaria di uno stimatore intervallare consideriamo la tabella C.4, che contiene la stima intervallare per il campione nella tabella C.3 ma anche medie campionarie e stime intervallari per altri 9 campioni di numerosità 30, proprio come quello nella tabella C.3. I 10 campioni sono contenuti nel file *table_c4.dat*.

Tabella C.4
Stime intervallari al 95% per 10 campioni di osservazioni

Campione	$\bar{y}$	Estremo inferiore	Estremo superiore
1	10,206	9,074	11,338
2	9,828	8,696	10,959
3	11,194	10,063	12,326
4	8,822	7,690	9,953
5	10,434	9,303	11,566
6	8,855	7,723	9,986
7	10,511	9,380	11,643
8	9,212	8,080	10,343
9	10,464	9,333	11,596
10	10,142	9,010	11,273

La tabella C.4 illustra la variabilità campionaria dello stimatore $\overline{Y}$. La media campionaria varia da un campione all'altro. In questa simulazione, o esperimento Monte Carlo, conosciamo il vero valore della media nella popolazione, $\mu = 10$ e le stime $\bar{y}$ sono centrate attorno a esso. La semi-ampiezza delle stime intervallari è $1{,}96\sigma/\sqrt{N}$. Si noti che mentre le stime puntuali $\bar{y}$ nella tabella C.4 cadono in prossimità del vero valore $\mu = 10$, non tutte le stime intervallari lo contengono. Gli intervalli ottenuti sui campioni 3, 4 e 6 non includono il vero valore $\mu = 10$. In 10 000 campioni simulati, tuttavia, la media di $\bar{y}$ è 10,004 e il 94,86% degli intervalli costruiti usando (C.13) contiene il vero valore del parametro $\mu = 10$.

Questi risultati mettono in luce ciò che possiamo e ciò che non possiamo affermare a proposito delle proprietà delle stime intervallari.

- Una stima intervallare qualsiasi può contenere o non contenere il vero valore del parametro nella popolazione.
- Usando *molti* campioni di numerosità N, il 95% degli intervalli costruiti applicando a ciascuno di essi la (C.13) con $(1 - \alpha) = 0{,}95$ conterrà il vero valore del parametro.
- Un livello di "confidenza" del 95% è la probabilità che lo stimatore intervallare fornirà un intervallo contenente il vero valore del parametro. La fiducia che nutriamo riguarda la procedura, non una stima intervallare qualsiasi.

Dato che il 95% degli intervalli costruiti usando la (C.13) contiene il vero valore del parametro $\mu = 10$, ci sorprenderebbe scoprire che il vero parametro cada al di fuori della stima intervallare costruita su un particolare campione. In effetti, il fatto che tre dei 10 intervalli nella tabella C.4 non contengano $\mu = 10$ è sorprendente, dato che su 10 campioni tenderemmo a pensare che il numero di stime intervallari al 95% che non contengono il vero μ dovrebbe essere al massimo pari a 1. Questo risultato tuttavia ci mostra che quello che può accadere con un campione qualsiasi, o con un numero ridotto di campioni, non corrisponde a ciò che intendiamo per proprietà campionarie. Queste proprietà ci dicono ciò che accade quando il numero di replicazioni dell'esperimento è molto elevato.

C.5.3. Stima intervallare: σ^2 ignota

La standardizzazione operata dalla (C.10) assume che la varianza nella popolazione σ^2 sia nota. Quando σ^2 è ignota è naturale sostituirla con lo stimatore $\hat{\sigma}^2$ definito dalla (C.7):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}$$

Questa sostituzione fa sì che la nuova variabile casuale standardizzata abbia distribuzione t (si veda l'appendice B.3.7) con $(N-1)$ gradi di libertà:

$$t = \frac{\overline{Y} - \mu}{\hat{\sigma}\sqrt{N}} \sim t_{(N-1)} \tag{C.14}$$

La notazione $t_{(N-1)}$ indica una distribuzione t con $N-1$ "gradi di libertà". Indichiamo con t_c il percentile $t_{(1-\alpha/2;\, N-1)}$ di livello $100(1-\alpha/2)\%$. Questo valore critico ha la proprietà che $P[t_{(N-1)} \leq t_{(1-\alpha/2;\, N-1)}] = 1-\alpha/2$. I valori critici della distribuzione t sono riportati nella tabella 2 dell'appendice D. Se t_c è un valore critico della distribuzione t:

$$P\left(-t_c \leq \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \leq t_c\right) = 1 - \alpha$$

Con qualche passaggio algebrico otteniamo:

$$P\left(\overline{Y} - t_c \frac{\hat{\sigma}}{\sqrt{N}} \leq \mu \leq \overline{Y} + t_c \frac{\hat{\sigma}}{\sqrt{N}}\right) = 1 - \alpha$$

Lo stimatore intervallare di μ al $100(1-\alpha)\%$ è dato da:

$$\overline{Y} \pm t_c \frac{\hat{\sigma}}{\sqrt{N}} \quad \text{o} \quad \overline{Y} \pm t_c \text{se}(\overline{Y}) \tag{C.15}$$

A differenza dello stimatore intervallare (C.13) valido nel caso di σ^2 nota, l'intervallo (C.15) ha centro *e ampiezza* che variano da un campione all'altro.

NOTA: L'intervallo di confidenza (C.15) è basato sull'ipotesi che la popolazione abbia distribuzione normale, in modo che anche $\overline{Y}$ abbia la stessa distribuzione. Se la popolazione non è normale possiamo invocare il teorema del limite centrale e affermare che in "grandi" campioni $\overline{Y}$ ha distribuzione approssimativamente normale; la figura C.3 ci ha mostrato che un campione può essere considerato "grande" anche se contiene solo 30 osservazioni. In questo caso useremo la (C.15) riconoscendo la possibilità di un errore di approssimazione, soprattutto in campioni poco numerosi.

C.5.4. Una simulazione (continua)

La tabella C.5 contiene le stime di σ^2 e le relative stime intervallari costruite usando la (C.15) per gli stessi 10 campioni usati per costruire la tabella C.4. Con una numerosità campionaria $N = 30$ e un livello di confidenza del 95%, il valore critico della distribuzione t è dato da $t_c = t_{(0{,}975;\, 29)} = 2{,}045$. Le stime $\bar{y}$ sono le

Tabella C.5
Stime intervallari usando (C.15) per 10 campioni di osservazioni

Campione	$\bar{y}$	$\hat{\sigma}^2$	Estremo inferiore	Estremo superiore
1	10,206	9,199	9,073	11,338
2	9,828	6,876	8,849	10,807
3	11,194	10,330	9,994	12,394
4	8,822	9,867	7,649	9,995
5	10,434	7,985	9,379	11,489
6	8,855	6,230	7,923	9,787
7	10,511	7,333	9,500	11,523
8	9,212	14,687	7,781	10,643
9	10,464	10,414	9,259	11,669
10	10,142	17,689	8,571	11,712

stesse della tabella C.4; quelle della varianza, $\hat{\sigma}^2$, variano attorno al vero valore $\sigma^2 = 10$. Di questi 10 intervalli, quelli relativi ai campioni 4 e 6 non contengono il vero valore del parametro $\mu = 10$. Ciò nonostante, su 10 000 campioni simulati il 94,82% di essi contiene la vera media nella popolazione.

C.5.5. Stima intervallare usando i dati sulla larghezza del bacino

Nei paragrafi precedenti abbiamo un ingegnere cui viene posto il problema empirico di progettare i sedili di un aeroplano. Dato un campione casuale di numerosità $N = 50$, abbiamo stimato che l'ampiezza media del bacino di un adulto statunitense è $\bar{y} = 17{,}1582$ pollici; la stima della varianza della popolazione, inoltre, è $\hat{\sigma}^2 = 3{,}265$ e di conseguenza la stima dello scarto quadratico medio è $\hat{\sigma} = 1{,}807$. Lo standard error della media è $\hat{\sigma}/\sqrt{N} = 1{,}807/\sqrt{50} = 0{,}2556$. Il valore critico per la stima intervallare proviene da una distribuzione t con $N-1 = 49$ gradi di libertà. Questo valore non è riportato nella tabella 2 dell'appendice D, ma il nostro software indica un valore critico esatto $t_c = t_{(0{,}975;\,49)} = 2{,}0095752$, che arrotondiamo a $t_c = 2{,}01$. Per costruire una stima intervallare al 95% usiamo la (C.15), sostituendo agli stimatori le rispettive stime e ottenendo:

$$\begin{aligned} \bar{y} \pm t_c \frac{\hat{\sigma}}{\sqrt{N}} &= 17{,}1582 \pm 2{,}01 \frac{1{,}807}{\sqrt{50}} \\ &= [16{,}6447;\ 17{,}6717] \end{aligned}$$

Questa stima suggerisce che la media della popolazione cade nell'intervallo compreso fra 16,645 e 17,672 pollici. Anche se non possiamo essere certi che questo intervallo contenga la vera larghezza media del bacino, sappiamo che la procedura usata per calcolarlo "funziona" il 95% delle volte; di conseguenza saremmo sorpresi di scoprire che l'intervallo ottenuto non contenga la vera media nella popolazione μ.

C.6. Verifica d'ipotesi sulla media di una popolazione

Le procedure di verifica d'ipotesi mettono a confronto una congettura o ipotesi sulla popolazione con l'informazione contenuta in un campione di osservazioni. Le congetture che verificheremo in questo paragrafo riguardano la media di una popolazione normale. Nell'esempio del problema affrontato dal progettista di sedili, supponiamo che a partire dal 1970 gli aeroplani siano stati costruiti assumendo una larghezza media del bacino nella popolazione di 16,5 pollici. Questo numero può essere considerato valido ancora oggi?

C.6.1. Componenti di una verifica d'ipotesi

Le verifiche d'ipotesi usano l'informazione campionaria su un parametro – più precisamente, la sua stima puntuale e il suo standard error – per trarre una conclusione a proposito dell'ipotesi. In ogni procedura di test devono essere presenti cinque componenti:

COMPONENTI DI UNA VERIFICA D'IPOTESI

- Un'ipotesi *nulla* H_0
- Un'ipotesi *alternativa* H_1
- Una *statistica* test
- Una regione di *rifiuto*
- Una conclusione

C.6.1.a. Ipotesi nulla

L'ipotesi "nulla", indicata con H_0 (*H-zero*), specifica un valore c per un parametro. L'ipotesi nulla è formulata come $H_0 : \mu = c$. Un'ipotesi nulla è una congettura che viene ritenuta valida a meno che l'evidenza empirica non ci convinca della sua falsità, nel qual caso l'ipotesi nulla viene *rifiutata*.

C.6.1.b. Ipotesi alternativa

A ogni ipotesi nulla è associata un'ipotesi a essa logicamente alternativa, H_1, che verrà accettata se l'ipotesi nulla viene rifiutata. L'ipotesi alternativa è flessibile e dipende in certa misura dal problema esaminato. Per l'ipotesi nulla $H_0 : \mu = c$ sono possibili tre alternative:

- $H_1 : \mu > c$. Se rifiutiamo l'ipotesi nulla $\mu = c$, accettiamo l'alternativa che μ sia maggiore di c.
- $H_1 : \mu < c$. Se rifiutiamo l'ipotesi nulla $\mu = c$, accettiamo l'alternativa che μ sia minore di c.
- $H_1 : \mu \neq c$. Se rifiutiamo l'ipotesi nulla $\mu = c$, accettiamo l'alternativa che il valore di μ sia diverso da (non uguale a) quello di c.

C.6.1.c. Statistica test

L'informazione campionaria sull'ipotesi nulla è incorporata nel valore campionario di una **statistica test**. Sulla base del valore di quest'ultima decideremo se rifiutare o non rifiutare l'ipotesi nulla. Una statistica test ha una caratteristica molto particolare: la sua distribuzione di probabilità è completamente nota se l'ipotesi nulla è vera; in caso contrario la distribuzione è ignota ma comunque diversa da quella sotto H_0.

Consideriamo l'ipotesi nulla $H_0 : \mu = c$. Se il campione proviene da una distribuzione normale di media μ e varianza σ^2:

$$t = \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)}$$

Se l'ipotesi nulla $H_0 : \mu = c$ è vera:

$$t = \frac{\overline{Y} - c}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)} \tag{C.16}$$

Se l'ipotesi nulla non è vera, la statistica t nella (C.16) non ha la consueta distribuzione t.

NOTA: La distribuzione della statistica test nella (C.16) è basata sull'ipotesi che la popolazione abbia distribuzione normale. Se la popolazione non è normale possiamo invocare il teorema del limite centrale e affermare che in "grandi" campioni $\overline{Y}$ ha distribuzione approssimativamente normale. In questo caso useremo la (C.16) riconoscendo la possibilità di un errore di approssimazione, soprattutto in campioni poco numerosi.

C.6.1.d. Regione di rifiuto

La regione di rifiuto dipende dalla forma dell'ipotesi alternativa e consiste nell'intervallo di valori della statistica test che induce a rifiutare l'ipotesi nulla. Questi valori sono *improbabili* (hanno probabilità bassa di essere osservati) se l'ipotesi nulla è vera. La logica sottostante è la seguente: se otteniamo un valore della statistica test che appartiene a una regione con bassa probabilità, è poco verosimile che la statistica test abbia la distribuzione ipotizzata e, di conseguenza, è improbabile che l'ipotesi nulla sia vera. Se l'ipotesi alternativa è vera, i valori della statistica test tendono a essere anormalmente elevati o anormalmente piccoli. Il significato esatto di "elevati" e "piccoli" è deciso dalla scelta di una probabilità α, chiamata **livello di significatività** del test, che precisa che cosa intendiamo per "evento *improbabile*". Il livello di significatività del test α viene di solito scelto pari a 0,01, 0,05 o 0,10.

C.6.1.e. Conclusione

Per completare una verifica d'ipotesi bisogna formulare una conclusione che consiste nella decisione se rifiutare o meno l'ipotesi nulla. Vi raccomandiamo tuttavia di abituarvi a spiegare bene che cosa significhi la vostra conclusione nel contesto del problema economico che state esaminando – in altre parole, a interpretare i risultati in maniera comprensibile da un punto di vista economico. Questo passaggio dovrebbe rappresentare uno snodo cruciale in qualsiasi analisi statistica stiate intraprendendo.

Passiamo ora a discutere i passaggi necessari allo svolgimento delle diverse versioni di una verifica d'ipotesi.

C.6.2. Test a una coda con alternativa "maggiore di" (>)

Se è vera l'ipotesi alternativa $H_1 : \mu > c$, il valore della statistica test t definita dalla (C.16) tende a essere "più grande del normale" per la distribuzione t. Indichiamo con t_c il percentile di livello $100(1-\alpha)\%$ di una distribuzione t con $N-1$ gradi di libertà, $t_{(1-\alpha;\, N-1)}$. In questo caso $P(t \leq t_c) = 1 - \alpha$, dove α è il livello di significatività del test. Se la statistica t è maggiore o uguale di t_c, rifiutiamo $H_0 : \mu = c$ e accettiamo l'alternativa $H_1 : \mu > c$. Questa situazione è illustrata dalla figura C.5.

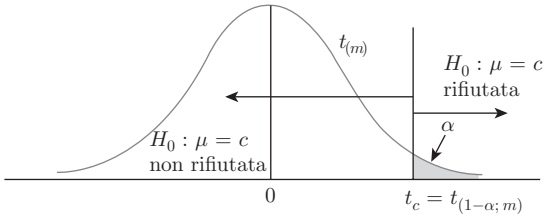

Figura C.5
Regione di rifiuto per un test a una coda di $H_0 : \mu = c$ rispetto a $H_1 : \mu > c$.

Se l'ipotesi nulla $H_0 : \mu = c$ è *vera*, la statistica test (C.16) ha distribuzione t e i suoi valori tenderanno a concentrarsi nella sua regione centrale, dove si trova la maggior parte della probabilità. Se $t < t_c$, non esiste evidenza empirica contraria all'ipotesi nulla e di conseguenza essa non viene rifiutata.

C.6.3. Test a una coda con alternativa "minore di" ($<$)

Se è vera l'ipotesi alternativa $H_1 : \mu < c$, il valore della statistica test t definita dalla (C.16) tende a essere "più piccolo del normale" per la distribuzione t. Il valore critico $-t_c$ è il percentile di livello $100\alpha\%$ di una distribuzione t con $N-1$ gradi di libertà, $t_{(\alpha;\, N-1)}$. In questo caso $P(t \leq -t_c) = \alpha$, dove α è il livello di significatività del test. Questa situazione è illustrata dalla figura C.6. Se $t \leq -t_c$, rifiutiamo $H_0 : \mu = c$ e accettiamo l'alternativa $H_1 : \mu < c$. Se $t > -t_c$, non rifiutiamo H_0.

STRATAGEMMA MNEMONICO: Per un test a una coda la regione di rifiuto si trova nella direzione indicata dall'ipotesi alternativa. Se l'alternativa è del tipo "$>$", si rifiuta nella coda destra; se l'alternativa è del tipo "$<$", si rifiuta nella coda sinistra.

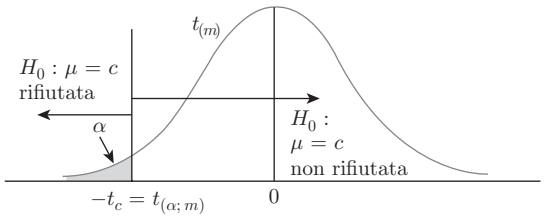

Figura C.6
Regione di rifiuto per un test a una coda di $H_0 : \mu = c$ rispetto a $H_1 : \mu < c$.

C.6.4. Test a due code con alternativa "diverso da" ($\neq$)

Se è vera l'ipotesi alternativa $H_1 : \mu \neq c$, il valore della statistica test t definita dalla (C.16) tende a essere "più piccolo o più grande del normale" per la distribuzione t. La regione di rifiuto è composta dalle due "code" della distribuzione t e questo test è detto "a due code". La figura C.7 illustra i valori critici del test di $H_0 : \mu = c$ rispetto a $H_1 : \mu \neq c$. Il valore critico è il percentile di livello $100(1-\alpha/2)\%$ di una distribuzione t con $N-1$ gradi di libertà, $t_{(1-\alpha/2;\, N-1)}$. In questo caso $P(t \geq t_c) = P(t \leq -t_c) = \alpha/2$.

Se il valore della statistica test t cade nella regione di rifiuto, cioè in una delle code della distribuzione $t_{(N-1)}$, rifiutiamo l'ipotesi nulla $H_0 : \mu = c$ e accettiamo

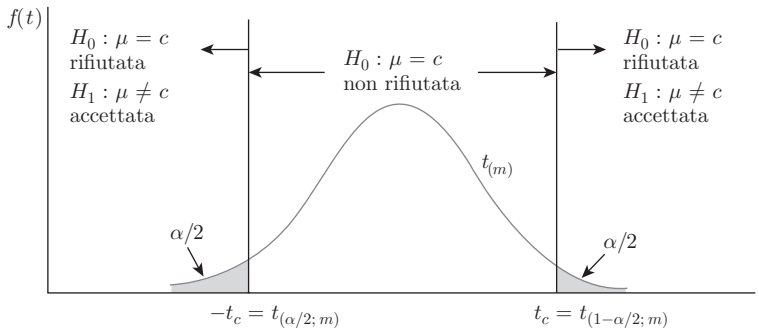

Figura C.7
Regione di rifiuto per un test di $H_0 : \mu = c$ rispetto a $H_1 : \mu \neq c$.

l'alternativa $H_1 : \mu \neq c$. Se il valore della statistica t cade nella regione di non rifiuto, compresa fra i valori critici $-t_c$ e t_c, non rifiutiamo l'ipotesi nulla $H_0 : \mu = c$.

C.6.5. Esempio di test a una coda usando i dati sulla larghezza del bacino

Come esempio consideriamo il test dell'ipotesi nulla che la larghezza media del bacino nella popolazione composta dagli adulti statunitensi sia di 16,5 pollici rispetto all'alternativa che sia *maggiore* di questo valore. Nell'effettuare un test vi consigliamo di seguire sempre il seguente schema in cinque passaggi.

1. L'ipotesi nulla è $H_0 : \mu = 16{,}5$. L'ipotesi alternativa è $H_1 : \mu > 16{,}5$.
2. La statistica test è $t = (\overline{Y} - 16{,}5)/(\hat{\sigma}/\sqrt{N}) \sim t_{(N-1)}$ *se l'ipotesi nulla è vera.*
3. Scegliamo un livello di significatività $\alpha = 0{,}05$. Il valore critico $t_c = t_{(0{,}95;\,49)}$ è pari a 1,6766 per una distribuzione t con $N - 1 = 49$ gradi di libertà. Di conseguenza rifiuteremo l'ipotesi nulla a favore dell'alternativa se $t \geq 1,68$.
4. Usando il campione in questione otteniamo una stima di μ pari a $\overline{y} = 17{,}1582$ e una stima della varianza $\hat{\sigma}^2 = 3{,}2653$ e di conseguenza $\hat{\sigma} = 1{,}807$. Il valore della statistica test è dato da:
$$t = \frac{17{,}1582 - 16{,}5}{1{,}807/\sqrt{50}} = 2{,}5756$$
5. *Conclusione*: Dato che $t = 2{,}5756 > 1{,}68$, *rifiutiamo* l'ipotesi nulla. L'informazione campionaria in nostro possesso è *incompatibile* con l'ipotesi $\mu = 16.5$. Per un livello di significatività $\alpha = 0{,}05$ accettiamo l'ipotesi alternativa che la larghezza media del bacino nella popolazione sia maggiore di 16,5 pollici.

C.6.6. Esempio di test a due code usando i dati sulla larghezza del bacino

Consideriamo ora il test dell'ipotesi nulla che la larghezza media del bacino nella popolazione degli adulti statunitensi sia di 17 pollici rispetto all'alternativa che sia *diversa* da questo valore. Il test è composto dai cinque passaggi seguenti.

1. L'ipotesi nulla è $H_0 : \mu = 17$. L'ipotesi alternativa è $H_1 : \mu \neq 17$.
2. La statistica test è $t = (\overline{Y} - 17)/(\hat{\sigma}/\sqrt{N}) \sim t_{(N-1)}$ *se l'ipotesi nulla è vera.*
3. Scegliamo un livello di significatività $\alpha = 0{,}05$. In un test a due code viene assegnata una probabilità $\alpha/2 = 0{,}025$ a ciascuna delle code della distribuzione. Il valore critico è il percentile di livello 97,5% della distribuzione t con $N - 1 = 49$ gradi di libertà: $t_c = t_{(0{,}975;\,49)} = 2{,}01$, che lascia nella coda superiore una probabilità pari al 2,5%. Di conseguenza rifiutiamo l'ipotesi nulla a favore dell'alternativa se $t \geq 2{,}01$ o se $t \leq -2{,}01$.

4. Usando il campione in questione otteniamo una stima di μ pari a $\bar{y} = 17{,}1582$ e una stima della varianza $\hat{\sigma}^2 = 3{,}2653$ e di conseguenza $\hat{\sigma} = 1{,}807$. Il valore della statistica test è dato da $t = (17{,}1582 - 17)/(1{,}807/\sqrt{50}) = 0{,}6191$.
5. *Conclusione*: Dato che $-2{,}91 < t = 0{,}6191 < 2{,}91$, *non rifiutiamo* l'ipotesi nulla. L'informazione campionaria in nostro possesso è *compatibile* con l'ipotesi che la larghezza media del bacino nella popolazione sia $\mu = 17$.

NOTA IMPORTANTE: L'interpretazione del risultato di un test statistico richiede sempre una certa dose di cautela. Una delle regole fondamentali della verifica d'ipotesi è che un valore campionario della statistica test nella regione di non rifiuto non rende vera l'ipotesi nulla! Per capire perché, consideriamo un'altra ipotesi nulla $H_0 : \mu = c^*$, dove c^* è "vicino" a c. Se non possiamo rifiutare l'ipotesi $H_0 : \mu = c$, probabilmente non riusciremo a rifiutare neanche $H_0 : \mu = c^*$. Nell'esempio precedente, per $\alpha = 0{,}05$ non possiamo rifiutare l'ipotesi che μ sia 17, 16,8, 17,2 o 17.3. In effetti, in qualsiasi problema esistono molte ipotesi che non potremmo rifiutare, ma non per questo qualcuna di esse è necessariamente vera. Le affermazioni più deboli "non rifiutiamo l'ipotesi nulla" o "non possiamo rifiutare l'ipotesi nulla" hanno il vantaggio di non trasmettere un messaggio fuorviante.

C.6.7. Il p-value

Nell'illustrare il risultato di un test statistico è diventata pratica comune riportare il ***p*-value** del test. Con il p-value del test, indicato con p, possiamo determinare l'esito della verifica d'ipotesi confrontando p con il livello di significatività prescelto, α, *senza* dover cercare o calcolare i valori critici della statistica test. La regola è la seguente:

REGOLA DEL *p*-VALUE: Rifiutiamo l'ipotesi nulla se il p-value è inferiore o uguale al livello di significatività α. In altre parole, se $p \leq \alpha$, rifiutiamo H_0; se $p > \alpha$, H_0 non è rifiutata.

Se come livello di significatività avete scelto $\alpha = 0{,}01$, 0,05, 0,10 o qualsiasi altro valore, potete confrontarlo con il p-value del test e rifiutare o non rifiutare H_0 senza dover confrontare la statistica test con il valore critico t_c.

Il modo di calcolare il p-value dipende dall'ipotesi alternativa. Se t è il valore campionario (non il valore critico t_c) della statistica t con $N - 1$ gradi di libertà:

- se $H_1 : \mu > c$, $p =$ probabilità a destra di t;
- se $H_1 : \mu < c$, $p =$ probabilità a sinistra di t;
- se $H_1 : \mu \neq c$, $p =$ *somma* delle probabilità a destra di $|t|$ *e* a sinistra di $-|t|$.

La direzione dell'ipotesi alternativa indica la coda o le code della distribuzione nelle quali deve essere calcolato il p-value.

Nel paragrafo C.6.5 abbiamo usato i dati sulla larghezza del bacino per verificare $H_0 : \mu = 16{,}5$ rispetto all'alternativa $H_1 : \mu > 16{,}5$. Il valore campionario della statistica test era $t = 2{,}5756$. In questo caso, dato che l'alternativa è del tipo "maggiore di" ($>$), il p-value del test è la probabilità che una variabile casuale t

con $N - 1 = 49$ gradi di libertà sia maggiore di 2,5756. Questa probabilità non può essere trovata nelle consuete tavole dei valori critici della distribuzione t, ma può essere calcolata facilmente usando il computer. I software econometrici e statistici, ma anche i fogli elettronici come Excel, offrono alcuni semplici comandi che consentono di valutare la *funzione di ripartizione* (*fdr*) (si veda il Piccolo manuale di probabilità, paragrafo P.2) per una varietà di distribuzioni di probabilità. Se $F_X(x)$ è la *fdr* di una variabile casuale X, allora $P(X \leq c) = F_X(c)$, qualunque sia c. Data la funzione di ripartizione della distribuzione t, possiamo calcolare il p-value desiderato con la:

$$p = P[t_{(49)} \geq 2{,}576] = 1 - P[t_{(49)} \leq 2{,}576] = 0{,}0065$$

Dato il p-value possiamo immediatamente concludere che per $\alpha = 0{,}01$ o $0{,}05$ l'ipotesi nulla è rifiutata a favore dell'alternativa, ma che per $\alpha = 0{,}001$ l'ipotesi nulla non verrebbe rifiutata.

La figura C.8 illustra la logica alla base della regola del p-value. Se la probabilità a destra di $t = 2{,}5756$ è pari a 0,0065, il valore critico t_c che lascia nella coda destra una probabilità di 0,01 $[t_{(0{,}99;\,49)}]$ o di 0,05 $[t_{(0{,}95;\,49)}]$ deve trovarsi alla sinistra di 2,5756. In questo caso, in cui il p-value è minore o uguale di α, deve necessariamente essere vero che $t \geq t_c$ e per entrambi i livelli di significatività H_0 dovrebbe essere rifiutata. D'altro canto, il valore critico associato a $\alpha = 0{,}001$ deve cadere alla destra di 2,5756, il che significa che per questo livello di significatività l'ipotesi nulla non verrebbe rifiutata.

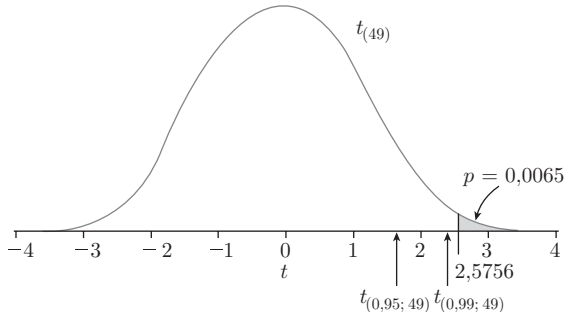

Figura C.8
p-value per un test a coda destra.

Per un test a due code la regione di rifiuto si trova in entrambe le code della distribuzione t e il p-value deve analogamente essere calcolato nelle due code. Per i dati sulla larghezza del bacino abbiamo considerato l'ipotesi nulla $H_0 : \mu = 17$ rispetto all'alternativa $H_1 : \mu \neq 17$, ottenendo un valore della statistica test $t = 0{,}6191$. Il p-value è dato da:

$$p = P[t_{(49)} \geq 0{,}6191] + P[t_{(49)} \leq -0{,}6191] = 2 \times 0{,}2694 = 0{,}5387$$

Dato che il p-value è 0,5387, maggiore di $\alpha = 0{,}05$, non possiamo rifiutare l'ipotesi nulla $H_0 : \mu = 17$ per $\alpha = 0{,}05$ e neppure per qualsiasi altro livello di significatività utilizzato di solito. Il p-value del test a due code è illustrato nella figura C.9.

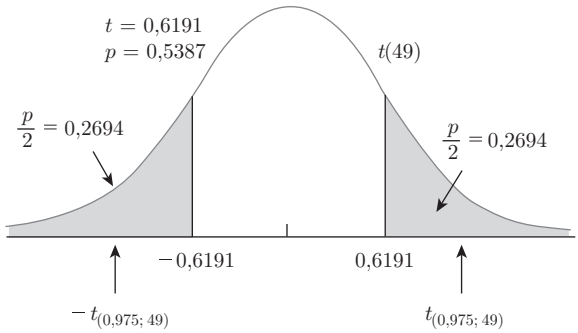

Figura C.9
p-value per un test a due code.

C.6.8. Formulazione di ipotesi nulla e ipotesi alternativa: un commento

Un test statistico non può dimostrare la verità dell'ipotesi nulla. Quando non possiamo rifiutarla, tutto ciò che la verifica d'ipotesi consente di dedurre è che l'informazione nel campione è *compatibile* con l'ipotesi nulla. D'altro canto, un test può spingerci a *rifiutare* l'ipotesi nulla anche se la probabilità α che questa decisione sia errata perché H_0 è in realtà vera è fissata a un livello decisamente basso. Di conseguenza, il rifiuto di un'ipotesi nulla rappresenta una conclusione più forte del non riuscire a rifiutarla.

L'ipotesi nulla viene di solito formulata in modo che se la nostra teoria è corretta saremo in grado di rifiutarla. L'ingegnere aeronautico, per esempio, ha elaborato il proprio progetto assumendo (ipotesi nulla) che la larghezza media del bacino nella popolazione sia di 16,5 pollici. Un'osservazione superficiale tuttavia suggerisce che nel corso del tempo gli individui adulti stanno in media progressivamente ingrassando. Se questo è vero, anche i sedili dovrebbero essere più larghi. Questa costosa modifica dovrebbe essere implementata solo in presenza di evidenza statistica che confermi l'aumento della larghezza media del bacino nella popolazione. Usando una verifica d'ipotesi ci piacerebbe scoprire se esiste davvero evidenza empirica contraria alla nostra "teoria" attuale, oppure se i dati sono compatibili con essa. Dato questo obiettivo, formuliamo l'ipotesi nulla che la media nella popolazione sia di 16,5 pollici, $H_0 : \mu = 16{,}5$, rispetto all'alternativa che sia maggiore, $H_1 : \mu > 16{,}5$. In questo caso se rifiutiamo l'ipotesi nulla abbiamo dimostrato che si è verificato un aumento "statisticamente significativo" della larghezza media del bacino.

Questa ipotesi nulla potrebbe essere considerata troppo limitata, dato che un'altra possibilità è che il valore corrente della larghezza media del bacino nella popolazione sia inferiore a 16,5 pollici. Il test dell'ipotesi nulla $H_0 : \mu \leq 16{,}5$ rispetto all'alternativa $H_1 : \mu > 16{,}5$ coincide con quello usato per verificare $H_0 : \mu = 16{,}5$ rispetto all'ipotesi alternativa $H_1 : \mu > 16{,}5$. La statistica test e la regione di rifiuto sono assolutamente le stesse. Per un test a una coda possiamo formulare l'ipotesi nulla equivalentemente in un modo o nell'altro.

Infine, è importante specificare l'ipotesi nulla e l'ipotesi alternativa prima di analizzare o persino di raccogliere il campione di osservazioni. Se non lo facciamo è possibile incorrere in errori nella formulazione dell'ipotesi alternativa. Supponiamo di voler verificare se $\mu > 16{,}5$ e di osservare una media campionaria $\bar{y}$ pari a 15,5. Questo risultato significa forse che l'alternativa dovrebbe essere specificata come $\mu < 16{,}5$, in modo da essere coerente con la stima? La risposta è negativa: l'ipotesi alternativa deve corrispondere alla congettura che vogliamo verificare: $\mu > 16{,}5$.

C.6.9. Errori di prima e di seconda specie

Sia che rifiutiamo sia che non rifiutiamo un'ipotesi nulla, esiste sempre la possibilità di commettere un errore. Questa eventualità è inevitabile. In qualunque verifica d'ipotesi esistono due circostanze in cui viene presa la decisione corretta e altre due in cui ne prendiamo una sbagliata.

DECISIONI CORRETTE

L'ipotesi nulla è *falsa* e noi decidiamo di *rifiutarla*.
L'ipotesi nulla è *vera* e noi decidiamo di *non rifiutarla*.

DECISIONI SBAGLIATE

L'ipotesi nulla è *vera* e noi decidiamo di *rifiutarla* (errore di prima specie).
L'ipotesi nulla è *falsa* e noi decidiamo di *non rifiutarla* (errore di seconda specie).

Quando rifiutiamo un'ipotesi nulla, corriamo un rischio detto "di prima specie". La probabilità di errore di prima specie è α, il livello di significatività del test. Quando l'ipotesi nulla è vera, la statistica t cade nella regione di rifiuto con probabilità α. Di conseguenza, in un test un'ipotesi nulla vera viene rifiutata il $100\alpha\%$ delle volte. Il fatto che siamo in grado di controllare la probabilità di errore di prima specie scegliendo il livello di significatività del test è indubbiamente una buona notizia.

Quando non rifiutiamo un'ipotesi nulla corriamo un rischio detto "di seconda specie". La probabilità che un test statistico non rifiuti un'ipotesi nulla falsa è diversa da zero e il suo valore rappresenta la probabilità di errore di seconda specie che non è sotto il nostro controllo e non può neppure essere calcolata, dato che dipende dal vero valore di μ che è ignoto. Sappiamo tuttavia che:

- La probabilità di errore di seconda specie varia inversamente con il livello di significatività del test, α, che rappresenta la probabilità d'errore di prima specie. Se scegliete di ridurre α, sappiate che la probabilità d'errore di seconda specie aumenterà.
- Se l'ipotesi nulla è $\mu = c$, e se il vero (e ignoto) valore di μ è *vicino* a c, la probabilità d'errore di seconda specie è elevata.
- Quanto maggiore è la numerosità campionaria, tanto minore sarà la probabilità d'errore di seconda specie per un dato livello di errore di prima specie α.

Un esempio facile da memorizzare della differenza fra errori di prima e di seconda specie proviene dai sistemi legali in vigore in molti paesi, fra i quali USA e Italia. In qualsiasi processo l'imputato è considerato innocente. Ciò costituisce l'ipotesi "nulla", mentre l'ipotesi alternativa è che sia colpevole. Se condanniamo un innocente, significa che abbiamo rifiutato un'ipotesi nulla vera, commettendo un errore di prima specie. Se non condanniamo un colpevole perché non possiamo rifiutare un'ipotesi nulla falsa, stiamo commettendo un errore di seconda specie. Qual è l'errore più costoso in questo contesto? È meglio mandare in prigione un innocente o lasciare libero un colpevole? In questo caso, è meglio ridurre a un valore molto piccolo la probabilità d'errore di prima specie.

C.6.10. Relazione fra verifica d'ipotesi e intervalli di confidenza

Fra i test d'ipotesi a due code e le stime intervallari esiste una relazione algebrica che può essere utile conoscere. Supponiamo che stiate verificando l'ipotesi nulla $H_0 : \mu = c$ rispetto all'alternativa $H_1 : \mu \neq c$. Se non rifiutiamo l'ipotesi nulla per un livello di significatività pari a α, il valore c cadrà all'interno della stima intervallare di μ associata al livello di confidenza $100(1-\alpha)\%$. Viceversa, se rifiutiamo l'ipotesi nulla, c cadrà al di fuori della stima intervallare al $100(1-\alpha)\%$ di μ. Questa relazione algebrica è vera perché l'ipotesi nulla non viene rifiutata quando $-t_c \leq t \leq t_c$ e cioè quando:

$$-t_c \leq \frac{\overline{Y} - c}{\hat{\sigma}/\sqrt{N}} \leq t_c$$

che con qualche passaggio algebrico diventa:

$$\overline{Y} - t_c \frac{\hat{\sigma}}{\sqrt{N}} \leq c \leq \overline{Y} + t_c \frac{\hat{\sigma}}{\sqrt{N}}$$

Gli estremi di questo intervallo sono gli stessi che definiscono la stima intervallare di μ per un livello di confidenza $100(1-\alpha)\%$. Di conseguenza, per ogni valore di c all'interno dell'intervallo di confidenza non potremo rifiutare $H_0 : \mu = c$ rispetto all'alternativa $H_1 : \mu \neq c$. Per qualunque valore di c al di fuori dell'intervallo rifiuteremo $H_0 : \mu = c$ e accetteremo l'alternativa $H_1 : \mu \neq c$.

Questa relazione può essere utile se i risultati che vi sono stati consegnati contengono solo un intervallo di confidenza mentre a voi interessa determinare quale sarebbe l'esito di un test a due code.

C.7. Altri utili test

In questo paragrafo riassumeremo in maniera molto sintetica qualche altro test. Oltre a rappresentare strumenti utili per verificare le rispettive ipotesi nulle, questi test servono anche a illustrare l'uso di statistiche con distribuzione chi quadro o F. Queste distribuzioni sono state introdotte nell'appendice B.3.

C.7.1. Test sulla varianza della popolazione

Sia Y una variabile casuale di distribuzione normale, $Y \sim \mathcal{N}(\mu, \sigma^2)$. Assumiamo di disporre di un campione casuale di numerosità N, $Y_1, Y_2, \ldots, Y_N$, tratto da questa popolazione. Lo stimatore della media della popolazione è $\overline{Y} = \sum Y_i/N$ e lo stimatore corretto della varianza della popolazione è $\hat{\sigma}^2 = \sum(Y_i - \overline{Y})^2/(N-1)$. Per verificare l'ipotesi nulla $H_0 : \sigma^2 = \sigma_0^2$ usiamo la statistica test:

$$V = \frac{(N-1)\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{(N-1)}$$

Se l'ipotesi nulla è vera la statistica test ha distribuzione chi quadro con $N-1$ gradi di libertà. Se l'ipotesi alternativa è $H_1 : \sigma^2 > \sigma_0^2$, il test che dobbiamo svolgere è a una coda. Se come livello di significatività scegliamo $\alpha = 0{,}05$, l'ipotesi nulla è rifiutata se $V \geq \chi^2_{(0{,}95;\, N-1)}$, dove $\chi^2_{(0{,}95;\, N-1)}$ è il 95-esimo percentile della distribuzione chi quadro con $N-1$ gradi di libertà. Questi valori possono essere individuati nella tabella 3 dell'appendice D o calcolati usando un software statistico. Se l'ipotesi alternativa è $H_1 : \sigma^2 \neq \sigma_0^2$, dobbiamo svolgere un test a

due code e l'ipotesi nulla è rifiutata se $V \geq \chi^2_{(0{,}975;\, N-1)}$ o se $V \leq \chi^2_{(0{,}025;\, N-1)}$. La distribuzione chi quadro è asimmetrica con una lunga coda a destra; di conseguenza non possiamo sfruttare una proprietà di simmetria per determinare il valore critico a sinistra in funzione di quello a destra.

C.7.2. Test di uguaglianza delle medie di due popolazioni

Consideriamo due popolazioni normali $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Per stimare i parametri ignoti e verificare un'ipotesi sulla differenza fra le medie $\mu_1 - \mu_2$ dobbiamo estrarre campioni casuali di osservazioni da entrambe le popolazioni. Estraiamo un campione di numerosità N_1 dalla prima e uno di numerosità N_2 dalla seconda. Usando il primo campione, otteniamo una media campionaria $\overline{Y}_1$ e una varianza campionaria $\hat{\sigma}_1^2$; usando il secondo, una media $\overline{Y}_2$ e una varianza $\hat{\sigma}_2^2$. La procedura da seguire per verificare l'ipotesi nulla $H_0 : \mu_1 - \mu_2 = c$ dipende dal fatto che le varianze siano o meno uguali fra loro.

Caso 1: Varianze della popolazione uguali Se le varianze della popolazione sono uguali, e dunque $\sigma_1^2 = \sigma_2^2 = \sigma_p^2$, usiamo l'informazione in entrambi i campioni per stimare la varianza comune σ_p^2. Lo "stimatore congiunto della varianza" è definito da:

$$\hat{\sigma}_p^2 = \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}$$

Se l'ipotesi nulla $H_0 : \mu_1 - \mu_2 = c$ è vera:

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2) - c}{\sqrt{\hat{\sigma}_p^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}} \sim t_{(N_1 + N_2 - 2)}$$

Come di consueto possiamo considerare un'alternativa a una coda come $H_1 : \mu_1 - \mu_2 > c$ oppure a due code, $H_1 : \mu_1 - \mu_2 \neq c$.

Caso 2: Varianze della popolazione diverse Se le varianze nella popolazione non sono uguali non possiamo usare lo stimatore congiunto della varianza. Come statistica test useremo allora:

$$t^* = \frac{(\overline{Y}_1 - \overline{Y}_2) - c}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

La distribuzione esatta di questa statistica test non è né normale né la solita distribuzione t. La distribuzione di t^* può essere approssimata da una distribuzione t con gradi di libertà dati da:

$$\text{gdl} = \frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{\frac{(\hat{\sigma}_1^2/N_1)^2}{N_1 - 1} + \frac{(\hat{\sigma}_2^2/N_2)^2}{N_2 - 1}}$$

Questa è solo una delle tante approssimazioni che compaiono nei testi di statistica; il vostro software potrebbe benissimo usarne una diversa.

C.7.3. Test del rapporto delle varianze di due popolazioni

Date due popolazioni normali indicate con $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$, consideriamo il test dell'ipotesi nulla $H_0 : \sigma_1^2/\sigma_2^2 = 1$. Se l'ipotesi nulla è vera, le varianze delle popolazioni sono uguali. La statistica test viene derivata dalle due proprietà $(N_1-1)\hat{\sigma}_1^2/\sigma_1^2 \sim \chi^2_{(N_1-1)}$ e $(N_2-1)\hat{\sigma}_2^2/\sigma_2^2 \sim \chi^2_{(N_2-1)}$. Nell'appendice B.3.8 abbiamo definito la variabile casuale F come il rapporto di due variabili casuali chi quadro indipendenti divise per i rispettivi gradi di libertà. In questo caso il rapporto da considerare è dato da:

$$F = \frac{\dfrac{(N_1-1)\hat{\sigma}_1^2/\sigma_1^2}{N_1-1}}{\dfrac{(N_2-1)\hat{\sigma}_2^2/\sigma_2^2}{N_2-1}} = \frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2} \sim F_{(N_1-1;\, N_2-1)}$$

Se l'ipotesi nulla è $H_0 : \sigma_1^2/\sigma_2^2 = 1$ è vera, la statistica test $F = \hat{\sigma}_1^2/\hat{\sigma}_2^2$ ha distribuzione F con $N_1 - 1$ gradi di libertà al numeratore e $N_2 - 1$ al denominatore. Se l'ipotesi alternativa è $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$, dobbiamo svolgere un test a due code. Con un livello di significatività $\alpha = 0{,}05$, rifiutiamo l'ipotesi nulla se $F \geq F_{(0{,}975;\, N_1-1;\, N_2-1)}$ oppure se $F \leq F_{(0{,}025;\, N_1-1;\, N_2-1)}$, dove $F_{(\alpha;\, N_1-1;\, N_2-1)}$ è il percentile di livello $100\alpha\%$ della distribuzione F con i gradi di libertà indicati. Se l'alternativa è a una coda, $H_1 : \sigma_1^2/\sigma_2^2 > 1$, rifiuteremo l'ipotesi nulla se $F \geq F_{(0{,}95;\, N_1-1;\, N_2-1)}$.

C.7.4. Test di normalità di una popolazione

I test sulle medie e le varianze che abbiamo illustrato partono dall'ipotesi che le popolazioni abbiano distribuzione normale. Questa osservazione suggerisce immediatamente due quesiti. Quali sono le proprietà di questi test se la popolazione non è normale? Possiamo verificare la normalità di una popolazione? La risposta alla prima domanda è che i test funzionano piuttosto bene anche se la popolazione non è normale, a condizione che i campioni siano sufficientemente numerosi. Che cosa si intende esattamente per "sufficientemente numerosi"? Purtroppo a questa domanda è impossibile rispondere in maniera semplice, perché la risposta precisa dipende dal grado di "non normalità" che caratterizza la vera distribuzione della popolazione. La risposta alla seconda domanda precedente è invece sì, possiamo verificare la normalità di una popolazione. Per molto tempo gli statistici hanno dedicato grande attenzione a questo problema e hanno sviluppato una varietà di test; sfortunatamente la teoria sottostante e gli stessi test sono molto complessi e decisamente al di là degli obiettivi di questo volume.

Possiamo tuttavia presentare un test leggermente meno ambizioso. La distribuzione normale è simmetrica e ha una distribuzione caratterizzata da un grado di concentrazione al centro e di spessore delle code che fa sì che il suo indice di curtosi sia pari a 3. Possiamo dunque verificare la presenza di scostamenti dalla normalità esaminando l'asimmetria e la curtosi nel campione. Se l'asimmetria non è prossima a 0 e la curtosi non è vicina a 3, la normalità della popolazione può essere rifiutata. Nel paragrafo C.4.2 abbiamo sviluppato gli indici campionari di asimmetria e curtosi:

$$\widehat{\text{asimmetria}} = A = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3}$$

$$\widehat{\text{curtosi}} = C = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$$

La statistica test di **Jarque-Bera** consente di verificare congiuntamente la presenza di queste due caratteristiche:

$$JB = \frac{N}{6}\left[A^2 + \frac{(C-3)^2}{4}\right]$$

Se la vera distribuzione è simmetrica e ha curtosi pari a 3, e potrebbe dunque essere normale, la statistica test JB ha distribuzione chi quadro con due gradi di libertà se la numerosità campionaria è sufficientemente elevata. Se $\alpha = 0{,}05$, il valore critico della distribuzione $\chi^2_{(2)}$ è 5,99. Se $JB \geq 5{,}99$, rifiutiamo l'ipotesi nulla e concludiamo che i dati sono non normali. In questo caso sappiamo che la popolazione presenta caratteristiche incompatibili con la normalità, ma non sappiamo quale sia la sua vera distribuzione.

Nel caso dei dati sulla larghezza del bacino gli indici campionari di asimmetria e di curtosi sono già stati calcolati nel paragrafo C.4.3. Sostituendo questi valori nella formula della statistica test JB, otteniamo:

$$JB = \frac{N}{6}\left[A^2 + \frac{(C-3)^2}{4}\right] = \frac{50}{6}\left[(-0{,}0138)^2 + \frac{(2{,}3315-3)^2}{4}\right] = 0{,}9325$$

Dato che $JB = 0{,}9325$ è minore del valore critico 5,99, concludiamo che non è possibile rifiutare la normalità dei dati sulla larghezza del bacino. Il p-value di questo test è l'area nella coda della distribuzione $\chi^2_{(2)}$ a destra di 0,9325 ed è dato da:

$$p = P\left[\chi^2_{(2)} \geq 0{,}9325\right] = 0{,}6273$$

C.8. Introduzione alla stima di massima verosimiglianza[1]

La tecnica di stima di massima verosimiglianza è una procedura di inferenza molto importante che può essere usata quando la distribuzione della popolazione è nota. In questo paragrafo ne introdurremo l'idea usando un'illustrazione molto semplice ma ciò nonostante illuminante. Considerate l'esempio seguente, ispirato al gioco chiamato "La ruota della fortuna". Immaginate di essere uno dei partecipanti e di avere di fronte a voi due ruote, ciascuna delle quali è in parte annerita e in parte non annerita (si veda la figura C.10). Nel gioco viene fatta girare una delle ruote; si vince se un indicatore fisso punta verso un'area annerita, in caso contrario si perde. Sulla ruota A l'area annerita è il 25% dell'area complessiva; la probabilità di vincere è dunque pari a 1/4. Sulla ruota B l'area annerita è il 75% del totale e la probabilità di vincere è data da 3/4. Nel gioco cui partecipate viene scelta e fatta girare tre volte una delle ruote, ottenendo come risultati VINCI, VINCI e PERDI, *senza mostrare* quale ruota è stata utilizzata; l'obiettivo finale è indovinare quale ruota ha generato i risultati. Voi quale scegliereste?

Intuitivamente potremmo precedere nel modo seguente: indichiamo con p la probabilità di vittoria quando la ruota viene fatta girare. Scegliere fra le ruote A e B equivale a scegliere fra $p = 1/4$ e $p = 3/4$. Stiamo stimando p ed esistono solo due possibili stime; la nostra scelta deve essere basata sulle osservazioni disponibili. Calcoliamo la probabilità della sequenza osservata per ciascuna delle ruote.

[1] Questo paragrafo contiene materiale di livello avanzato.

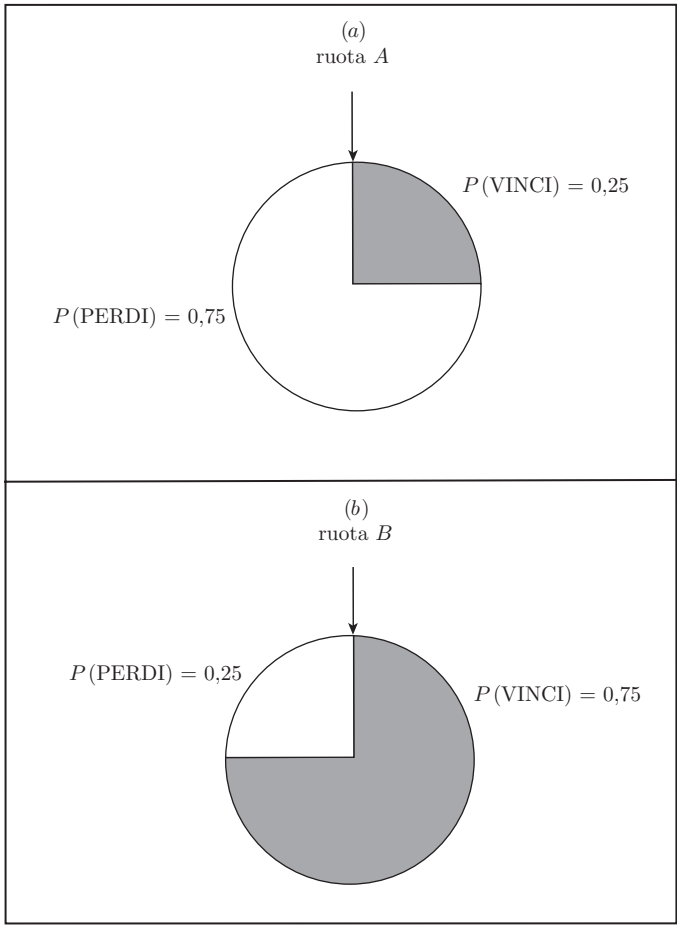

Figura C.10
Il gioco della ruota della fortuna.

Per la ruota A, con $p = 1/4$, la probabilità di osservare VINCI, VINCI e PERDI è data da:

$$\frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = 0{,}0469$$

In altre parole, la probabilità, o **verosimiglianza**, di osservare la sequenza VINCI, VINCI e PERDI se $p = 1/4$ è 0,0469.

Per la ruota B, con $p = 3/4$, la probabilità di osservare VINCI, VINCI e PERDI è data da:

$$\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = 0{,}1406$$

La probabilità, o verosimiglianza, di osservare la sequenza VINCI, VINCI e PERDI se $p = 3/4$ è 0,1406.

Se dovessimo scegliere fra la ruota A e la B sulla base delle osservazioni disponibili sceglieremmo la B perché è quella che ha probabilità più alta di aver generato i dati osservati. In altre parole, è più *verosimile* che sia stata girata la ruota B piuttosto che la ruota A e $\hat{p} = 3/4$ è la **stima di massima verosimiglianza** di p. Il **principio di massima verosimiglianza** suggerisce di cercare il valore dei

parametri che massimizzano la probabilità, o la verosimiglianza, di osservare i dati effettivamente disponibili.

Supponiamo ora che p possa essere una probabilità qualsiasi compresa fra 0 e 1 e non solo 1/4 e 3/4. Abbiamo di fronte una ruota di cui una parte, quella che rappresenta la probabilità di ottenere VINCI, è annerita ma non conosciamo a quanto ammonti la sua area in percentuale sulla superficie totale. In tre lanci osserviamo VINCI, VINCI e PERDI. Qual è il valore più verosimile di p? La probabilità di osservare VINCI, VINCI e PERDI è la verosimiglianza L ed è data da:

$$L(p) = p \times p \times (1-p) = p^2 - p^3 \tag{C.17}$$

La verosimiglianza L dipende dalla probabilità ignota p di ottenere VINCI; questo spiega perché abbiamo usato la notazione $L(p)$ che suggerisce l'esistenza di una relazione funzionale. Vorremmo trovare il valore di p che massimizza la verosimiglianza di ottenere i tre esiti effettivamente osservati. La figura C.11 illustra la funzione di verosimiglianza (C.17) e il valore di p che massimizza questa funzione. Questo valore è indicato con $\hat{p}$ ed è chiamato stima di massima verosimiglianza di p. Per calcolarlo possiamo usare qualche passaggio matematico. Derivando $L(p)$ rispetto a p, otteniamo:

$$\frac{dL(p)}{dp} = 2p - 3p^2$$

Uguagliamo questa derivata a zero:

$$2p - 3p^2 = 0 \quad \Rightarrow \quad p(2-3p) = 0$$

Questa equazione ha due soluzioni, $p = 0$ e $p = 2/3$. Il valore che massimizza $L(p)$ è $\hat{p} = 2/3$, la stima di massima verosimiglianza di p. In altre parole, fra tutti i possibili valori di p compresi fra 0 e 1, quello che massimizza la probabilità di osservare due vittorie e una sconfitta (l'ordine di queste osservazioni è irrilevante) è $\hat{p} = 2/3$.

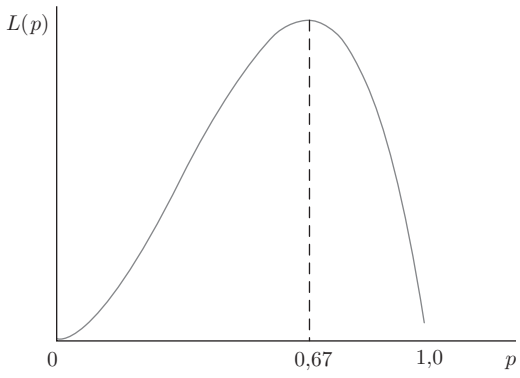

Figura C.11
Una funzione di verosimiglianza.

Come possiamo derivare una formula più generale che possa essere usata per qualsiasi insieme di osservazioni? Nell'appendice B.3.1 abbiamo introdotto la distribuzione di Bernoulli. Consideriamo la variabile casuale X che assume valore $x = 1$ (VINCI) e $x = 0$ (PERDI) con probabilità p e $1 - p$. Da un punto di vista matematico, la funzione di probabilità per questa variabile casuale può essere espressa come:

$$P(X = x) = f(x|p) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Se facciamo girare la "ruota" N volte, osserveremo gli N valori campionari $x_1, x_2, \ldots, x_N$. Assumendo che gli N tentativi siano indipendenti fra loro, possiamo costruire la funzione di probabilità congiunta:

$$\begin{aligned} f(x_1, \ldots, x_N|p) &= f(x_1|p) \times \ldots \times f(x_N|p) \\ &= p^{\sum x_i}(1-p)^{N-\sum x_i} \\ &= L(p|x_1, \ldots, x_N) \end{aligned} \tag{C.18}$$

La funzione di probabilità congiunta rappresenta la probabilità di osservare una particolare sequenza di risultati ed è la generalizzazione della (C.17). L'ultima riga indica che da un punto di vista algebrico la funzione di probabilità congiunta è equivalente alla **funzione di verosimiglianza** $L(p|x_1, \ldots, x_N)$. La notazione sottolinea che la funzione di verosimiglianza dipende dalla probabilità ignota p *date* le osservazioni campionarie. Per semplificare la notazione nel seguito indicheremo la funzione di verosimiglianza solo con $L(p)$.

Nel gioco della "ruota della fortuna" la stima di massima verosimiglianza è il valore di p che massimizza $L(p)$. Per calcolare matematicamente questa stima useremo una stratagemma che semplifica i calcoli. Il valore di p che massimizza $L(p) = p^2(1-p)$ è lo stesso che massimizza la **funzione di logverosimiglianza** $\log L(p) = 2\log(p) + \log(1-p)$, dove "log" indica il logaritmo naturale. Il grafico della funzione di logverosimiglianza è illustrato nella figura C.12. Confrontate questa figura con la C.11: il massimo della funzione di verosimiglianza è $L(\hat{p}) = 0{,}1481$ e quello della funzione di logverosimiglianza è $\log L(\hat{p}) = -1{,}9095$. Entrambi questi valori si verificano per $\hat{p} = 2/3 = 0{,}6667$.

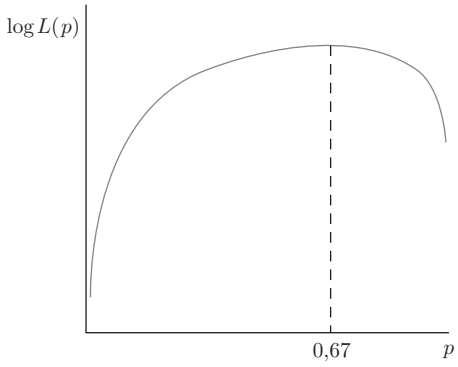

Figura C.12
Una funzione di logverosimiglianza.

Questa proprietà vale per tutte le funzioni di verosimiglianza e logverosimiglianza e per tutti i valori dei loro parametri; per questo motivo la stima di massima verosimiglianza sarà sempre formulata sotto forma di massimizzazione della funzione di logverosimiglianza. Nel problema generale che stiamo considerando, la funzione di logverosimiglianza è il logaritmo di (C.18):

$$\begin{aligned} \log L(p) &= \sum_{i=1}^{N} \log f(x_i|p) \\ &= \left(\sum_{i=1}^{N} x_i\right) \log(p) + \left(N - \sum_{i=1}^{N} x_i\right) \log(1-p) \end{aligned} \tag{C.19}$$

La derivata prima è data da:

$$\frac{d \log L(p)}{dp} = \frac{\sum x_i}{p} - \frac{N - \sum x_i}{1 - p}$$

Uguagliando la derivata a zero e sostituendo p con $\hat{p}$, il valore che massimizza $\log L(p)$, otteniamo:

$$\frac{\sum x_i}{\hat{p}} - \frac{N - \sum x_i}{1 - \hat{p}} = 0$$

Per risolvere questa equazione moltiplichiamo entrambi i termini per $\hat{p}(1 - \hat{p})$:

$$(1 - \hat{p}) \sum x_i - \hat{p} \left(N - \sum x_i \right) = 0$$

Infine, risolvendo rispetto a $\hat{p}$, otteniamo:

$$\hat{p} = \frac{\sum x_i}{N} = \bar{x} \tag{C.20}$$

Lo stimatore $\hat{p}$ è la **percentuale campionaria di vittorie**: $\sum x_i$ è il numero complessivo di osservazioni pari a 1 (VINCI) su N tentativi. Come potete osservare $\hat{p}$ è anche la media campionaria delle x_i. Questo risultato è del tutto generale: ogni volta che l'esito osservato può assumere solo due valori diversi con probabilità p e $1 - p$, la stima di massima verosimiglianza basata su un campione di N osservazioni è la percentuale campionaria (C.20). Questa strategia di stima può essere usata da un individuo incaricato di condurre un sondaggio per stimare la quota di popolazione che intende votare per il candidato A invece che per il candidato B, oppure da uno studioso di medicina che desidera stimare la quota di popolazione caratterizzata da una particolare mutazione genetica, o da un ricercatore di marketing che vuole scoprire se per i cereali della colazione la popolazione dei consumatori preferisce una scatola blu o una verde. In quest'ultimo caso, supponiamo di selezionare in maniera casuale 200 consumatori di cereali e di chiedere loro se preferiscono una scatola blu o una verde. Se 75 dichiarano di preferire la scatola blu, la stima della quota di individui nella popolazione che preferisce la scatola blu è pari a $\hat{p} = \sum x_i / N = 75/200 = 0{,}375$. Di conseguenza, la stima suggerisce che il 37,5% della popolazione di consumatori preferisce una scatola blu.

C.8.1. Inferenza con gli stimatori di massima verosimiglianza

Come possiamo effettuare verifiche d'ipotesi e costruire stime intervallari se usiamo la stima di massima verosimiglianza? La risposta a questa domanda sfrutta alcune importanti proprietà degli stimatori costruiti seguendo il principio di massima verosimiglianza. Consideriamo un problema del tutto generale: sia X una variabile casuale (discreta o continua) con funzione di densità o di probabilità $f(x|\theta)$, dove θ è un parametro ignoto. La funzione di logverosimiglianza, costruita sulla base di un campione casuale $x_1, \ldots, x_N$ di numerosità N, è data da:

$$\log L(\theta) = \sum_{i=1}^{N} \log f(x_i|\theta)$$

Se la funzione di densità o di probabilità della variabile casuale è sufficientemente regolare e se valgono alcune condizioni tecniche, in grandi campioni lo stimatore di massima verosimiglianza $\hat{\theta}$ del parametro θ ha distribuzione di probabilità

approssimativamente normale, di valore atteso θ e varianza $V = \text{Var}(\hat{\theta})$, la cui espressione sarà discussa fra breve. In altre parole, possiamo scrivere che:

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, V) \tag{C.21}$$

dove il simbolo $\overset{a}{\sim}$ sta per "asintoticamente distribuito come". Con il termine "asintotico" indichiamo una proprietà dello stimatore valida quando la numerosità campionaria N diventa elevata, oppure quando $N \to \infty$. Affermare che uno stimatore è asintoticamente normale implica che la sua distribuzione di probabilità, che potrebbe essere ignota per piccoli campioni, diventa approssimativamente normale in campioni molto numerosi. Questo risultato è analogo a quello stabilito dal teorema del limite centrale che abbiamo discusso nel paragrafo C.3.4.

Non sorprendentemente, usando la proprietà di normalità (C.21) possiamo costruire facilmente una statistica t e da essa ottenere immediatamente sia un intervallo di confidenza sia una statistica test. In particolare, se desideriamo verificare l'ipotesi nulla $H_0 : \theta = c$ rispetto a un'ipotesi alternativa a una o due code possiamo usare la statistica test:

$$t = \frac{\hat{\theta} - c}{\text{se}(\hat{\theta})} \overset{a}{\sim} t_{(N-1)} \tag{C.22}$$

Se l'ipotesi nulla è vera, questa statistica t ha una distribuzione che in grandi campioni può essere approssimata con una distribuzione t con $N - 1$ gradi di libertà. I passaggi che costituiscono il test sono esattamente gli stessi descritti nel paragrafo C.6.

Se t_c indica il percentile $t_{(1-\alpha/2;\, N-1)}$ di livello $100(1-\alpha/2)\%$ della distribuzione $t_{(N-1)}$, l'intervallo di confidenza per θ al $100(1-\alpha)\%$ è definito da:

$$\hat{\theta} \pm t_c \text{se}(\hat{\theta})$$

Questo intervallo di confidenza può essere interpretato esattamente come quelli nel paragrafo C.5.

NOTA: I risultati asintotici (C.21) e (C.22) valgono solo in grandi campioni, grazie al fatto che per numerosità campionarie elevate la distribuzione della statistica t può essere approssimata da una distribuzione t con $N - 1$ gradi di libertà. Se N è veramente elevato, la distribuzione $t_{(N-1)}$ converge alla distribuzione normale standardizzata $\mathcal{N}(0, 1)$ e il percentile $t_{(1-\alpha/2;\, N-1)}$ converge al corrispondente percentile della distribuzione normale standardizzata. A torto o a ragione, i risultati asintotici sono usati spesso anche quando la numerosità campionaria N non è elevata. In questo caso è preferibile usare i valori critici della distribuzione t, che attraverso la correzione prodotta dai gradi di libertà tengono conto della dimensione ridotta del campione nel costruire stime intervallari e nell'effettuare verifiche d'ipotesi.

C.8.2. Varianza dello stimatore di massima verosimiglianza

Sia nell'espressione della statistica test sia in quella dell'intervallo di confidenza uno degli ingredienti principali è lo standard error $\text{se}(\hat{\theta})$. Come possiamo calcolare

questa quantità? Uno standard error è la radice quadrata di una stima della varianza. Fino a ora abbiamo evitato di discutere come calcolare la varianza dello stimatore di massima verosimiglianza, $V = \text{Var}(\hat{\theta})$. La varianza V è data dall'inversa del valore atteso della derivata seconda della funzione di verosimiglianza, cambiato di segno:

$$V = \text{Var}(\hat{\theta}) = \left[-E\left(\frac{d^2 \log L(\theta)}{d\theta^2} \right) \right]^{-1} \tag{C.23}$$

Questa espressione ha l'aria piuttosto terrificante e questo è esattamente il motivo per il quale abbiamo aspettato tanto per tirarla in ballo. Qual è il suo significato? Prima di tutto, la derivata seconda misura la curvatura della funzione di logverosimiglianza. Una derivata seconda è, letteralmente, la derivata della derivata. La derivata prima misura la pendenza o il tasso di variazione di una funzione. La derivata seconda (la derivata della derivata prima) misura il tasso di variazione della pendenza. Per avere un massimo la funzione di logverosimiglianza deve avere una forma a "scodella rovesciata" come quelle illustrate nella figura C.13.

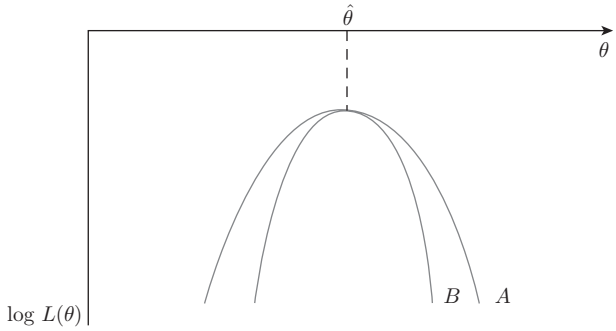

Figura C.13
Due funzioni di logverosimiglianza.

In un punto qualsiasi alla sinistra del massimo la pendenza della funzione di logverosimiglianza è positiva; in qualunque punto alla sua destra è negativa. A mano a mano che ci spostiamo da sinistra verso destra la pendenza *diminuisce* (diventa meno positiva o più negativa) e di conseguenza la derivata seconda deve essere negativa. Un valore assoluto più elevato della derivata seconda implica che la pendenza cambia più rapidamente, il che indica una funzione di logverosimiglianza con una maggiore curvatura. Questa osservazione è importante. Nella figura C.13 le due funzioni di logverosimiglianza A e B sono massime in corrispondenza dello stesso valore $\hat{\theta}$. Immaginate di essere uno scalatore che sta salendo lungo il crinale di una di queste due montagne. Per quale delle due montagne sarà più facile capire di essere arrivati alla vetta? Nel caso della logverosimiglianza B la posizione del punto di massimo è chiara e molto più facile da individuare di quella della logverosimiglianza A. Una funzione con una curvatura maggiore è caratterizzata da una "zona ambigua" meno estesa attorno al vertice. La minore estensione della regione ambigua implica che esiste minore incertezza per quanto riguarda la posizione del valore di θ che massimizza la funzione, $\hat{\theta}$; in termini di stima, una minore incertezza significa una maggiore precisione e una varianza più piccola. La funzione di logverosimiglianza con curvatura maggiore, con derivata seconda più elevata in valore assoluto, produce una stima di massima verosimiglianza più accurata e uno stimatore di massima verosimiglianza con una varianza minore. La varianza

V dello stimatore di massima verosimiglianza, dunque, varia inversamente con la derivata seconda (cambiata di segno). Il valore atteso E è necessario perché questa quantità dipende dai dati ed è dunque una variabile casuale; per questo motivo dobbiamo calcolarne la media su tutti i possibili campioni.

C.8.3. Distribuzione della quota campionaria

È il momento di un esempio. All'inizio del paragrafo C.8 abbiamo introdotto una variabile casuale X che assume i valori $x = 1$ e $x = 0$ con probabilità p e $1 - p$. La funzione di logverosimiglianza per un campione di osservazioni indipendenti di X è descritta dalla (C.19). In questo problema il parametro che stiamo stimando è la quota p di valori $x = 1$ nella popolazione. Sappiamo già che lo stimatore di massima verosimiglianza di p è la quota campionaria $\hat{p} = \sum x_i/N$. La derivata seconda della funzione di logverosimiglianza (C.19) è data da:

$$\frac{d^2 \log L(p)}{dp^2} = -\frac{\sum x_i}{p^2} - \frac{N - \sum x_i}{(1-p)^2} \tag{C.24}$$

Per calcolare la varianza dello stimatore di massima verosimiglianza ci serve il "valore atteso" dell'espressione (C.24). Nel valore atteso consideriamo i valori x_i come variabili casuali, dato che variano da un campione all'altro. Possiamo ottenere il valore atteso di questa variabile casuale discreta usando la (P.9) nel Piccolo manuale di probabilità:

$$E(x_i) = 1 \times P(x_i = 1) + 0 \times P(x_i = 0) = 1 \times p + 0 \times (1 - p) = p$$

Usando una generalizzazione di (P.16) (il valore atteso di una somma è la somma dei valori attesi e le costanti moltiplicative possono essere raccolte fuori dal valore atteso) possiamo ora calcolare il valore atteso della derivata seconda come:

$$\begin{aligned} E\left[\frac{d^2 \log L(p)}{dp^2}\right] &= -\frac{\sum E(x_i)}{p^2} - \frac{N - \sum E(x_i)}{(1-p)^2} \\ &= -\frac{Np}{p^2} - \frac{N - Np}{(1-p)^2} \\ &= -\frac{N}{p(1-p)} \end{aligned}$$

La varianza dello stimatore di massima verosimiglianza di p, la quota campionaria $\hat{p}$ di $x = 1$, è dunque data da:

$$V = \mathrm{Var}(\hat{p}) = \left[-E\left(\frac{d^2 \log L(p)}{dp^2}\right)\right]^{-1} = \frac{p(1-p)}{N}$$

La distribuzione asintotica della quota campionaria, valida in grandi campioni, è:

$$\hat{p} \overset{a}{\sim} \mathcal{N}\left[p, \frac{p(1-p)}{N}\right]$$

Per stimare la varianza V dobbiamo sostituire la vera quota di $x = 1$ nella popolazione con la sua stima:

$$\hat{V} = \frac{\hat{p}(1-\hat{p})}{N}$$

Lo standard error che ci serve per effettuare verifiche d'ipotesi e costruire stime intervallari è la radice quadrata di questa stima della varianza:

$$\text{se}(\hat{p}) = \sqrt{\hat{V}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Come esempio numerico supponiamo che un dirigente di un'impresa produttrice di cereali per la prima colazione ritenga che il 40% della popolazione preferisca una scatola blu. Per verificare questa congettura costruiamo l'ipotesi nulla $H_0 : p = 0{,}4$ e usiamo l'alternativa a due code $H_1 : p \neq 0{,}4$. Se l'ipotesi nulla è vera, la statistica test $t = (\hat{p} - 0{,}4)/\text{se}(\hat{p})$ ha approssimativamente distribuzione t: $t \overset{a}{\sim} t_{(N-1)}$. Per un campione di numerosità $N = 200$, il valore critico della distribuzione $t_{(199)}$ è $t_c = t_{(0{,}975;\,199)} = 1{,}96$. Di conseguenza, rifiutiamo l'ipotesi nulla se nel campione $t \geq 1{,}96$ o $t \leq -1{,}96$. Se 75 degli intervistati preferiscono una scatola blu, la quota campionaria è data da $\hat{p} = 75/200 = 0{,}375$. Lo standard error di questa stima è:

$$\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = \sqrt{\frac{0{,}375 \times 0{,}625}{200}} = 0{,}0342$$

Il valore della statistica test è:

$$t = \frac{\hat{p} - 0{,}4}{\text{se}(\hat{p})} = \frac{0{,}375 - 0{,}4}{0{,}0342} = -0{,}7303$$

Questo valore appartiene alla regione di non rifiuto, $-1{,}96 < t = -0{,}7303 < 1{,}96$, e di conseguenza non rifiutiamo l'ipotesi nulla $p = 0{,}4$. Le osservazioni campionarie sono compatibili con la congettura che il 40% della popolazione dei consumatori preferisca una scatola blu.

La stima intervallare al 95% della quota p di consumatori nella popolazione che preferiscono una scatola blu è data da:

$$\hat{p} \pm 1{,}96\,\text{se}(\hat{p}) = 0{,}375 \pm 1{,}96(0{,}0342) = [0{,}3075;\ 0{,}4425]$$

Questi risultati suggeriscono che una quota compresa fra il 30,8% e il 44,3% della popolazione di consumatori preferisca una scatola blu.

C.8.4. Procedure di test asintotiche

Quando si usa la stima di massima verosimiglianza è possibile usare tre procedure alternative di test; la scelta fra queste dipende da quale sia la più conveniente nel caso specifico. I test sono *asintoticamente equivalenti* e in grandi campioni forniscono gli stessi risultati. Supponiamo di essere interessati a verificare l'ipotesi nulla $H_0 : \theta = c$ rispetto all'ipotesi alternativa $H_1 : \theta \neq c$. Nella (C.22) abbiamo definito una statistica test usata per effettuare il test; qual è l'idea che ne sta alla base? Il test in pratica misura la distanza $\hat{\theta} - c$ fra la stima $\hat{\theta}$ e il valore ipotizzato sotto H_0, c. Questa distanza viene normalizzata per lo standard error di $\hat{\theta}$ per tenere conto della precisione con la quale abbiamo stimato θ. Se la distanza fra la stima $\hat{\theta}$ e il valore ipotizzato c è grande significa che nei dati è presente evidenza empirica contraria all'ipotesi nulla e se la distanza è abbastanza grande concluderemo che l'ipotesi nulla non è vera.

Per costruire la statistica test è possibile utilizzare altri modi di misurare la distanza fra $\hat{\theta}$ e c; ognuno dei tre principi di test ne adotta uno diverso.

C.8.4.a. Test del rapporto di verosimiglianza (RV)

Consideriamo la figura C.14, che illustra una funzione di logverosimiglianza, la stima di massima verosimiglianza $\hat{\theta}$ e il valore ipotizzato sotto H_0, c. Si noti che la distanza fra $\hat{\theta}$ e c si riflette nella distanza fra il valore della funzione di logverosimiglianza in corrispondenza della stima di massima verosimiglianza, $\log L(\hat{\theta})$, e quello in corrispondenza del valore sotto H_0, $\log L(c)$. Per motivi che saranno chia-

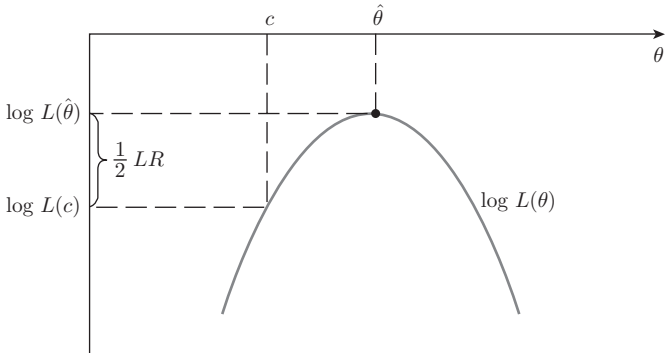

Figura C.14
Test del rapporto di verosimiglianza.

riti fra poco, indichiamo la differenza fra questi due valori della logverosimiglianza come $(1/2)RV$. Se la stima $\hat{\theta}$ è vicina a c, la differenza fra i valori di $\log L(\theta)$ sarà piccola; se $\hat{\theta}$ è lontana da c, la differenza sarà grande. Questa osservazione è alla base della **statistica del rapporto di verosimiglianza**, pari a due volte la differenza fra $\log L(\hat{\theta})$ e $\log L(c)$:

$$RV = 2[\log L(\hat{\theta}) - \log L(c)] \tag{C.25}$$

Usando alcune proprietà statistiche avanzate è possibile dimostrare che se l'ipotesi nulla è vera la statistica RV ha distribuzione chi quadro (si veda l'appendice B.3.6) con $J = 1$ gradi di libertà. In situazioni più generali, J è il numero di ipotesi sotto H_0 e può essere maggiore di 1. Se l'ipotesi nulla è falsa, la statistica RV assume valori elevati; di conseguenza per un livello di significatività α l'ipotesi nulla H_0 è rifiutata se $RV \geq \chi^2_{(1-\alpha;\, J)}$, dove $\chi^2_{(1-\alpha;\, J)}$ (illustrato nella figura C.15) è il percentile di livello $100(1-\alpha)\%$ della distribuzione chi quadro con J gradi di libertà. I percentili di livello 90%, 95% e 99% della distribuzione chi quadro per valori diversi del numero di gradi di libertà sono riportati nella tabella 3 dell'appendice D.

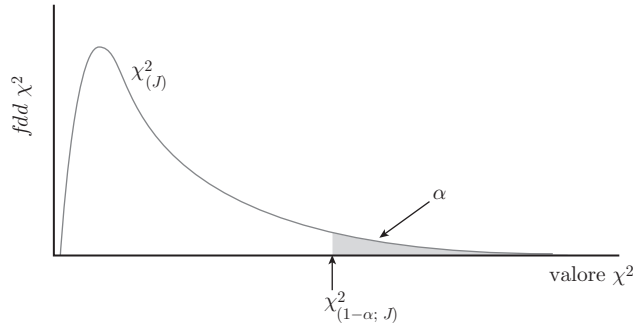

Figura C.15
Valore critico per una distribuzione chi quadro.

La formula (C.19) descrive la funzione di logverosimiglianza usata per stimare la quota p di valori $x = 1$ nella popolazione. Il valore di p che massimizza questa funzione è la quota campionaria $\hat{p} = \sum x_i/N$. Di conseguenza, il valore massimo della logverosimiglianza è dato da:

$$\begin{aligned}\log L(\hat{p}) &= \left(\sum_{i=1}^{N} x_i\right)\log\hat{p} + \left(N - \sum_{i=1}^{N} x_i\right)\log(1-\hat{p}) \\ &= N\hat{p}\log\hat{p} + (N - N\hat{p})\log(1-\hat{p}) \\ &= N[\hat{p}\log\hat{p} + (1-\hat{p})\log(1-\hat{p})]\end{aligned}$$

dove abbiamo usato la proprietà secondo la quale $\sum x_i = N\hat{p}$. Nel caso dello studio del colore migliore della scatola dei cereali, $\hat{p} = 0{,}375$ e $N = 200$; di conseguenza:

$$\begin{aligned}\log L(\hat{p}) &= 200[0{,}375 \times \log(0{,}375) + (1 - 0{,}375)\log(1 - 0{,}375)] \\ &= -132{,}3126\end{aligned}$$

Se assumiamo vera $H_0 : p = 0{,}4$, il valore della logverosimiglianza diventa:

$$\begin{aligned}\log L(0{,}4) &= \left(\sum_{i=1}^{N} x_i\right)\log(0{,}4) + \left(N - \sum_{i=1}^{N} x_i\right)\log(1 - 0{,}4) \\ &= 75 \times \log(0{,}4) + (200 - 75) \times \log(0{,}6) \\ &= -132{,}5750\end{aligned}$$

Il nostro problema è quello di capire se $-132{,}3126$ è significativamente diverso da $-132{,}5750$. La statistica test RV, definita dalla (C.25), è pari a:

$$RV = 2[\log L(\hat{p}) - \log L(0{,}4)] = 2 \times [-132{,}3126 - (-132{,}5750)] = 0{,}5247$$

Se l'ipotesi nulla $p = 0{,}4$ è vera, la statistica RV ha distribuzione $\chi^2_{(1)}$. Se scegliamo $\alpha = 0{,}05$, il valore critico del test è $\chi^2_{(0{,}95;\,1)} = 3{,}84$, il 95-esimo percentile della distribuzione $\chi^2_{(1)}$. Dato che $0{,}5247 < 3{,}84$, l'ipotesi nulla non deve essere rifiutata.

C.8.4.b. Test di Wald

La figura C.14 mostra chiaramente che la distanza $(1/2)RV$ dipende dalla curvatura della funzione di logverosimiglianza. La figura C.16 illustra due funzioni di

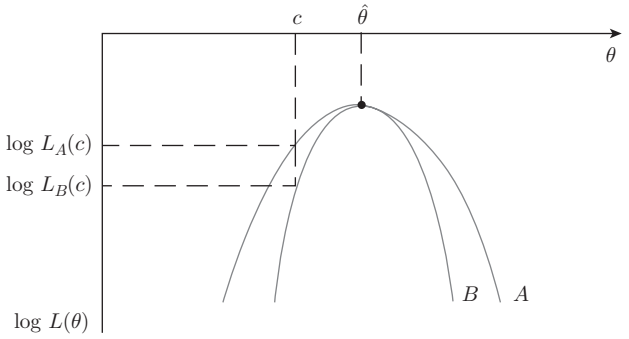

Figura C.16
Statistica di Wald.

verosimiglianza e la distanza $(1/2)RV$ fra i loro valori in corrispondenza del valore ipotizzato c. Le due logverosimiglianze hanno entrambe lo stesso massimo in $\hat{\theta}$, ma i loro valori in c sono diversi.

Nel caso della logverosimiglianza B, quella a curvatura maggiore, la distanza $\hat{\theta} - c$ si riflette in un valore più elevato di $(1/2)RV$. Questa osservazione suggerisce di costruire una statistica test ponderando la distanza $\hat{\theta} - c$ con una misura della curvatura della logverosimiglianza, la sua derivata seconda cambiata di segno. Questo è esattamente ciò che fa il test di Wald:

$$W = (\hat{\theta} - c)^2 \left[-\frac{d^2 \log L(\theta)}{d\theta^2} \right] \tag{C.26}$$

Il valore di questa statistica test è maggiore per la funzione di verosimiglianza B (quella a curvatura maggiore) che per la funzione di logverosimiglianza A (a curvatura minore).

Se l'ipotesi nulla è vera, la statistica di Wald (C.26) ha distribuzione $\chi^2_{(1)}$ e l'ipotesi nulla viene rifiutata se $W \geq \chi^2_{(1-\alpha;\,1)}$. In situazioni più generali possiamo verificare congiuntamente $J > 1$ ipotesi e in questo caso lavoreremo con una distribuzione chi quadro con J gradi di libertà, come quella rappresentata nella figura C.15.

Fra la curvatura della funzione di logverosimiglianza e la precisione dello stimatore di massima verosimiglianza esiste un collegamento: quanto maggiore è la curvatura, tanto minore sarà la varianza V definita dalla (C.23) e tanto più precisa la stima di massima verosimiglianza, a testimonianza del fatto che abbiamo più **informazione** sul parametro ignoto θ. Viceversa, quanta più informazione abbiamo su θ, tanto minore sarà la varianza dello stimatore $\hat{\theta}$. Usando questa idea possiamo considerare l'inversa della varianza V come una misura di informazione:

$$I(\theta) = -E \left[\frac{d^2 \log L(\theta)}{d\theta^2} \right] = V^{-1} \tag{C.27}$$

Questa notazione indica che la misura di informazione $I(\theta)$ è una funzione del parametro θ. Se sostituiamo la misura di informazione al posto della derivata seconda nella (C.26), la definizione della statistica di Wald, otteniamo:

$$W = (\hat{\theta} - c)^2 I(\theta) \tag{C.28}$$

In grandi campioni le due versioni della statistica di Wald coincidono. È interessante riformulare la (C.28) come:

$$W = (\hat{\theta} - c)^2 V^{-1} = (\hat{\theta} - c)^2 / V \tag{C.29}$$

Per implementare il test di Wald usiamo la stima della varianza:

$$\hat{V} = [I(\hat{\theta})]^{-1} \tag{C.30}$$

Calcolando ora la radice quadrata otteniamo esattamente la statistica t (C.22):

$$\sqrt{W} = \frac{\hat{\theta} - c}{\sqrt{\hat{V}}} = \frac{\hat{\theta} - c}{\text{se}(\hat{\theta})} = t$$

In altre parole, il test t è anche un test di Wald.

Nell'esempio relativo alla scelta fra scatola blu e scatola verde, sappiamo che la stima di massima verosimiglianza è $\hat{p} = 0{,}375$. Per calcolare la statistica test di Wald consideriamo:

$$I(\hat{p}) = \hat{V}^{-1} = \frac{N}{\hat{p}(1-\hat{p})} = \frac{200}{0{,}375(1-0{,}375)} = 853{,}3333$$

dove $V = p(1-p)/N$ e $\hat{V}$ sono stati ottenuti nel paragrafo C.7.3. Il valore campionario della statistica di Wald è dunque dato da:

$$W = (\hat{p} - c)^2 I(\hat{p}) = (0{,}375 - 0{,}4)^2 \times 853{,}3333 = 0{,}5333$$

In questo caso il valore della statistica di Wald è vicino a quello della statistica RV e la conclusione dei due test è la stessa. Si noti inoltre che quando si verifica un'unica ipotesi la statistica di Wald è pari al quadrato della statistica t, $W = t^2 = (-0{,}7303)^2 = 0{,}5333$.

C.8.4.c. Test dei moltiplicatori di Lagrange (LM)

La terza procedura di test basata sullo stimatore di massima verosimiglianza è quella dei moltiplicatori di Lagrange (LM, da *Lagrange Multiplier*). La figura C.17 illustra un altro modo di misurare la distanza fra $\hat{\theta}$ e c. La pendenza della funzione di logverosimiglianza, talvolta chiamata *score*, è data da:

$$s(\theta) = \frac{d \log L(\theta)}{d\theta} \tag{C.31}$$

La notazione $s(\theta)$ indica che la pendenza della funzione di logverosimiglianza dipende dal valore di θ. In corrispondenza del massimo la pendenza della funzione di logverosimiglianza è nulla, $s(\hat{\theta}) = 0$. Il test LM esamina la pendenza della logverosimiglianza in c. La logica del test è che se $\hat{\theta}$ è vicino a c, la pendenza della logverosimiglianza valutata in c, $s(c)$, dovrebbe essere prossima a zero. In effetti verificare $\theta = c$ è equivalente a verificare $s(c) = 0$.

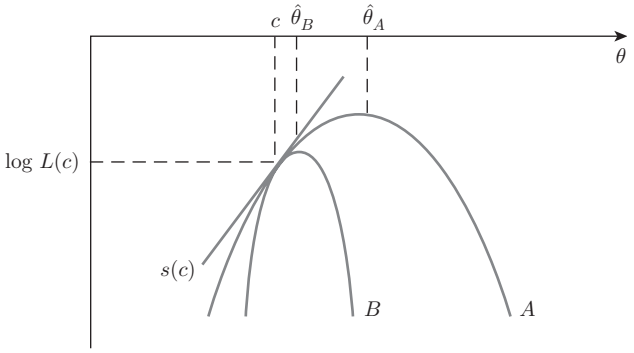

Figura C.17
Intuizione alla base del test dei moltiplicatori di Lagrange.

La differenza fra c e la stima di massima verosimiglianza $\hat{\theta}_B$ (che massimizza $\log L_B$) è minore della differenza fra c e $\hat{\theta}_A$. Al contrario del test di Wald, una curvatura maggiore della funzione di logverosimiglianza implica una differenza minore fra la stima di massima verosimiglianza e c. Se usiamo l'informazione $I(\theta)$

per misurare la curvatura (una curvatura maggiore implica più informazione), la statistica test dei moltiplicatori di Lagrange può essere scritta come:

$$LM = \frac{[s(c)]^2}{I(\theta)} = [s(c)]^2[I(\theta)]^{-1} \tag{C.32}$$

Il valore della statistica LM per la funzione di logverosimiglianza A (con curvatura minore) è maggiore di quello per la logverosimiglianza B (con curvatura maggiore). Se l'ipotesi nulla è vera, la statistica test LM (C.32) ha distribuzione $\chi^2_{(1)}$ e la regione di rifiuto è la stessa dei test RV e di Wald. I test LM, RV e di Wald sono asintoticamente equivalenti e in campioni sufficientemente numerosi conducono alla stessa conclusione.

Per poter applicare il test LM dobbiamo valutare la misura di informazione in corrispondenza del punto $\theta = c$; la statistica diventa dunque:

$$LM = [s(c)]^2[I(c)]^{-1}$$

Nei casi in cui la stima di massima verosimiglianza è difficile da ottenere (di solito in problemi più complicati) il test LM ha il vantaggio di non richiedere il calcolo di $\hat{\theta}$. D'altro canto, il test di Wald definito dalla (C.28) usa la misura d'informazione valutata in corrispondenza della stima di massima verosimiglianza $\hat{\theta}$:

$$W = (\hat{\theta} - c)^2 I(\hat{\theta})$$

e di solito viene preferito quando è facile calcolare la stima di massima verosimiglianza e la sua varianza. La statistica test del rapporto di verosimiglianza (C.25) richiede il calcolo della funzione di logverosimiglianza in corrispondenza sia della stima $\hat{\theta}$ sia del valore c ipotizzato sotto H_0. Dato che, come abbiamo già osservato, i tre test sono asintoticamente equivalenti, la scelta di quale utilizzare viene spesso effettuata sulla base di un semplice criterio di convenienza. Nelle situazioni più complesse, tuttavia, questa regola banale potrebbe non essere la migliore. Il test del rapporto di verosimiglianze è quello relativamente più affidabile nella maggior parte delle circostanze; per questo motivo, nel dubbio, è quello da preferire.

Nell'esempio della scatola verde/scatola blu, il valore dello *score*, basato sulla derivata prima riportata subito dopo la (C.19) e valutato in corrispondenza del valore ipotizzato c è dato da:

$$s(0{,}4) = \frac{\sum x_i}{c} - \frac{N - \sum x_i}{1 - c} = \frac{75}{0{,}4} - \frac{200 - 75}{1 - 0{,}4} = -20{,}8333$$

Il valore in c della misura d'informazione è:

$$I(0{,}4) = \frac{N}{c(1-c)} = \frac{200}{0{,}4(1 - 0{,}4)} = 833{,}3333$$

Il valore della statistica test LM è dunque dato da:

$$LM = [s(0{,}4)]^2[I(0{,}4)]^{-1} = [-20{,}8333]^2[833{,}3333]^{-1} = 0{,}5208$$

Nel nostro esempio, dunque, i valori delle statistiche test RV, LM e di Wald sono molto simili e portano alla medesima conclusione. Questo risultato non deve sorprendere, dato che la numerosità campionaria $N = 200$ è elevata e il modello da stimare è piuttosto semplice.

C.9. Altri risultati algebrici

C.9.1. Derivazione dello stimatore dei minimi quadrati

In questo paragrafo mostreremo come usare il principio dei minimi quadrati per mostrare che la media campionaria è uno stimatore della media della popolazione. Indichiamo con $y_1, y_2, \ldots, y_N$ il campione di osservazioni. La media della popolazione è $E(Y) = \mu$. Il principio dei minimi quadrati suggerisce di calcolare il valore di μ che minimizza:

$$S = \sum_{i=1}^{N} (y_i - \mu)^2$$

dove S è la somma dei quadrati degli scarti fra i valori osservati e μ.

La giustificazione alla base di questo approccio può essere dedotta dall'esempio seguente. Supponiamo che andiate a fare spese visitando alcuni negozi lungo una certa strada. La vostra intenzione è quella di fare acquisti in un negozio e tornare all'auto per depositarvi quanto appena acquistato; una volta fatto questo visitate un secondo negozio e tornate nuovamente all'auto e così via. Dopo aver visitato ciascun negozio tornate sempre all'auto. Dove vi converrebbe parcheggiare per minimizzare la distanza complessiva percorsa nei tragitti fra l'auto e i negozi visitati? Il vostro obiettivo è quello di minimizzare la *distanza* percorsa. Immaginate la strada lungo la quale si trovano i negozi come una retta orientata dotata di unità di misura. La distanza euclidea fra un negozio situato in y_i e la vostra auto situata al punto μ è data da:

$$d_i = \sqrt{(y_i - \mu)^2}$$

La distanza al quadrato, con la quale è matematicamente più semplice lavorare, è:

$$d_i^2 = (y_i - \mu)^2$$

Per minimizzare la distanza quadratica complessiva fra il parcheggio μ e i negozi situati in $y_1, y_2, \ldots, y_N$ dovreste minimizzare:

$$S(\mu) = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_i - \mu)^2$$

che è esattamente una funzione definita da una somma di quadrati. Il principio dei minimi quadrati, dunque, è in realtà equivalente al principio della minima *distanza quadratica*.

Dato che i valori campionari y_i sono noti, la funzione somma dei quadrati $S(\mu)$ dipende dal parametro ignoto μ. Sviluppando i quadrati e calcolando le sommatorie dei diversi termini, otteniamo:

$$S(\mu) = \sum_{i=1}^{N} y_i^2 - 2\mu \sum_{i=1}^{N} y_i + N\mu^2 = a_0 - 2a_1\,\mu + a_2\,\mu^2$$

Usando i dati nella tabella C.1 abbiamo:

$$a_0 = \sum y_i^2 = 14\,880{,}1909 \qquad a_1 = \sum y_i = 857{,}9100 \qquad a_2 = N = 50$$

Il grafico della parabola descritta dalla somma dei quadrati è illustrato nella figura C.18. Il valore di μ corrispondente al minimo sembra essere leggermente alla destra di 17. Calcoliamone ora il valore esatto.

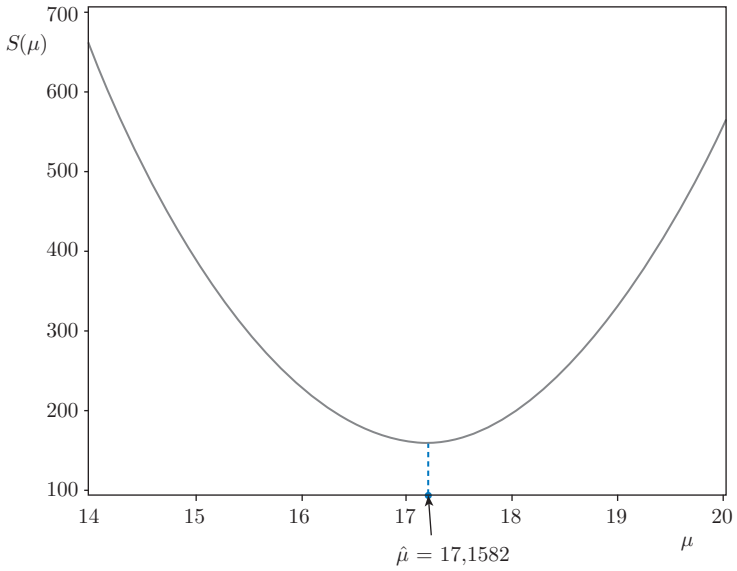

Figura C.18
Parabola della somma dei quadrati per i dati sulla larghezza del bacino.

Il valore di μ che minimizza $S(\mu)$ è la "stima dei minimi quadrati". Sappiamo che il minimo di una funzione si trova dove la pendenza, descritta dalla derivata prima della funzione, è nulla; di conseguenza, uguagliando la derivata prima di $S(\mu)$ a zero e risolvendo possiamo ottenere il valore esatto di μ che minimizza la somma dei quadrati. La derivata di $S(\mu)$ è data da:

$$\frac{dS(\mu)}{d\mu} = -2a_1 + 2a_2\mu$$

La stima dei minimi quadrati di μ, indicata con $\hat{\mu}$, può essere individuata ponendo a zero questa derivata:

$$-2a_1 + 2a_2\hat{\mu} = 0$$

Risolvendo rispetto a $\hat{\mu}$ otteniamo la formula della stima dei minimi quadrati:

$$\hat{\mu} = \frac{a_1}{a_2} = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}$$

La stima dei minimi quadrati della media di una popolazione è dunque data dalla media campionaria, $\bar{y}$. Questa formula è del tutto generale e può essere usata in qualsiasi campione. Ciò significa che lo stimatore dei minimi quadrati è:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} Y_i}{N} = \overline{Y}$$

Per i dati sulla larghezza del bacino contenuti nella tabella C.1:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{857{,}9100}{50} = 17{,}1582$$

La stima della larghezza media del bacino nella popolazione degli adulti statunitensi è dunque pari a 17,1582 pollici (43,6 centimetri circa).

C.9.2. Stimatori BLU

Una delle proprietà più importanti della media campionaria (che è anche uno stimatore dei minimi quadrati) è che si stratta del migliore fra tutti gli stimatori che sono sia *lineari* sia *corretti*. Il fatto che $\overline{Y}$ sia il migliore stimatore lineare e corretto (BLUE, da *Best Linear Unbiased Estimator*) spiega il motivo per il quale è così utilizzato. In questo contesto per migliore intendiamo lo stimatore con varianza minima fra tutti quelli lineari e corretti. Fra uno stimatore con varianza piccola e uno con varianza grande sceglieremo sempre il primo, perché in questo modo aumentano le possibilità di ottenere una stima vicina alla vera media nella popolazione μ. Questa importante proprietà dello stimatore dei minimi quadrati è vera *se* i valori campionari Y_i hanno tutti la stessa media μ e la stessa varianza σ^2, $Y_i \sim (\mu, \sigma^2)$, e sono incorrelati fra loro; non è necessario che la popolazione abbia distribuzione normale. L'osservazione che $\overline{Y}$ è BLUE è così importante che ne dimostreremo la validità.

La media campionaria è una media ponderata delle osservazioni:

$$\begin{aligned}\overline{Y} &= \sum_{i=1}^{N} Y_i/N = \frac{1}{N}\,Y_1 + \frac{1}{N}\,Y_2 + \ldots + \frac{1}{N}\,Y_N \\ &= a_1 Y_1 + a_2 Y_2 + \ldots + a_N Y_N \\ &= \sum_{i=1}^{N} a_i Y_i\end{aligned}$$

dove i pesi sono dati da $a_i = 1/N$. Le medie ponderate sono anche combinazioni lineari; per questo motivo definiamo la media campionaria come una **combinazione lineare**. In effetti, qualsiasi stimatore che possa essere formulato come $\sum_{i=1}^{N} a_i Y_i$ è uno stimatore lineare. Supponiamo per esempio che i pesi a_i^* siano costanti diverse da $a_i = 1/N$. Possiamo allora definire un altro stimatore lineare di μ come:

$$\tilde{Y} = \sum_{i=1}^{N} a_i^* Y_i$$

Per assicurarci che $\tilde{Y}$ sia diverso da $\overline{Y}$ definiamo:

$$a_i^* = a_i + c_i = \frac{1}{N} + c_i$$

dove i c_i sono costanti non tutte uguali a zero. Abbiamo così che:

$$\begin{aligned}\tilde{Y} &= \sum_{i=1}^{N} a_i^* Y_i = \sum_{i=1}^{N} \left(\frac{1}{N} + c_i\right) Y_i \\ &= \sum_{i=1}^{N} \frac{1}{N}\,Y_i + \sum_{i=1}^{N} c_i Y_i \\ &= \overline{Y} + \sum_{i=1}^{N} c_i Y_i\end{aligned}$$

Il valore atteso di questo nuovo stimatore è dato da:

$$E(\tilde{Y}) = E\left(\overline{Y} + \sum_{i=1}^{N} c_i Y_i\right) = \mu + \sum_{i=1}^{N} c_i E(Y_i)$$
$$= \mu + \mu \sum_{i=1}^{N} c_i$$

Perché lo stimatore $\tilde{Y}$ sia corretto è necessario che $\sum_{i=1}^{N} c_i = 0$. Dato che vogliamo confrontare la media campionaria con altri stimatori lineari e corretti, assumeremo che questa condizione sia verificata. Calcoliamo ora la varianza di $\tilde{Y}$. Lo stimatore lineare e corretto con varianza minima sarà considerato il migliore.

$$\mathrm{Var}(\tilde{Y}) = \mathrm{Var}\left(\sum_{i=1}^{N} a_i^* Y_i\right) = \mathrm{Var}\left[\sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right) Y_i\right] = \sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)^2 \mathrm{Var}(Y_i)$$
$$= \sigma^2 \sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)^2 = \sigma^2 \sum_{i=1}^{N}\left(\frac{1}{N^2} + \frac{2}{N} c_i + c_i^2\right)$$
$$= \sigma^2\left(\frac{1}{N} + \frac{2}{N}\sum_{i=1}^{N} c_i + \sum_{i=1}^{N} c_i^2\right) = \frac{\sigma^2}{N} + \sigma^2 \sum_{i=1}^{N} c_i^2 \quad \left(\text{dato che} \sum_{i=1}^{N} c_i = 0\right)$$
$$= \mathrm{Var}(\overline{Y}) + \sigma^2 \sum_{i=1}^{N} c_i^2$$

Di conseguenza, la varianza di $\tilde{Y}$ è sempre maggiore di quella di $\overline{Y}$ a meno che i valori di tutti i c_i siano nulli, nel qual caso $\tilde{Y} = \overline{Y}$.

C.10. Stima kernel della densità

Gli econometrici lavorano quasi sempre con dati estratti da distribuzioni ignote. Per fare un esempio, la figura C.19 illustra le distribuzioni empiriche di due campioni sotto forma di istogrammi. Le osservazioni relative alle variabili X e Y sono contenute nel file *kernel.dat*. La domanda che ci poniamo in questo paragrafo è se sia possibile stimare le funzioni di densità che hanno generato queste osservazioni. La conoscenza di queste distribuzioni è importante per l'inferenza statistica.

Esistono due strategie principali per stimare una densità: possiamo usare uno stimatore parametrico o uno stimatore kernel non parametrico. Nell'**approccio parametrico** ci basiamo su funzioni di densità con forme funzionali ben definite

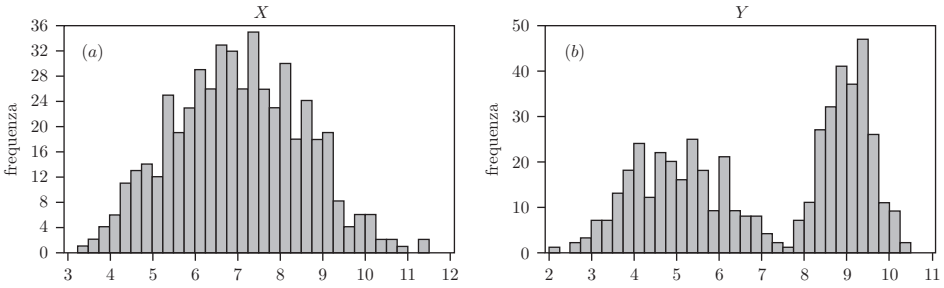

Figura C.19
Istogrammi di variabili: (*a*) variabile unimodale X, (*b*) variabile bimodale Y.

caratterizzate da alcuni parametri ignoti. Per esempio, la funzione di densità $f(\cdot)$ della distribuzione normale ha una forma funzionale particolare definita da due parametri – la media μ e lo scarto quadratico medio σ:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Una volta stimate la media e la varianza con $\hat{\mu}$ e $\hat{\sigma}$, possiamo sostituire queste stime nella formula della densità normale e ottenere:

$$\widehat{f(x)} = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right]$$

La figura C.20 illustra un'applicazione di questo approccio; le funzioni di densità normali sono sovrapposte agli istogrammi dei dati. Nel volume abbiamo applicato questo approccio parametrico alla discussione del Teorema del limite centrale (C.3.4) e dei modelli ARCH (capitolo 14).

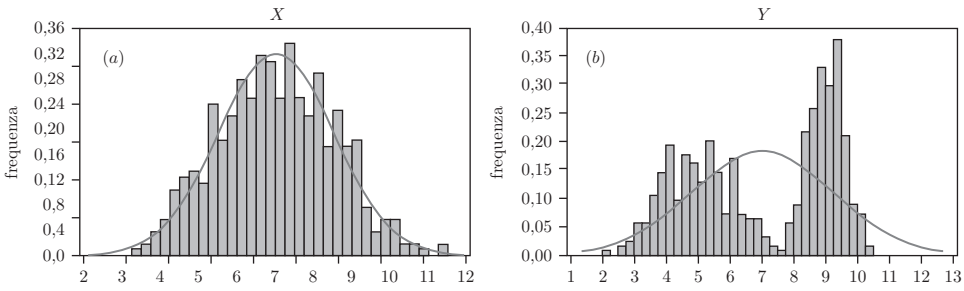

Figura C.20
Stimatore parametrico della densità: (*a*) variabile unimodale X, (*b*) variabile bimodale Y.

L'istogramma della variabile X, a sinistra nella figura C.20, è unimodale e la densità normale sembra adattarsi bene alla distribuzione dei dati. Al contrario, l'istogramma della variabile Y illustrato nel grafico di destra della figura C.20 è bimodale e la distribuzione normale approssima piuttosto male la funzione di densità sottostante. Anziché tentare di stimare questa densità usando qualche altra forma funzionale parametrica, useremo uno stimatore kernel non parametrico per catturare la forma dei dati usando una funzione continua regolare.

I **metodi non parametrici** non richiedono di ipotizzare una forma funzionale specifica (per esempio la formula della distribuzione normale) per approssimare la distribuzione. Al posto di questa ipotesi, useremo alcune funzioni chiamate **kernel** che stimano la densità ignota "regolarizzando" la distribuzione empirica.

La logica dell'approccio non parametrico può essere compresa intuitivamente riflettendo su come sono costruiti gli istogrammi. La figura C.21 illustra due istogrammi per il campione di osservazioni su Y. Quello a sinistra usa nove intervalli (in altre parole, l'istogramma è composto da nove rettangoli) con base larga 1, mentre quello a destra usa molti più intervalli, ciascuno dei quali ha base larga 0,1. Nel primo istogramma a ciascun rettangolo è associata una frequenza di osservazione maggiore perché un numero più elevato di osservazioni ricade nella base più larga. In particolare, se x_k è il punto centrale della base del k-esimo rettangolo e h è la sua larghezza, l'intervallo di valori rappresentati dal rettangolo è $x_k \pm h/2$ e la frequenza di osservazione n_k è il numero di osservazioni che appartiene a quell'intervallo. La somma di tutte le frequenze è pari alla numerosità

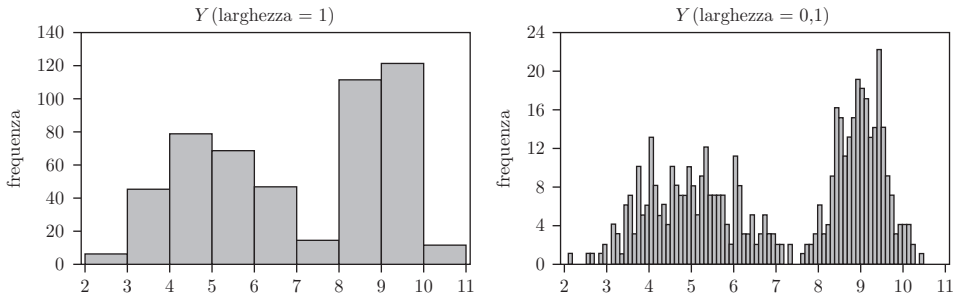

Figura C.21
Istogrammi con diverse larghezze dei rettangoli: (*a*) larghezza 1, (*b*) larghezza 0,1.

campionaria n, mentre la somma delle aree è pari a nh, dato che ogni area vale $n_k h$ e $\sum_k n_k h = nh$. Si noti inoltre che le forme dei due istogrammi sono simili ma che quello con rettangoli più larghi è "più regolare" (ha meno picchi e valli).

Possiamo interpretare l'istogramma come uno stimatore della funzione di densità, $\widehat{f(x)}$, dove x assume qualsiasi valore nel supporto di x e:

$$\widehat{f(x)} = \frac{1}{nh} \sum_{i=1}^{n} 1(A_i)$$

dove l'espressione $1(A_i)$ rappresenta una **funzione indicatrice** che assume valore 1 se A_i è vera e 0 altrimenti; A_i è la condizione che x_i appartenga allo stesso intervallo di x. Supponiamo per esempio di voler calcolare $\widehat{f(x)}$ per un valore di x che appartiene al k-esimo intervallo. In questo caso A_i è vera per tutti gli x tali che $x_k - h/2 < x_i < x_k + h/2$. Per le x nel k-esimo rettangolo, dunque, $\sum_{i=1}^{n} 1(A_i) = n_k$ e lo stimatore a istogramma della densità sarà pari a $\widehat{f(x)} = n_k/nh$. Il divisore nh garantisce che la somma delle aree di tutti i rettangoli sia pari a 1.

Consideriamo ora un altro stimatore di densità che invece di un numero predeterminato di intervalli con punti centrali x_k considera un solo intervallo con punto centrale x e conta il numero di osservazioni comprese fra i due estremi $x \pm h/2$. Se ripetiamo questo processo per tutti i valori di x, possiamo immaginare che lo stimatore crei un numero infinito di intervalli che si sovrappongono all'interno dell'insieme dei possibili valori di x. In questo caso lo stimatore della densità è definito da:

$$\widehat{f(x)} = \frac{1}{nh} \sum_{i=1}^{n} 1\left(x - \frac{h}{2} < x_i < x + \frac{h}{2}\right) = \frac{1}{nh} \sum_{i=1}^{n} 1\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right)$$

In pratica, sommando su tutte le osservazioni la funzione indicatrice ci garantisce che vengano "contate" solo le osservazioni rilevanti. Questa stima della densità tuttavia non è liscia perché a ciascuna osservazione è assegnato un peso uguale a 0 oppure a 1, in altre parole, o è compresa o è esclusa dall'intervallo, a seconda che la condizione specificata nella funzione indicatrice sia verificata o meno.

Supponiamo ora di sostituire questa semplice regola di conteggio con una funzione di ponderazione più sofisticata, chiamata **kernel**:

$$\widehat{f(x)} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

dove K è un kernel, h è un parametro di regolarizzazione chiamato **ampiezza di banda** e x è uno qualsiasi dei possibili valori della variabile casuale X. Esistono molte funzioni kernel; una di queste è la gaussiana, la cui espressione è:

$$K\left(\frac{x_i - x}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - x}{h}\right)^2\right]$$

La figura C.22 illustra l'applicazione di questo stimatore kernel alle osservazioni della variabile Y contenute nel file *kernel.dat* e considerando quattro possibili valori dell'ampiezza di banda. Si noti il modo in cui questo parametro controlla la forma della funzione di densità. Quanto minore è l'ampiezza di banda, tanto migliore sarà l'adattamento ai dati, ma questa relazione è caratterizzata da un trade-off fra il numero di "gobbe" catturate dalla stima e il suo grado di regolarità (in altre parole, quanto la stima sia "liscia"). Intuitivamente, diminuire l'ampiezza di banda equivale a diminuire la larghezza della base dei rettangoli nell'istogramma e la funzione kernel è simile a un "contatore" – anche se un kernel attribuisce un peso inferiore alle osservazioni più lontane dal punto x considerato. (Immaginate che la riduzione dell'ampiezza di banda vi sposti dall'istogramma di destra a quello di sinistra della figura C.21 e poi cercate di visualizzare l'effetto dell'uso del kernel in termini di lisciaggio dei rettangoli.) La funzione di densità kernel (gaussiana) con ampiezza di banda 0,4 sembra riuscire a catturare la bimodalità delle osservazioni.

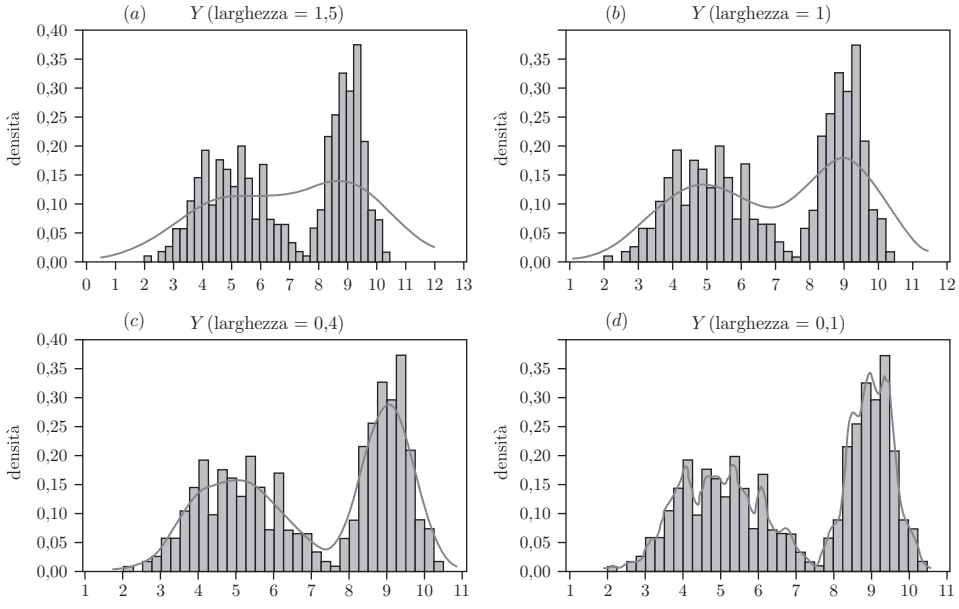

Figura C.22
Applicazione di uno stimatore non parametrico della densità:
(*a*) ampiezza di banda 1,5;
(*b*) ampiezza di banda 1;
(*c*) ampiezza di banda 0,4;
(*d*) ampiezza di banda 0,1.

Esiste una vasta letteratura che studia la scelta ottimale dell'ampiezza di banda così come diverse estensioni dei metodi non parametrici all'analisi di regressione. Un paio di riferimenti bibliografici utili per questi argomenti sono Pagan, A. e A. Ullah, *Nonparametric Econometrics*, Cambridge University Press, 1999; e Li Q. e J.S. Racine, *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, 2007.

C.11. Esercizi

Alla pagina web `http://online.universita.zanichelli.it/hillecon` sono disponibili le risposte agli esercizi indicati con un asterisco.

C.1 Supponete che $Y_1, Y_2, \ldots, Y_N$ sia un campione casuale tratto da una popolazione di media μ e varianza σ^2. Invece di usare tutte le N osservazioni, considerate uno stimatore di μ che usa solo le prime due:

$$Y^* = \frac{Y_1 + Y_2}{2}$$

(a) Mostrate che Y^* è uno stimatore lineare.
(b) Mostrate che Y^* è uno stimatore corretto.
(c) Calcolate la varianza di Y^*.
(d) Spiegate perché la media campionaria di tutte le N osservazioni è uno stimatore di μ migliore di Y^*.

C.2 Supponete che Y_1, Y_2, Y_3 sia un campione casuale tratto da una popolazione $\mathcal{N}(\mu, \sigma^2)$. Per stimare μ considerate lo stimatore ponderato:

$$\tilde{Y} = \frac{1}{2}\,Y_1 + \frac{1}{3}\,Y_2 + \frac{1}{6}\,Y_3$$

(a) Mostrate che $\tilde{Y}$ è uno stimatore lineare.
(b) Mostrate che $\tilde{Y}$ è uno stimatore corretto.
(c) Calcolate la varianza di $\tilde{Y}$ e confrontatela con quella della media campionaria $\overline{Y}$.
(d) Vi sembra che come stimatore di μ $\tilde{Y}$ sia equivalente a $\overline{Y}$?
(e) Assumendo $\sigma^2 = 9$, calcolate per ciascuno dei due stimatori la probabilità di trovarsi a meno di 1 da μ.

C.3* Le vendite orarie di pollo fritto di Louisiana Fried Chicken hanno distribuzione normale di media 2000 e scarto quadratico medio 500 (entrambi i parametri sono espressi in numero di pezzi venduti). Qual è la probabilità che in un giorno lavorativo di nove ore venga venduto un numero di pezzi maggiore di 20 000?

C.4 Il salario d'ingresso sul mercato del lavoro di un neolaureato in economia ha media 47 000 e scarto quadratico medio 8000 (entrambe le cifre sono espresse in dollari). Qual è la probabilità che un campione casuale di 40 neolaureati abbia un salario medio campionario maggiore di 50 000 dollari?

C.5* Il direttore di un negozio sta esaminando un nuovo sistema di contabilità la cui convenienza di costo dipende dal fatto che il saldo medio mensile di conto corrente sia superiore a 170 dollari. Viene selezionato casualmente un campione di 400 conti correnti. Il saldo medio campionario è 178 dollari e lo scarto quadratico medio è 65 dollari. Può il direttore concludere che il nuovo sistema sarà meno costoso?
(a) Rispondete a questa domanda effettuando una verifica d'ipotesi. Usate un livello di significatività $\alpha = 0{,}05$.
(b) Calcolate il p-value del test.

C.6 Un professore di econometria ritiene che per ogni ora di lezione frontale gli studenti dovrebbero aspettarsi di avere bisogno di due ore di studio ed esercitazione individuale. Ciò significa che per un corso con tre ore settimanali di lezione gli studenti dovrebbero lavorare autonomamente sei ore. Il professore sceglie casualmente otto studenti da una classe e chiede loro quante ore di studio hanno dedicato alla sua materia durante la settimana precedente. Le osservazioni campionarie così ottenute sono 1, 3, 4, 4, 6, 6, 8, 12.

(a) Se assumete che la popolazione abbia distribuzione normale e considerate un livello di significatività di 0,05, può il professore concludere che gli studenti stiano studiando in media più di sei ore?

(b) Costruite un intervallo di confidenza al 90% del numero medio di ore studiate a settimana nella popolazione.

C.7 Le moderne strategie di gestione del personale tentano di contenere i costi del lavoro assumendo e licenziando lavoratori per andare incontro alle fluttuazioni della domanda. I lavoratori appena assunti tuttavia non sono produttivi come quelli esperti. Supponete che gli operai con esperienza lavorativa pregressa possano assicurare una produzione di 500 pezzi al giorno. Un responsabile conclude che da un punto di vista di costo è conveniente mantenere la prassi corrente basata su un'elevata rotazione dei dipendenti se, dopo una settimana di addestramento, un nuovo assunto riesce a produrre 450 pezzi al giorno. Viene osservato un campione di $N = 50$ operai appena addestrati. Indichiamo con Y_i il numero di pezzi prodotti da ciascuno di essi in un giorno scelto casualmente. La media campionaria è $\bar{y} = 460$ e la stima dello scarto quadratico medio è $\hat{\sigma} = 38$.

(a) Effettuate un test per verificare se per un livello di significatività del 5% esiste evidenza empirica a sostegno della congettura secondo la quale le procedure di assunzione attuali sono economicamente convenienti. Fate molta attenzione a come formulate l'ipotesi nulla e l'ipotesi alternativa.

(b) In che cosa consisterebbe di preciso un errore di prima specie in questo esempio? Vi sembra che si tratti di un errore costoso?

(c) Calcolate il p-value del test.

C.8* Per confrontare fra loro un certo numero di schemi pensionistici per i propri dipendenti un'azienda multinazionale deve conoscere l'età media della propria forza lavoro. Assumete che l'età dei dipendenti segua una distribuzione normale. Dato che l'azienda ha migliaia di dipendenti, è necessario selezionarne un campione. Se lo scarto quadratico medio delle età è noto e pari a $\sigma = 21$ anni, quale deve essere la numerosità campionaria per garantire che una stima intervallare al 95% dell'età media abbia ampiezza non superiore a 4 anni?

C.9 Considerate la variabile casuale discreta Y che assume i valori $y = 1$, 2, 3 e 4 rispettivamente con probabilità 0,1, 0,2, 0,3 e 0,4.

(a) Tracciate il grafico della sua funzione di probabilità.

(b) Calcolate il valore atteso di Y.

(c) Calcolate la varianza di Y.

(d) Se estraiamo un campione casuale di numerosità $N = 3$ da questa distribuzione, quali sono la media e la varianza della media campionaria $\overline{Y} = (Y_1 + Y_2 + Y_3)/3$?

C.10 Questo esercizio consiste in una versione particolarmente semplice di esperimento di simulazione riferito all'esercizio C.9. Se necessario può essere svolto in gruppo o di fronte a tutta la classe. Chiedete a uno studente di preparare 10 foglietti identici numerati come quelli della tabella seguente.

1	2	2	3	3
3	4	4	4	4

(a) Estraete un foglietto a caso e registrate il suo valore, preferibilmente memorizzando ciascun risultato in un file di dati che esaminerete con il vostro software. Effettuate in totale 10 estrazioni, reinserendo nel mucchio il foglietto estratto e mescolando per bene. Confrontate la media dei valori ottenuti con il valore atteso calcolato al punto (b) dell'esercizio C.9. Estraete con reinserimento altri 10 numeri. Qual è la media dei 20 valori?

(b) Calcolate la varianza campionaria dei 20 valori ottenuti al punti (a). Confrontate questa varianza con la vera varianza ottenuta al punto (c) dell'esercizio C.9.

(c) Estraete con reinserimento tre foglietti a caso. Calcolate la media di questi $N = 3$ numeri, $\overline{Y} = (Y_1 + Y_2 + Y_3)/3$. Ripetete il processo almeno $NCAMP = 20$ volte, ottenendo $NCAMP$ medie, $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_{NCAMP}$. Calcolate la media e la varianza campionaria di questi $NCAMP$ valori e confrontatele con il valore atteso e la varianza della media campionaria ottenute al punto (d) dell'esercizio C.9.

(d) Memorizzate gli $NCAMP$ valori $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_{NCAMP}$ in un file di dati. Standardizzate questi valori sottraendo loro la vera media e dividendo per il vero scarto quadratico medio della media, ottenuti al punto (d) dell'esercizio C.9. Usate il vostro software per tracciare un istogramma. Discutete il Teorema del limite centrale e spiegate il collegamento fra il Teorema e la figura che avete creato.

(e) Ripetete i passaggi dei punti (c) e (d) usando $NCAMP$ campioni di più di $N = 3$ foglietti, magari cinque o sette. Commentate le differenze fra il nuovo istogramma e quello ottenuto al punto (d).

(f) Discutete il significato dei termini "variabilità campionaria" e "distribuzione campionaria" nel contesto degli esperimenti che avete svolto.

C.11 Nel famoso Fulton Fish Market di New York City le vendite di merlano (una varietà di pescato) variano da un giorno all'altro. Il file *fultonfish.dat* raccoglie osservazioni della quantità giornaliera venduta (in libbre) per un periodo di diversi mesi.

(a) Usando i dati relativi ai giorni di lunedì verificate l'ipotesi nulla che la quantità media venduta sia maggiore o uguale a 10 000 libbre al giorno rispetto all'alternativa che sia minore di 10 000 libbre. Usate un livello di significatività α del 5%. Accertatevi di (i) formulare dettagliatamente l'ipotesi nulla e l'ipotesi alternativa, (ii) descrivere la statistica test e la sua distribuzione, (iii) indicare la regione di rifiuto, facendone anche un grafico, (iv) formulare chiaramente la vostra conclusione e (v) calcolare il p-value del test. Fate anche un grafico che illustri il p-value.

(b) Assumete che le vendite giornaliere di martedì (X_2) e mercoledì (X_3) abbiano distribuzione normale di medie μ_2 e μ_3 e varianze σ_2^2 e σ_3^2. Assumete che le vendite di martedì e mercoledì siano indipendenti. Verificate l'ipotesi che le varianze σ_2^2 e σ_3^2 siano uguali fra loro rispetto all'alternativa che la varianza sia maggiore di martedì. Usate un livello di significatività α del 5%. Accertatevi di (i) formulare dettagliatamente l'ipotesi nulla e l'ipotesi alternativa, (ii) descrivere la statistica test e la sua distribuzione, (iii) indicare la regione di rifiuto, facendone anche un grafico, (iv) formulare chiaramente la vostra conclusione e (v) calcolare il p-value del test. Fate anche un grafico che illustri il p-value.

(c) Vogliamo verificare l'ipotesi che le vendite medie giornaliere di martedì e di mercoledì siano uguali rispetto all'alternativa che siano diverse. Effettuate questo test usando un livello di significatività del 5% e basandovi sui risultati ottenuti al punto (b) per scegliere la versione più appropriata del test (si veda il paragrafo C.7).

(d) Indichiamo le vendite giornaliere nei giorni di lunedì, martedì, mercoledì, giovedì e venerdì rispettivamente con X_1, X_2, X_3, X_4 e X_5. Assumete che $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ e che le vendite in giorni diversi siano indipendenti. Definite la quantità complessivamente venduta in una settimana come $W = X_1 + X_2 + X_3 + X_4 + X_5$. Derivate valore atteso e varianza di W. Accertatevi di descrivere con precisione i calcoli che avete fatto e di giustificare le vostre conclusioni.

(e) ♦ Con riferimento al punto (d) precedente, indichiamo $E(W)$ con μ. Supponiamo che stiate stimando μ con:

$$\hat{\mu} = \overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \overline{X}_4 + \overline{X}_5$$

dove $\overline{X}_i$ è la media campionaria per l'i-esimo giorno. Derivate la distribuzione di probabilità di $\hat{\mu}$ e costruite una stima intervallare approssimata (valida in grandi campioni) al 95% di μ. Giustificate la validità del vostro stimatore intervallare.