SAPIENZA
UNIVERSITÀ DI ROMA

INTERVALLI DI CONFIDENZA PER MEDIA E VARIANZA DI DISTRIBUZIONI NON NORMALI

Facoltà di Ingegneria dell'informazione, Informatica e Statistica
Laurea Triennale in Statistica Gestionale

Paola Maria Lepore
Matricola 2021453

Relatrice
Dott.ssa Stefania Gubbiotti

Anno Accademico 2023/2024

Tesi discussa il 11 Luglio 2024
di fronte a una commissione esaminatrice composta da:

Prof. Fulvio De Santis (presidente)
Prof. ssa Stefania Gubbiotti

INTERVALLI DI CONFIDENZA PER MEDIA E VARIANZA DI DISTRIBUZIONI NON NORMALI
Tesi compilativa. Sapienza Università di Roma

Questa tesi è stata composta con LaTeX e la classe Sapthesis.

Email dell'autore: lepore.2021453@studenti.uniroma1.it

Sommario

Questa tesi si propone di analizzare, sulla base del lavoro del professor Curto [3], dei nuovi intervalli di confidenza per la media, la differenza di medie, la varianza e il rapporto di varianze di popolazioni provenienti da distribuzioni non normali.
L'analisi si svolge mediante un confronto tra gli intervalli derivanti dalla teoria classica con quelli derivanti dai nuovi metodi proposti dal professor Curto [3].
Ci si è posti principalmente cinque obiettivi, tutti relativi all'individuazione di un ottimo intervallo di confidenza per i parametri individuati. A tal fine sono stati simulati in un caso 10000, nell'altro 1000, campioni di Monte Carlo di differenti dimensioni per diverse distribuzioni teoriche.
Dai risultati ottenuti è stato possibile individuare nei nuovi intervalli proposti da Curto [3] degli ottimi stimatori nel caso di varianza e rapporto di varianza, e dei buoni stimatori nel caso della media. Lo stesso risultato non è stato raggiunto in senso assoluto per la differenza di medie, per cui rimangono validi tutti gli intervalli proposti.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

Gli intervalli di confidenza sono un importante strumento statistico per la stima di parametri: essi forniscono un insieme di valori che, con una determinata probabilità, possono essere assegnati al parametro della popolazione.
Questa tesi si pone l'obiettivo di confrontare gli intervalli di confidenza per media e varianza di distribuzioni non normali determinati attraverso nuovi metodi proposti dal professor J. D. Curto [3] con i metodi classici.
Formalmente, dato un campione casuale $(X_1, X_2, \ldots, X_n)$ e il parametro θ da stimare, lo stimatore per intervallo di θ sarà: $P(L_1 < \theta < L_2) = 1-\alpha$, dove L_1 ed L_2 sono due statistiche campionarie. Allora se $(x_1, x_2, \ldots, x_n)$ è il campione osservato e l_1 ed l_2 i valori assunti dalle statistiche in corrispondenza di tale campione, l'intervallo di confidenza per θ di livello $(1-\alpha)$ sarà (l_1, l_2) [7].
Quindi l'intervallo di confidenza per il parametro θ è un intervallo di valori che contiene il parametro θ con una certa probabilità pari a $(1-\alpha)$. Il vantaggio nell'uso di stime intervallari risiede nel fatto che esse permettono di inglobare l'imprecisione dovuta al processo di stima vero e proprio, al contrario delle stime puntuali.
In pratica, se si ripetesse l'estrazione del campione numerose volte, e per ognuna di queste estrazioni si determinasse l'intervallo di confidenza, si otterebbe che l'$(1-\alpha)*100\%$ degli intervalli conterrà il vero valore del parametro di interesse. Più precisamente la probabilità con cui l'intervallo contiene il vero valore del parametro di interesse è detto livello di confidenza ed è pari a $(1-\alpha)$. Di conseguenza α, detto livello di significatività, è la probabilità di commettere un errore , cioè di considerare il parametro interno all'intervallo quando in realtà non vi appartiene.
Le caratteristiche principali dello stimatore per intervallo sono due: livello di confidenza ed ampiezza dell'intervallo, $A = L_2 - L_1$. La situazione ideale è tale per cui il livello di confidenza sia molto elevato così da avere una maggiore fiducia che l'intervallo contenga il vero parametro. Allo stesso tempo però, affinchè l'intervallo risulti effettivamente informativo, è necessario che l'ampiezza sia la più piccola possibile, in modo da avere informazioni più precise sul parametro. Per questi motivo i livelli di α più comunemente usati sono 0.025, 0.05 e 0,1, in quanto garantiscono un buon livello di confidenza, rispettivamente pari a 0.975, 0.95 e 0.9, ed una ragionevole ampiezza dell'intervallo.
I parametri di cui il professor Curto [3] si è interessato sono media, differenza di medie di due popolazioni aventi la stessa distribuzione, varianza e rapporto di varianze di

due popolazioni aventi la stessa distribuzione.
Partendo dal confronto tra le diverse distribuzioni, sarà possibile determinare quale sia l'intervallo migliore per ciascuno di questi parametri, assumendo come livello di confidenza nominale $(1 - \alpha) = 0.95$.
Il confronto si svolgerà in termini di probabilità di copertura stimate per ciascuno stimatore individuato. Per probabilità di copertura si intende la probabilità con cui l'intervallo di confidenza stimato contiene il vero parametro della popolazione, concetto strettamente legato al livello di confidenza.
Dai risultati sarà evidente che le distribuzioni non normali che destano maggiore interesse, in quanto presentano più criticità nella stima delle probabilità di copertura, saranno le distribuzioni leptocurtiche, in particolare la distribuzione χ^2_{n-1}.
Si procede simulando, nella prima parte, che riguarda la varianza, 10000 campioni di Monte Carlo di dimensioni differenti, $(10, 20, 30, 40, 50, 100)$, da diverse distribuzioni teoriche. Le distribuzioni scelte, oltre la normale standard, $N(0, 1)$, in quanto punto di riferimento per il confronto, sono: χ^2_1, $t(5)$ ed $Exp(1)$ per le distribuzioni leptocurtiche, e $Unif(0, 1)$, $Beta(3, 3)$ per le distribuzioni platicurtiche.
Nella seconda parte, che riguarda la media, sono stati simulati 1000 campioni di Monte Carlo di dimensioni differenti da diverse distribuzioni teoriche. Le dimensioni del campione sono $(10, 20, 30, 40, 50, 100)$, e, nel caso della differenza di medie, il secondo campione avrà le seguenti dimensioni $(15, 25, 35, 45, 55, 110)$. Le distribuzioni scelte saranno: $Unif(0, 1)$ e $Beta(3, 3)$ per le distribuzioni platicurtiche, χ^2_1 ed $Exp(1)$ per le distribuzioni leptocurtiche e $logN(0, 1)$.

Capitolo 2

Intervalli di confidenza per la varianza

Gli intervalli di confidenza per la varianza forniscono un insieme di valori all'interno dei quali ci si aspetta di trovare il valore vero di σ^2, la varianza.
Se si estrae un campione casuale di ampiezza n, $(X_1, X_2, \ldots, X_n)$, da una popolazione avente distribuzione normale, $X_i \sim N(\mu, \sigma^2)$, la variabile casuale V ha distribuzione χ^2_{n-1}, dove:

$$V = \frac{(n-1)S^2}{\sigma^2},$$

e S^2 è la varianza corretta, $S^2 = \frac{N}{(N-1)}\sigma^2$.
Allora si ricava dalla suddetta statistica il seguente stimatore intervallare di σ^2:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right] \tag{2.1}$$

dove $\chi^2_{1-\alpha/2}$ e $\chi^2_{\alpha/2}$ sono i quantili rispettivamente di livello $(1-\alpha/2)$ e $(\alpha/2)$ della distribuzione χ^2_{n-1}, con $(n-1)$ gradi di libertà. Per ricavare questo intervallo basta considerare la probabilità di copertura fissata pari al livello di confidenza:

$$P(\chi^2_{\alpha/2} < V < \chi^2_{1-\alpha/2}) = 1 - \alpha$$

Da cui, attraverso le seguenti uguaglianze, si ricavano gli estremi dell'intervallo precedentemente individuato:

$$1 - \alpha = P\left[\chi^2_{\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{1-\alpha/2}\right] = P\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right].$$

2.1 Stimatori della curtosi

L'intervallo di confidenza individuato non sempre risulta la scelta migliore, in quanto è fortemente sensibile anche alle più piccole violazioni dell'ipotesi di normalità. Sfruttando i risultati di Scheffè, si può dimostrare che la probabilità di copertura asintotica di tale intervallo per alcune distribuzioni non normali risulta molto bassa, come nel caso della distribuzione $t(5)$.
A tal proposito il professor Curto [3], riprendendo i risultati di Bonett [2] e Shoemaker [10], suggerisce la costruzione dell'intervallo con un metodo alternativo in modo tale che, nel caso di normalità, l'intervallo sia approssimativamente esatto, mentre nel caso di distribuzioni non normali la probabilità di copertura sia simile al livello di confidenza scelto.
Siano le variabili $(Y_1, Y_2, \ldots, Y_n)$ continue, indipendenti e identicamente distribuite con media, varianza e momento quarto finiti.
Detta $Var(y_i) = \sigma^2$, la varianza di σ^2 può essere espressa come segue:

$$Var(\sigma^2) = \frac{\sigma^4}{n}\left[\gamma_4 - \frac{(n-3)}{(n-1)}\right]$$

dove σ è la deviazione standard, γ_4 è l'indice di curtosi, cioè $\gamma_4 = \frac{\mu_4}{\sigma^4}$ e μ_4 è il momento quarto centrato della popolazione, cioè $\mu_4 = E[(X-\mu)^4]$. Per poter analizzare distribuzioni non necessariamente normali, risulta fondamentale stabilizzare la varianza, cioè ridurre la sua dipendenza dai dati e dalla loro media. Per questo motivo si ricorre alla trasformazione logaritmica di σ^2, da cui si ricava, attraverso un'approssimazione suggerita da Shoemaker [10], la seguente forma della varianza:

$$V[\ln(\sigma^2)] \approx \frac{1}{(n-1)}\left[\gamma_4 - \frac{(n-3)}{n}\right].$$

In particolare risulta necessario stimare le quantità incognite, quindi γ_4 e σ, il cui stimatore risulta pari a

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{(n-1)}.$$

La stima di γ_4 risulta più complessa e interessante.
L'indice di curtosi, γ_4, è un indice statistico che misura il grado di diversità tra la distribuzione di probabilità effettiva e il modello teorico rappresentato dalla distribuzione normale. In particolare la valutazione della curtosi si basa sulla forma della distribuzione stessa, concentrandosi sull'ampiezza delle code, quindi sui valori più estremi. L'indice di curtosi più usato è l'indice di Pearson, che, come visto in precedenza, è pari a $\frac{\mu_4}{\sigma^4}$: valori elevati dell'indice suggeriscono un allontanmento sempre maggiore dalla curva di Gauss.
Le distribuzioni non normali si distinguono in distribuzioni ipernormali, che eccedono rispetto alla normale attorno al valore medio, e iponormali, che presentano un minore addensamento attorno al valore medio [7].
In termini di curtosi invece, si possono distinguere i seguenti tipi di distribuzioni [4]:

- leptocurtiche: distribuzioni con code più allungate e pesanti rispetto alla normale, con elevata presenza di outliers e un picco più acuto. Si possono associare a distribuzioni ipernormali e l'indice di Pearson sarà maggiore di 3;

- mesocurtiche: distribuzioni che somigliano ad una normale, quindi le osservazioni si disporranno uniformemente attorno al valore medio. L'indice di Pearson sarà pari a 3;
- platicurtiche: distribuzioni tali da avere code più leggere e appiattite rispetto alla normale, quindi con una minore probabilità di osservare valori estremi in quanto i dati sono più uniformemente distribuiti, con un picco piuttosto piatto. Si possono associare a distribuzioni iponormali e l'indice di Pearson è minore di 3, generalmente prossimo allo 0.

Per poter stimare γ_4 Curto [3] ha individuato tre stimatori:

- $\hat{\gamma}_4(1)$, la curtosi campionaria, detto anche stimatore di Pearson, è lo stimatore naturale della curtosi ed è basato sulla media campionaria:

$$\hat{\gamma}_4(1) = \frac{n\sum_{i=1}^{n}(X_i - \hat{\mu})^4}{\left[\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right]^2}, \qquad \hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

- $\hat{\gamma}_4(2)$, stimatore proposto da Bonett [2], asintoticamente equivalente allo stimatore di Pearson, basato sulla media troncata:

$$\hat{\gamma}_4(2) = \frac{n\sum_{i=1}^{n}(X_i - \hat{\mu}_m)^4}{\left[\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right]^2}, \;\; \hat{\mu}_m = \frac{1}{(N-2s)}\left[\sum_{i=1}^{n} x_{(i)}\right], \;\; t = \frac{1}{[2(n-4)^{1/2}]}$$

  dove $\hat{\mu}_m$ è la media troncata con percentuale di taglio pari a t, utile ad individuare le s osservazioni più piccole e le s più grandi da eliminare [7].

- $\hat{\gamma}_4(3)$, stimatore basato sulla mediana, $\hat{\mu}_{med}$, indice più robusto rispetto alla media:

$$\hat{\gamma}_4(3) = \frac{n\sum_{i=1}^{n}(X_i - \hat{\mu}_{med})^4}{\left[\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right]^2}.$$

Per poter individuare lo stimatore migliore da adottare nella stima della varianza si considerano una serie di simulazioni campionarie per diverse distribuzioni teoriche, per ognuna delle quali si confronteranno le distorsioni di ciascuno dei tre stimatori, calcolate come $\hat{\gamma}_4(j) - \gamma_4(j)$.
Si simulano 10000 campioni di Monte Carlo di dimensioni differenti per ciascuna delle distribuzioni di interesse: i risultati sono riportati in Tabella 2.1.

Distorsione degli stimatori: $\hat{\gamma}_4(1)$, $\hat{\gamma}_4(2)$, $\hat{\gamma}_4(3)$.				
DISTRIBUZIONE	n	$\hat{\gamma}_4(1)$	$\hat{\gamma}_4(2)$	$\hat{\gamma}_4(3)$
$N(0,1)$	10	-0.542	-0.195	0.053
	20	-0.285	-0.169	0.048
	30	-0.194	-0.134	0.034
	40	-0.128	-0.077	0.048
	50	-0.104	-0.069	0.041
	100	-0.067	-0.052	0.006
$Unif(0,1)$	10	0.206	0.519	0.908
	20	0.125	0.194	0.541
	30	0.084	0.111	0.378
	40	0.060	0.083	0.291
	50	0.052	0.067	0.239
	100	0.027	0.033	0.128
$Beta(3,3)$	10	-0.087	0.214	0.496
	20	-0.003	0.072	0.318
	30	-0.001	0.032	0.222
	40	0.002	0.030	0.172
	50	0.008	0.026	0.143
	100	0.001	0.008	0.071
χ^2_1	10	-11.235	-9.152	-8.280
	20	-9.295	-7.608	-5.678
	30	-7.947	-6.496	-3.978
	40	-7.117	-5.452	-2.977
	50	-6.422	-4.927	-2.130
	100	-4.461	-2.994	0.189
$t(5)$	10	-6.170	-5.664	-5.437
	20	-5.403	-5.143	-4.937
	30	-4.937	-4.756	-4.586
	40	-4.639	-4.480	-4.355
	50	-4.330	-4.195	-4.086
	100	-3.567	-3.482	-3.419
$Exp(1)$	10	-5.849	-4.545	-3.959
	20	-4.565	-3.528	-2.335
	30	-3.837	-2.970	-1.447
	40	-3.223	-2.216	-0.747
	50	-2.824	-1.914	-0.246
	100	-1.834	-0.964	0.916

Tabella 2.1 Sono stati simulati 10000 campioni di Monte Carlo di diverse dimensioni provenienti dalle seguenti distribuzioni teoriche: $N(0,1)$, $Unif(0,1)$, $Beta(3,3)$, χ^2_1, $t(5)$ ed $Exp(1)$. Per ciascuna di esse è stata stimata la distorisione dei tre stimatori.

Dalla Tabella 2.1, si evince che, nel caso di distribuzioni mesocurtiche, quindi della normale, la distorsione è piuttosto bassa, in particolare all'aumentare della numerosità campionaria essa tende a diminuire, indicando così un miglioramento nella precisione di stima. Lo stimatore con minima distorsione, per qualsiasi dimensione

n del campione, è $\hat{\gamma}_4(3)$.
Per quanto riguarda le distribuzioni leptocurtiche, χ^2_1, $t(5)$ ed $Exp(1)$, è evidente che lo stimatore di Pearson, nonostante sia lo stimatore più usato, ha una distorsione negativa piuttosto elevata, risultando così il peggiore tra le alternative proposte. Anche in questo caso lo stimatore basato sulla mediana risulta il migliore avendo, il più delle volte, distorsione minima. In effetti media troncata e mediana sono stimatori più robusti rispetto alla media aritmetica, quindi subiscono una minore influenza dei valori anomali che, in questo tipo di distribuzione, sono generalmente molto presenti.
In questi casi la distorsione è quasi sempre negativa, quindi gli stimatori tendono, in media, a sottostimare il valore del parametro, restituendo cioè un valore dell'indice di curtosi più basso del valore reale.
In merito alle distribuzioni platicurtiche, $Unif(0,1)$ e $Beta(3,3)$, lo stimatore più appropriato sembra essere lo stimatore di Pearson che, per l'appunto, presenta valori della distorsione molto bassi. Al contrario di quanto accade nelle distribuzioni con code molto pesanti, in questo caso gli altri due stimatori tendono ad avere una distorsione positiva, seppur non eccessivamente elevata, che dimostra una sovrastima del valore vero del parametro, restituendo un valore più elevato della curtosi osservata.
La numerosità campionaria esercita una forte influenza nel calcolo della distorisone, infatti maggiore è il valore di n, più precisa sarà la stima, e i risultati in Tabella 2.1 lo dimostrano.
I valori minimi della distorsione sono restituiti con una numerosità pari a 100, la più alta considerata, ma è chiaro che, aumentando ancora di più il suo valore, le stime ottenute saranno ancora più precise con una distorsione tendente allo 0.
Da questo confronto si può concludere che non esiste uno stimatore migliore in assoluto, in quanto le caratteristiche della distribuzione, come l'ampiezza delle code, influiscono fortemente sull'accuratezza delle stime che riguardano la varianza.
Quindi, per ciascun tipo di distribuzione, si può suggerire uno stimatore più appropriato: per le distribuzioni leptocurtiche è consigliabile $\hat{\gamma}_4(3)$, a causa della robustezza della mediana, mentre per le distribuzioni platicurtiche è consigliabile lo stimatore di Pearson in quanto la probabilità di osservare valori anomali che influiscano sulla media è molto bassa.

2.2 Intervalli di confidenza basati sugli stimatori della curtosi

Per costruire gli intervalli di confidenza per la varianza il professor Curto [3] ha proposto una trasformazione logaritmica del parametro, e quindi dell'intervallo 2.1. La trasformazione logaritmica permette infatti di superare il problema della distribuzione di σ^2 che risulta positivamente asimmetrica, quindi caratterizzata da una coda ampia sulla destra e un valore della mediana maggiore della media, migliorando l'approssimazione normale. In accordo con Bartlett e Kendall [1], tra le condizioni necessarie quando si analizza la varianza di una variabile trasformata, vi è proprio la normalità della nuova distribuzione, in questo caso di $\ln(\sigma^2)$.
Per avere un'idea del tipo di correzione apportata dalla trasformazione logaritmica, consideriamo un campione casuale,$(X_1, X_2, \ldots, X_n)$, di variabili provenienti da una popolazione normale standard, con $n = 50$.

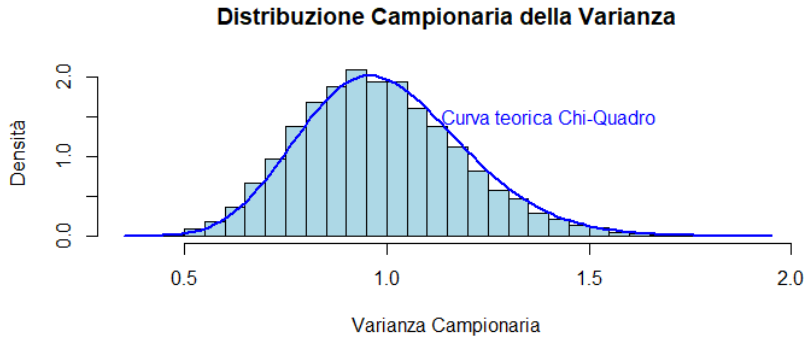

Figura 2.1 Distribuzione campionaria di σ^2 per campioni provenienti da una distribuzione normale standard con n = 50.

L'istogramma in Figura 2.1 risulta esemplificativo del ragionamento illustrato in merito alla varianza, in quanto è facilmente visibile come la coda destra della distribuzione sia più lunga rispetto alla coda sinistra. Infatti, la curva teorica della distribuione χ^2_{n-1}, rappresenta un'ottima approssimazione della distribuzione campionaria rappresentata dall'istogramma, allontanandosi così dalla condizione di normalità.
Applicando la trasformazione logaritmica alla varianza per costruire il nuovo intervallo di confidenza, risulterà evidente che $\ln(\sigma^2)$ si distribuirà approssimativamente come una normale.

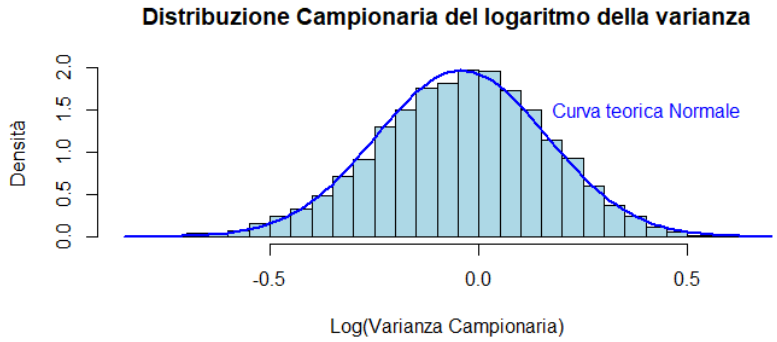

Figura 2.2 Distribuzione campionaria di $\ln(\sigma^2)$ per campioni provenienti da una distribuzione normale standard con n = 50.

Il grafico in Figura 2.2 rende evidente la necessità di adottare questa trasformazione della varianza: l'istogramma campionario del logaritmo delle varianze trova nella curva normale una buona approssimazione teorica. Tale distribuzione teorica sarà una normale avente come parametri la media e la deviazione standard delle varianze trasformate.
Quindi su questa nuova distribuzione si potranno costruire i due nuovi intervalli di confidenza proposti dal professor Curto [3]:

$$exp\left[\ln(\hat{\sigma}^2) - z_{1-\alpha/2}se(1) \;\; ; \;\; \ln(\hat{\sigma}^2) + z_{1-\alpha/2}se(1)\right]$$

$$exp\left[\ln(c\hat{\sigma}^2) - z_{1-\alpha/2}se(2) \;\; ; \;\; \ln(c\hat{\sigma}^2) - z_{1-\alpha/2}se(2)\right],$$

dove $z_{1-\alpha/2}$ è il quantile della distribuzione normale standardizzata di livello $(1-\alpha/2)$. In particolare $se(j)$ è lo standard error di $\ln(\sigma^2)$ determinato empiricamente e che si considera nelle due forme:

$$se(1) = \left\{\frac{1}{(n-1)}\left[\hat{\gamma}_4 - \frac{(n-3)}{n}\right]\right\}^{1/2}$$

$$se(2) = c\left\{\frac{1}{(n-1)}\left[\hat{\gamma}_4 - \frac{(n-3)}{n}\right]\right\}^{1/2}, \qquad c = \frac{n}{(n - z_{1-\alpha/2})}$$

dove c è l'aggiustamento campionario suggerito da Bonett [2] per rendere equivalenti le code della distribuzione.
Anche in questo caso si è interessati a un confronto tra i tre stimatori della varianza, ma in termini di probabilità di copertura dei nuovi intervalli proposti.
Il primo intervallo, denominato «normale», è determinato nella formula 2.1, quindi è individuato mediante il metodo classico.
Gli altri tre intervalli sono stati costruiti mediante gli stimatori della curtosi proposti dal professor Curto [3]. In particolare lo standard error di tali intervalli, come visto in precedenza, dipende da questi stimatori.

Probabilità di copertura stimate per gli $IC_{1-\alpha}$ per $\hat{\sigma}^2$ con $se(1)$					
DISTRIBUZIONE	n	NORMALE	$\hat{\gamma}_4(1)$	$\hat{\gamma}_4(2)$	$\hat{\gamma}_4(3)$
$N(0,1)$	10	0.949	0.890	0.901	0.925
	20	0.947	0.913	0.916	0.931
	30	0.952	0.924	0.926	0.937
	40	0.950	0.928	0.930	0.938
	50	0.952	0.935	0.936	0.943
	100	0.950	0.940	0.940	0.945
$Unif(0,1)$	10	0.994	0.946	0.957	0.970
	20	0.996	0.954	0.957	0.973
	30	0.997	0.956	0.958	0.972
	40	0.997	0.952	0.953	0.968
	50	0.997	0.957	0.959	0.970
	100	0.998	0.952	0.953	0.962
$Beta(3,3)$	10	0.979	0.923	0.936	0.946
	20	0.981	0.938	0.943	0.954
	30	0.982	0.943	0.945	0.956
	40	0.982	0.943	0.945	0.954
	50	0.983	0.945	0.946	0.953
	100	0.985	0.950	0.951	0.955
χ^2_1	10	0.646	0.643	0.743	0.779
	20	0.605	0.744	0.795	0.849
	30	0.591	0.773	0.811	0.874
	40	0.571	0.798	0.842	0.891
	50	0.574	0.819	0.852	0.904
	100	0.554	0.863	0.887	0.933
$t(5)$	10	0.874	0.810	0.840	0.855
	20	0.840	0.843	0.853	0.869
	30	0.826	0.855	0.861	0.872
	40	0.818	0.869	0.873	0.881
	50	0.799	0.877	0.880	0.886
	100	0.787	0.894	0.895	0.898
$Exp(1)$	10	0.771	0.719	0.788	0.818
	20	0.735	0.776	0.815	0.860
	30	0.705	0.809	0.839	0.880
	40	0.704	0.832	0.867	0.904
	50	0.703	0.848	0.873	0.912
	100	0.689	0.881	0.901	0.936

Tabella 2.2 Sono stati simulati 10000 campioni di Monte Carlo di diverse dimensioni provenienti dalle seguenti distribuzioni teoriche: $N(0,1)$, $Unif(0,1)$, $Beta(3,3)$, χ^2_1, $t(5)$ ed $Exp(1)$. Per ciascuna di esse è stata stimata la probabilità di copertura dei quattro intervalli proposti per $\hat{\sigma}^2$ con $se(1)$.

Anche in questo caso è facile notare che i risultati restituiti dalle stime campionarie in Tabella 2.2, differiscono notevolmente tra distribuzioni leptocurtiche e platicurtiche, dimostrando ancora una volta la forte influenza della forma della distribuzione su più aspetti.
La distribuzione normale presenta dei valori stimati che tendono a essere tanto più simili tanto più grande è la numerosità campionaria. Nonostante ciò l'intervallo che restituisce una probabilità di copertura maggiore, quindi generalmente migliore, è l'intervallo «normale», quindi il classico intervallo costruito sul campione non trasformato; tale probabilità risulta, approssimativamente equivalente alla probabilità nominale, avendo assunto $\alpha = 0.5$.
Un po' più articolato è il ragionamento sulle distribuzioni non normali.
Le stime delle probabilità di copertura nelle distribuzioni platicurtiche restituiscono valori molto elevati qualsiasi sia $\hat{\gamma}_4$, ma, anche in questo caso, il valore più elevato è restituito dall'intervallo «normale». Il comportamento di queste distribuzioni è detto conservativo in quanto il valore effettivo delle probabilità di copertura è maggiore del valore nominale, cioè del livello di confidenza dichiarato tra le ipotesi. Questo vale per entrambe le distribuzioni platicurtiche considerate, ma in particolar modo per la distribuzione uniforme, per la quale l'intervallo «normale» restituisce valori prossimi all'1.
Gli intervalli ottenuti mediante trasformazione della varianza non sempre hanno questa particolare proprietà, per esempio l'intervallo costruito sulla media risulta, di poco, il peggiore, mentre l'intervallo costruito sulla mediana risulta il migliore pur superando leggermente la probabilità di copertura nominale.
Piuttosto diverso è il ragionamento per le distribuzioni leptocurtiche, dette distribuzioni liberali, in quanto le probabilità effettive risultano notevolmente minori del livello nominale, $(1 - \alpha) = 0.95$.
Tra le distribuzioni considerate la più critica è la distribuzione χ^2_1, per la quale, in quasi tutti i casi, la probabilità di copertura restituita dalle simulazioni presenta valori particolarmente bassi. Questo non dovrebbe sorprendere in quanto la finalità stessa di questa analisi è quella di individuare dei nuovi metodi di stima di intervalli di confidenza per distribuzioni non normali, con una maggiore attenzione per le distribuzioni che presentano code più ampie, quindi per distribuzioni leptocurtiche. Proprio con questa finalità il professor Curto [3] ha proposto questi nuovi metodi di stime basati sulla trasformazione logaritmica della varianza.
Dall'analisi delle simulazioni in Tabella 2.2 sembra che i nuovi intervalli restituiscano valori più interessanti dell'intervallo «normale», in particolar modo per la distribuzione χ^2_1, i cui valori si avvicinano maggiormente alla probabilità nominale.
Poiché la media tende a risentire dei valori anomali, in modo particolare in una distribuzione asimmetrica come nel caso della distribuzione χ^2_1, con un'ampia coda a destra, il valore della media tenderà ad essere più spostato verso destra rispetto al valore della mediana, quindi quest'ultimo sarà minore del valore della media. Per questo motivo l'intervallo di confidenza che restituisce una probabilità stimata maggiore è quello basato sulla mediana.
Risultati analoghi si ottengono per le altre distribuzioni leptocurtiche tenute in considerazione, $t(5)$ ed $Exp(1)$, per le quali l'intervallo di confidenza migliore risulta sempre quello costruito sullo stimatore $\hat{\gamma}_4(3)$.
Quindi la tesi del professor Curto [3] risulta suffragata dai risultati campionari

ottenuti, dimostrando che una stima più accurata può essere ottenuta con questo nuovo intervallo di confidenza che tiene conto, in modo particolare, della forma tipica della distribuzione non normale di interesse.
Considerando $se(2)$, e quindi il secondo intervallo proposto da Curto [3], i cui risultati sono riportati in Tabella 2.3, si può giungere alle medesime conclusioni dei risultati in Tabella 2.2, con qualche piccolo appunto.
Il valore dello standard error determina la differenza sostanziale che vi è tra i due intervalli basati sulla trasformazione logaritmica, in quanto essi stessi dipendono dal valore dello stimatore considerato.
Di fatto, come è evidente dai risultati in Tabella 2.3, lo stimatore che restituisce intervalli con probabilità di copertura più elevata è, ancora una volta, $\hat{\gamma}_4(3)$, dimostrando che la mediana permette di ottenere uno stimatore più preciso, soprattutto nel caso di distribuzioni leptocurtiche.

Probabilità di copertura stimate per gli $IC_{1-\alpha}$ per $\hat{\sigma}^2$ con $se(2)$				
DISTRIBUZIONE	n	$\hat{\gamma}_4(1)$	$\hat{\gamma}_4(2)$	$\hat{\gamma}_4(3)$
$N(0,1)$	10	0.948	0.953	0.966
	20	0.939	0.942	0.954
	30	0.938	0.939	0.948
	40	0.939	0.941	0.948
	50	0.940	0.941	0.948
	100	0.941	0.942	0.945
$Unif(0,1)$	10	0.978	0.983	0.988
	20	0.968	0.971	0.981
	30	0.964	0.966	0.979
	40	0.964	0.965	0.976
	50	0.965	0.965	0.976
	100	0.957	0.950	0.965
$Beta(3,3)$	10	0.962	0.966	0.975
	20	0.956	0.959	0.972
	30	0.953	0.953	0.964
	40	0.952	0.953	0.964
	50	0.956	0.957	0.962
	100	0.951	0.952	0.957
χ^2_1	10	0.737	0.816	0.849
	20	0.769	0.822	0.872
	30	0.809	0.846	0.903
	40	0.820	0.855	0.904
	50	0.934	0.865	0.945
	100	0.869	0.894	0.936
$t(5)$	10	0.891	0.911	0.923
	20	0.874	0.885	0.897
	30	0.880	0.884	0.896
	40	0.889	0.893	0.899
	50	0.886	0.888	0.896
	100	0.899	0.901	0.902
$Exp(1)$	10	0.802	0.848	0.870
	20	0.818	0.850	0.884
	30	0.829	0.856	0.901
	40	0.846	0.875	0.911
	50	0.864	0.889	0.927
	100	0.885	0.906	0.942

Tabella 2.3 Sono stati simulati 10000 campioni di Monte Carlo di diverse dimensioni provenienti dalle seguenti distribuzioni teoriche: $N(0,1)$, $Unif(0,1)$, $Beta(3,3)$, χ^2_1, $t(5)$ ed $Exp(1)$. Per ciascuna di esse è stata stimata la probabilità di copertura dei quattro intervalli proposti per $\hat{\sigma}^2$ con $se(2)$.

La differenza dei due intervalli considerati risiede non solo nell'espressione dello standard error, ma anche nel fattore di correzione c di cui si tiene conto nel secondo caso.

Valori del fattore di correzione c						
n	10	20	30	40	50	100
c	1.244	1.109	1.070	1.052	1.041	1.020

Tabella 2.4 Sulla base delle simulazioni in Tabella 2.3, sono riportati i valori del fattore di correzione c usati nel nuovo intervallo di confidenza con $se(2)$.

Il fattore di correzione c rappresenta un piccolo aggiustamento dovuto alla dimensione del campione, in quanto campioni con numerosità più alta restituiscono stime più precise della probabilità di copertura dell'intervallo di confidenza. Per questo motivo c, pur rimanendo costante a parità di n, risulta avere un peso decrescente all'aumentare dell'ampiezza del campione, come si può notare dalla Tabella 2.4, permettendo di ottenere stime più accurate.
Dalle osservazioni della presente sezione, i risultati migliori sono stati ottenuti mediante il nuovo intervallo di confidenza proposto da Curto [3] che adotta l'aggiustamento campionario c e lo standard error $se(2)$: i risultati successivi saranno ottenuti mediante gli intervalli basati proprio su $se(2)$.
Il primo risultato di tali simulazioni, indipendentemente dallo standard error considerato, risulta il seguente: le probabilità di copertura effettive ottenute per gli intervalli basati sulla trasformazione logaritmica permettono di ottenere valori migliori, nel caso di distribuzioni leptocurtiche, rispetto agli intervalli costruiti sulla varianza non trasformata. In particolare l'intervallo più conservativo è quello costruito sullo stimatore della curtosi $\hat{\gamma}_4(3)$.
Per le altre distribuzioni rimane da preferire l'intervallo costruito sulla varianza non trasformata, in quanto, per la distribuzione normale restituisce valori effettivi circa pari al valore nominale, mentre per le distribuzioni platicurtiche restituisce valori effettivi, in alcuni casi, più elevati del valore nominale, $(1 - \alpha) = 0.95$.
L'impiego della trasformazione logaritmica ha come obiettivo proprio quello di ridurre l'asimmetria della distribuzione, in questo modo esse tendono anche a ridurre il valore dell'indice di curtosi. Questo risulta fondamentale perchè, come nel caso della distribuzione χ^2_1, ridurre la curtosi implica che sarà necessaria una minore numerosità campionaria per ottenere una buona probabilità di copertura.
Tutto ciò è evidente dai risultati campionari, prendendo come esempio la colonna «normale» della Tabella 2.2 e la Tabella 2.3 per la distribuzione χ^2_1, per una numerosità piccola il valore della probabilità di copertura risulta piuttosto basso sia con un intervallo classico sia con gli intervalli trasformati. Ma, aumentando il valore di n, la probabilità effettiva dell'intervallo classico diminuisce mentre la probabilità effettiva dell'intervallo di Curto aumenta, così che la differenza tra le due si amplifichi, dimostrando la maggiore precisione del nuovo intervallo.
Questo ragionamento non può essere esteso anche alle distribuzioni non leptocurtiche, in quanto, osservando i risultati delle medesime simulazioni per la distribzuzione $Beta(3, 3)$, presa come esempio, le probabilità migliori sono restituite dall'intervallo non trasformato.

Capitolo 3

Intervalli di confidenza per il rapporto tra varianze

Date due variabili casuali $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$, si è interessati a confrontare le due popolazioni in termini di varianza.
Per confrontare le varianze di due popolazioni provenienti dalla stessa distribuzione se ne considera il rapporto: se $\frac{\sigma_2^2}{\sigma_1^2} = 1$, le varianze sono uguali.
Si procede quindi ad individuare un buon intervallo di confidenza per il nuovo parametro di inetresse: $\frac{\sigma_2^2}{\sigma_1^2}$.
Date le varianze corrette S_1^2 ed S_2^2 e le numerosità campionarie dei due campioni n_1 ed n_2, la statistica

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

ha distribuzione F di Fisher[1], $F(r_1, r_2)$, dove $r_1 = (n_1 - 1)$ ed $r_2 = (n_2 - 1)$ sono i gradi di libertà.
Tutto questo è vero se la distribuzione delle variabili X_1 ed X_2 è normale, infatti se viene meno l'ipotesi di normalità, si procede approssimando la distribuzione F di Fisher correggendo opportunamente i gradi di libertà, come suggerito dall'approccio di Shoemaker [10]. Tale approccio consiste nel normalizzare la statistica F e poi abbinare i momenti della statistica F normalizzata ai momenti di una distribuzione normale. Per il teorema del limite centrale S^2, poiché è approssimativamente la media di variabili indipendenti e identicamente distribuite, avrà distribuzione normale per una numerosità campionaria abbastanza grande.
Come nel caso di una singola varianza, bisogna superare il problema dell'asimmetria della distribuzione di S^2 per una numerosità campionaria non abbastanza grande. Per questo motivo, analogamente alla Sezione 2.2, seguendo l'approccio di Bartelett e Kendall [1], si procede con la trasformazione logaritmica della varianza per rimuovere l'asimmetria e migliorare l'ipotesi di normalità.

[1] La distribuzione F di Fisher -Snedecor è una distribuzione asimmetrica a destra che dipende da due parametri, cioè i gradi di libertà delle distribuzioni χ^2 su cui è costruita. La variabile aleatoria F è ottenuta come rapporto tra due variabili aventi distribuzione χ^2 divise per i loro gradi di libertà. È usata principalmente per il confronto tra varianze di due variabili, come in questo caso [6].

Si considera la trasformazione logaritmica della statistica F di interesse:

$$\ln(F) = \ln(S_1^2) - \ln(S_2^2) - \ln(\sigma_1^2) + \ln(\sigma_2^2),$$

che, sotto le solite ipotesi, ha distribuzione normale con $var(\ln(F)) = \frac{2}{r_1} + \frac{2}{r_2}$.
Più in generale, per le distribuzioni non necessariamente normali, si assume:

$$var(\ln(F)) = var(\ln(S_1^2)) + var(\ln(S_2^2)).$$

Il metodo adottato da Shoemaker [10] e poi ripreso da Curto [3], è un adattamento della tecnica Satterwhite[2].
A questo punto si associano ai momenti normalizzati di F i momenti di una normale. Ponendo $\frac{2}{r_i} = var(\ln(S_i^2))$ e risovendo per r_i, si ottiene un aggiustamento necessario per i gradi di libertà:

$$r_i = \frac{2n_i}{\frac{\mu_4}{\sigma^4} - \frac{(n_i-3)}{(n_i-1)}} \tag{3.1}$$

dove μ_4 è il momento quarto centrato, σ è la deviazione standard ed n_i è la numerosità campionaria del campione i.
Le simulazioni campionarie in Tabella 3.1 permettono di confrontare una serie di intervalli di confidenza proposti per il parametro $\frac{\sigma_2^2}{\sigma_1^2}$, in modo da stabilire quale sia il milgiore tra essi. Gli intervalli considerati in tali simulazioni sono i seguenti.
L'intervallo denominato «normale» è l'intervallo di confidenza per il rapporto di varianze di due popolazioni provenienti dalla stessa distribuzione, sotto l'ipotesi di normalità. È l'intervallo classico di livello $(1-\alpha)$ costruito a partire dalla statistica F:

$$1 - \alpha = P\left[f_1 < F < f_2\right] = P\left[f_1 < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_2\right] = P\left[\frac{S_2^2}{S_1^2} f_1 < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2} f_2\right]$$

Da cui risulta il seguente intervallo di confidenza:

$$[IC]_{1-\alpha} = \left]\frac{S_2^2}{S_1^2} \cdot f_1; \frac{S_2^2}{S_1^2} \cdot f_2\right[,$$

dove $f_1 = f_{\alpha/2}$ ed $f_2 = f_{1-\alpha/2}$ sono i due quantili della distribuzione F-Fisher con $(n_1 - 1)$ ed $(n_2 - 1)$ gradi di libertà, F_{n_1-1,n_2-1}.
Il secondo intervallo di confidenza considerato, detto «F», è lo stesso intervallo appena descritto ma con un piccolo aggiustamento che riguarda i gradi di libertà; tale fattore correttivo è pari al valore r_i suggerito da Shoemaker [10], riportato nella formula 3.1.
I restati intervalli sono stati costruiti basandosi sulla trasformazione logaritmica della varianza.
Poiché $\ln(\sigma_2^2/\sigma_1^2) = \ln(\sigma_2^2) - \ln(\sigma_1^2)$, l'intervallo di confidenza per il rapporto tra varianze basato sulla trasformazione logaritmica di $[\ln(\sigma_2^2) - \ln(\sigma_1^2)]$, sarà:

$$exp\left\{\left[\ln(c_2\hat{\sigma}_2^2) - \ln(c_1\hat{\sigma}_1^2)\right] \pm z_{1-\alpha/2}\sqrt{se(2)_1^2 + se(2)_2^2}\right\}$$

[2]Il metodo Satterwhite è un metodo statistico usato nell'analisi della varianza (ANOVA) per la correzione dei gradi di libertà nel caso di eteroschedasticità

dove $z_{1-\alpha/2}$ è il quantile della distribuzione normale standardizzata di livello $(1-\alpha/2)$, con $\alpha = 0.05$, c_1 e c_2 sono rispettivamente i fattori di correzione per la numerosità campionaria n_1 e n_2 definiti nella sezione 2.2 ed $se(2)$ è lo standard error considerato nella Sezione 2.2.

Probabilità di copertura stimate per gli $IC_{1-\alpha}$ per σ_2^2/σ_1^2 con $se(2)$						
DISTRIBUZIONE	n	NORMALE	F	$\hat{\gamma}_4(1)$	$\hat{\gamma}_4(2)$	$\hat{\gamma}_4(3)$
$N(0,1)$	10	0.952	0.950	0.995	0.995	0.997
	20	0.951	0.947	0.991	0.992	0.995
	30	0.951	0.949	0.992	0.992	0.995
	40	0.951	0.950	0.992	0.993	0.995
	50	0.948	0.948	0.991	0.991	0.993
	100	0.950	0.950	0.994	0.994	0.994
$Unif(0,1)$	10	0.990	0.961	0.998	0.998	0.999
	20	0.994	0.955	0.997	0.998	0.999
	30	0.996	0.956	0.997	0.998	0.999
	40	0.996	0.957	0.997	0.997	0.998
	50	0.997	0.951	0.997	0.997	0.998
	100	0.997	0.947	0.995	0.995	0.996
$Beta(3,3)$	10	0.976	0.950	0.996	0.997	0.998
	20	0.979	0.951	0.994	0.995	0.997
	30	0.980	0.949	0.995	0.994	0.996
	40	0.979	0.948	0.994	0.996	0.996
	50	0.983	0.952	0.996	0.996	0.997
	100	0.982	0.949	0.994	0.994	0.996
χ_1^2	10	0.592	0.797	0.876	0.921	0.961
	20	0.516	0.802	0.879	0.909	0.965
	30	0.476	0.795	0.878	0.900	0.961
	40	0.432	0.765	0.871	0.888	0.957
	50	0.395	0.748	0.869	0.884	0.957
	100	0.263	0.645	0.824	0.847	0.935
$t(5)$	10	0.869	0.938	0.984	0.987	0.992
	20	0.838	0.946	0.981	0.984	0.989
	30	0.823	0.951	0.984	0.985	0.987
	40	0.819	0.955	0.985	0.986	0.989
	50	0.810	0.954	0.985	0.986	0.988
	100	0.775	0.955	0.988	0.988	0.989
$Exp(1)$	10	0.767	0.984	0.951	0.968	0.986
	20	0.728	0.919	0.958	0.968	0.986
	30	0.720	0.932	0.963	0.971	0.989
	40	0.710	0.937	0.968	0.974	0.992
	50	0.706	0.940	0.974	0.979	0.995
	100	0.689	0.949	0.983	0.987	0.997

Tabella 3.1 Sono stati simulati 10000 campioni di Monte Carlo di diverse dimensioni provenienti dalle seguenti distribuzioni teoriche: $N(0,1)$, $Unif(0,1)$, $Beta(3,3)$, χ_1^2, $t(5)$ ed $Exp(1)$. Per ciascuna è stata stimata la probabilità di copertura degli intervalli proposti per σ_2^2/σ_1^2.

Dai risultati ottenuti e illustrati in Tabella 3.1, risulta necessario, anche in questo caso, distinguere le distribuzioni in relazione all'ampiezza delle loro code.
Per la distribuzione normale tutti gli intervalli risultano piuttosto conservativi, no-

nostante sia evidente che le probabilità di copertura maggiori sono state ottenute mediante la trasformazione logaritmica.
Anche per le distribuzioni platicurtiche tutti gli intervalli risultano conservativi, e in particolare l'aggiustamento per i gradi di libertà non sembra avere alcuna necessità pratica, essendo le stime effettive approssimativamente pari alle stime nominali.
Le distribuzioni leptocurtiche hanno bisogno di un'analisi più approfondita, soprattutto per quanto riguarda gli intervalli basati sui dati non trasformati. In particolare per le distribuzioni $t(5)$ ed $Exp(5)$, l'aggiustamento di Shoemaker rende le stime più conservative, migliorando il risultato, ma il valore della probabilità effettiva di partenza non risulta eccessivamente basso.
Al contrario, per la distribuzione χ^2_1, che per sua natura ha code più ampie, questo aggiustamento diventa necessario dal momento che le probabilità effettive ottenute inizialmente risultano eccessivamente basse, ben lontane dal valore nominale.
Il fondamentale effetto correttivo dell'aggiustamento nel caso della distribuzione χ^2_1 è evidente anche dal confronto tra i grafici nelle Figure 3.1 e 3.2.

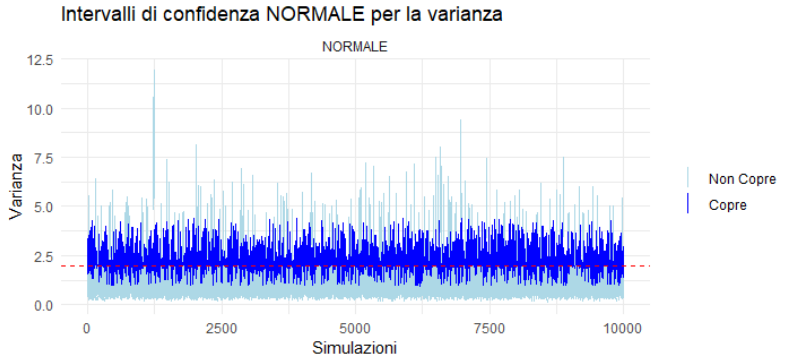

Figura 3.1 Rappresentazione grafica degli intervalli di confidenza per il rapporto di varianze sotto l'ipotesi di normalità.

Figura 3.2 Rappresentazione grafica degli intervalli di confidenza per il rapporto di varianze sotto l'ipotesi di normalità con la correzione 3.1.

Nel grafico in Figura 3.1 sono rappresentati gli intervalli costruiti con il metodo classico basato sul rapporto delle varianze corrette, senza l'applicazione di alcuna trasformazione. L'analisi visiva permette di notare che gli intervalli che non conten-

gono il vero valore del parametro $\frac{\sigma_2^2}{\sigma_1^2}$, rappresentati in celeste, sono numerosi.
Al contrario, nel grafico in Figura 3.2, in cui sono rappresentati gli intervalli costruiti con metodo classico ma con la correzione dei gradi di libertà in fomula 3.1, c'è una netta prevalenza del colore blu, associato agli intervalli contenenti il vero valore del parametro di interesse.
Risulta evidente da questo confronto l'effetto fondamentale dell'aggiustamento di Shoemaker [10], soprattutto nel caso di distribuzioni leptocurtiche, come la distribuzione χ_1^2.
La finalità vera e propria di questa correzione è quella di rendere la statistica F più robusta, così che il test di omogeneità della varianza, basato su tale statistica, abbia una potenza maggiore rispetto agli altri test possibili, proprio come analizzato da Shoemaker [10].
Quindi per le distribuzioni platicurtiche e per la distribuzione normale, gli intervalli proposti da Curto [3] costituiscono un efficace strumento di stima per le probabilità di copertura per la loro caratteristica conservatività.
Per le distribuzioni leptocurtiche, più liberali, risulta necessaria la trasformazione logaritmica per poter ottenere delle buone stime effettive usando il metodo classico, ma i risultati migliori sono ottenuti con l'intervallo trasformato che sfrutta lo stimatore $\hat{\gamma}_4(3)$, basato sulla mediana, proposto appunto dal professor Curto [3].

Capitolo 4

Intervalli di confidenza per la media

L'obiettivo di questa sezione è individuare un buon intervallo di confidenza per il parametro μ, la media aritmetica della popolazione, concentrandosi sulle distribuzioni non normali.
I dati reali sono raramente, per loro natura, distribuiti secondo una normale, risulteranno distorti, quindi diventa necessario ricorrere a una serie di trasformazioni per poter garantire una migliore interpretazione del fenomeno.
Nel caso di dati distorti e che, in particolare, presentano una distorsione sulla destra della distribuzione, si può ricorrere, come suggerito da Curto [3], alla trasformazione logaritmica. Tale trasformazione risulta la modalità più adatta per ottenere dei dati distribuiti, almeno approssimativamente, come una normale.
In ogni caso l'oggetto di interesse di questa analisi rimane la media dei dati originali, ricorrendo alla trasformazione logaritmica per migliorarne l'interpretazione, e per costruire dei nuovi intervalli.
Prima di costruire l'intervallo bisogna considerare alcune osservazioni sulla media dei dati originali.
Si considerano una variabile casuale X avente media μ_{aX} e varianza σ_X^2 e Y, la variabile trasformata mediante logaritmo, $Y = \log(X)$, avente media μ_{aY}.
Per poter sfruttare al meglio la relazione che lega queste due variabili, si considera la media geometrica di X e quindi la seguente uguaglianza:

$$\mu_{gX} = exp(\mu_{aY}),$$

che deriva dalle proprietà di μ_{gX}, in particolare dal suo logaritmo [7], che risulta:

$$\log(\mu_{gX}) = \log\left(\sqrt[n]{\prod_{i=1}^{n} X_i}\right) = \log\left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}.$$

Allora, poichè la media aritmetica della variabile trasformata, Y, è

$$\mu_{aY} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} log(X_i) = \frac{1}{n} log\left(\prod_{i=1}^{n} X_i\right) = log\left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}},$$

risultano valide le seguenti uguaglianze:

$$\log(\mu_{gX}) = \log\left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}} = \mu_{aY};$$

$$\mu_{gX} = \sqrt[n]{\prod_{i=1}^{n} X_i} = exp\left[log\left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}\right] = exp(\mu_{aY}).$$

Da tali equivalenze è evidente che la media geometrica dei dati originali, μ_{gX}, corrisponde all'esponenziale della media aritmetica dei dati trasformati, μ_{aY}.
Come suggerisce Feng [5], gli obiettivi della trasformazione logaritmica sono principalmente due:

- ridurre la distorsione: data una variabile originale con distribuzione log-normale, la variabile trasformata avrà distribuzione normale; questo vale, in generale, per diversi tipi di distribuzioni, nonostante in alcuni casi si potrebbe ottenere l'effetto opposto.
- ridurre la variabilità dovuta agli outliers: questo è vero perchè la media geometrica è tendenzialmente più robusta della media aritmetica, per cui risente meno della presenza di valori anomali.

Poichè l'obiettivo della trasformazione logaritmica che è più conforme allo scopo di questa sezione rimane quello di trovare una variabile con distribuzione normale, o approssimativamente normale, si considera una variabile di partenza avente distribuzione lognormale.
Data $X \sim logN(\mu, \sigma^2)$, la variabile trasformata $Y = log(X)$ avrà distribuzione normale, $Y \sim N(\mu, \sigma^2)$. Il valore atteso di X, è

$$E(X) = \mu_{aX} = exp\left(\mu + \frac{\sigma^2}{2}\right) = exp\left(\frac{\sigma^2}{2}\right) exp(\mu_{aY}) = exp\left(\frac{\sigma^2}{2}\right)\mu_{gX},$$

quindi, poichè $X = exp(Y)$, allora:

$$E(exp(Y)) = E(X) = exp\left(\frac{\sigma^2}{2}\right)\mu_{gX}.$$

Sapendo che $exp(\mu_{aY}) = \mu_{gX}$, si può sfruttare tale relazione per ricavare una buona misura di approssimazione della media della variabile iniziale X:

$$\mu_{aX} = exp\left(\frac{\sigma^2}{2}\right) exp(\mu_{aY}) = \gamma exp(\mu_{aY}) = \gamma\mu_{gX}.$$

In sintesi il parametro γ rappresenta un piccolo aggiustamento necessario per ottenere la perfetta equivalenza tra media aritmetica e media geometrica di X, ovviamente per distribuzioni che presentano distorsione.
Infatti poichè $\sigma^2 > 0$, $\exp\left(\frac{\sigma^2}{2}\right) > 1$: per valori non eccessivamente elevati della varianza questo fattore γ ha una piccola incidenza sulla media geometrica, al contrario

se aumenta la varianza esso tenderà ad aumentare di conseguenza.
Di seguito si analizzano quattro intervalli diversi, alcuni dei quali non dipendenti dai dati trasformati, per poter determinare quale tra questi è il migliore.
Il primo intervallo, denominato «normale», è il classico intervallo di confidenza per la media della popolazione, nel caso della variabile non trasformata.
Si assume che la statistica

$$T = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

ma, poichè σ è incognita, essa si approssima, per n piccolo, con una distribuzione t-Student con $(n-1)$ gradi di libertà, t_{n-1}.
Per ricavare l'intervallo basterà eguagliare la probabilità di copertura al livello di confidenza nominale, $(1-\alpha)$:

$$1-\alpha = P\left(t_{\alpha/2} < T < t_{1-\alpha/2}\right) = P\left(t_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < t_{1-\alpha/2}\right) =$$

$$P\left(\hat{\mu} - t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

in quanto $t_{\alpha/2} = -t_{1-\alpha/2}$. Allora l'intervallo per la media sarà il seguente:

$$[IC]_{1-\alpha} = \left[\hat{\mu}_{aX} - t_{1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \;\; ; \;\; \hat{\mu}_{aX} + t_{1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right].$$

Il secondo intervallo di interesse è denominato «Bonett», è l'intervallo «normale» che, però, considera un valore della deviazione standard differente, in quanto usa la stima proposta appunto da Bonett [2], $\hat{\sigma}_B$.
Per le distribuzioni non normali egli suggerisce di sfruttare la trasformazione logaritmica della varianza, per la sua proprietà di stabilizzazione, di cui si è parlato nella sezione 2.2. Inoltre se l'ipotesi di normalità è valida, si sostituisce il quantile $t_{1-\alpha/2}$ con il quantile della distribuzione normale standard, $z_{1-\alpha/2}$.
L'intervallo risultante sarà:

$$\left[\hat{\mu}_{aX} - z_{1-\alpha/2} \cdot \frac{\hat{\sigma}_B}{\sqrt{n}}; \hat{\mu}_{aX} + z_{1-\alpha/2} \cdot \frac{\hat{\sigma}_B}{\sqrt{n}}\right],\ \hat{\sigma}_B^2 = exp(log(c\hat{\sigma}^2)),\ c = \frac{n}{n - z_{1-\alpha/2}}.$$

Il terzo intervallo, denominato «Jhonson» si ottiene con una procedura leggermente più complessa, in quanto considera una correzione del quantile $t_{1-\alpha/2}$.
Come illustrato da Jhonson [8], tale correzione deriva dall'espansione di Cornish-Fisher con l'obiettivo di approssimare i quantili di una distribuzione non normale, in particolare asimmetrica con code molto pesanti, come nel caso di distribuzioni leptocurtiche.
L'espansione per una generica variabile X risulta:

$$CF(X) = \mu + \sigma\zeta + \frac{\mu_3}{\sigma^2}(\zeta^2 - 1) + \ldots,$$

dove ζ è una variabile proveniente da una distribuzione normale standard, μ è la media di X, σ^2 è la varianza di X e μ_3 è il momento terzo centrato di X, indice di simmetria della distribuzione.

Considerando allora il parametro di interesse $\hat{\mu}$ e assumendo che tutti i momenti della popolazione sono finiti, l'approssimazione a due termini di Cornish-Fisher per $\hat{\mu}$ è:

$$CF(\hat{\mu}) = \mu + \frac{\sigma}{\sqrt{n}}\zeta + \frac{\mu_3}{6n\sigma^2}(\zeta^2 - 1) + O(n-1).$$

Jhonson [8] parte da questa espansione del parametro per poter derivare l'approssimazione del quantile t, trascurando il termine $(\zeta^2 - 1)$, eliminando l'asimmetria, in quanto associato al coefficiente μ_3.
La forma di t modficata tramite l'espansione, è:

$$t_1 = \left[(\hat{\mu} - \mu) + \lambda + \gamma\left\{(\hat{\mu} - \mu)^2 - \left(\frac{\sigma^2}{n}\right)\right\}\right]\left(\frac{S^2}{n}\right)^{-\frac{1}{2}}.$$

I parametri sono scelti in modo tale che λ sia una funzione di n e γ in modo che ζ^2 sia pari a 0 per rendere nullo μ_3 e quindi per eliminare l'asimmetria.
Sostituendo il valore di λ e γ individuati, si ottiene il seguente t_1:

$$t_1 = \left[(\hat{\mu} - \mu) + \frac{\mu_3}{6\sigma^2 n} + \frac{\mu_3}{3\sigma^4}(\hat{\mu} - \mu)^2\right]\left(\frac{S^2}{n}\right)^{-\frac{1}{2}}.$$

Questa prima approssimazione di t risulta avere due problemi: l'elevata distorsione del numeratore e la forte correlazione tra numeratore e denominatore. Per poter costruire l'intervallo di confidenza infatti è necessario trovare un'espressione lineare rispetto a μ.
Quindi l'approssimazione finale del quantile t si ottiene trascurando il fattore $(\hat{\mu}-\mu)^2$, stimando i termini incogniti, μ_3 e σ, e sostituendo la varianza corretta, S^2, con la varianza della popolazione, σ^2, se si considera n abbastanza grande.
Allora le correzioni attuate per poter eliminare gli effetti di distorsione e asimmetria dovuti alla non normalità delle distribuzioni di interesse, permettono di individuare la seguente approssimazione di t:

$$t' = \left[(\hat{\mu} - \mu) + \frac{\mu_3}{6\sigma^2 n}\right]\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{2}}.$$

Di conseguenza l'intervallo di confidenza di livello $(1-\alpha)$ per μ_{aX}, media dei dati iniziali, è:

$$[IC]_{1-\alpha} = \left[\left(\hat{\mu}_{aX} + \frac{\hat{\mu}_3}{6\hat{\sigma}_X^2 n}\right) - t_{1-\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}} \;;\; \left(\hat{\mu}_{aX} + \frac{\hat{\mu}_3}{6\hat{\sigma}_X^2 n}\right) + t_{1-\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}\right].$$

L'ultimo è l'intervallo che, ai fini di tale analisi, è il più interessante. È stato proposto dal professor Curto [3] al fine di individuare dei nuovi intervalli, più accurati, per le distribuzioni asimmetriche e con code molto pesanti.
Data una variabile indipendente e identicamente distribuita, $X_i \sim logN(\mu, \sigma^2)$, ed $Y_i = log(X_i)$, con $Y_i \sim N(\mu, \sigma^2)$, media e varianza di Y_i saranno:

$$\mu_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(Y_i - \mu_Y)^2.$$

L'intervallo di confidenza di livello $(1-\alpha)$ per l'esponenziale della media aritmetica di Y, μ_{aY}, quindi per la media geometrica di X, μ_{gX}, sarà:

$$[IC]_{1-\alpha} = \left]exp\left(\hat{\mu}_{aY} - t_{1-\alpha/2} \cdot \frac{\hat{\sigma}_Y}{\sqrt{n}}\right); exp\left(\hat{\mu}_{aY} + t_{1-\alpha/2} \cdot \frac{\hat{\sigma}_Y}{\sqrt{n}}\right)\right[,$$

dove $\hat{\mu}_{aY}$ e $\hat{\sigma}^2_Y$ sono media e varianza di Y e $t_{\alpha/2} = t_{n-1,\alpha/2}$ è il quantile di una distribuzione t-Student con (n-1) gradi di libertà, t_{n-1}.
Per costruire l'intervallo finale bisogna tener conto di un piccolo fattore di aggiustamento per gli estremi.
Sapendo che maggiore è la numerosità campionaria, migliore sarà l'approssimazione di normalità di μ_{aY} per il teorema del limite centrale, allora la media μ_{gX} sarà approssimativamente distribuita come una lognormale per n molto grande, anche se la distribuzione di partenza di X non è lognormale, come nei casi di interesse.
L'intervallo finale sarà:

$$\left[\hat{\gamma}exp\left(\hat{\mu}_{aY} - t_{1-\alpha/2} \cdot \frac{\hat{\sigma}_Y}{\sqrt{n}}\right) \;\; ; \;\; \hat{\gamma}exp\left(\hat{\mu}_{aY} + t_{1-\alpha/2} \cdot \frac{\hat{\sigma}_Y}{\sqrt{n}}\right)\right].$$

Per determinare il valore di $\hat{\gamma}$ il professor Curto [3] ricorre alla procedura di regressione della media aritmetica suggerita da Woolridge [12].
Essenzialmente si costruisce un modello di regressione semplice senza intercetta dove si regredisce la media aritmetica, l'outcome, sulla media geometrica, regressore. Il coefficiente stimato della media geometrica sarà il valore associato a $\hat{\gamma}$. Quindi si costruiscono due vettori, uno per ognuna delle due medie, che verranno stimati per ciascuno dei campioni estratti, adottando il metodo del bootstrap.
Il professor Curto [3] si è posto il problema di derivare dei nuovi intervalli per le distribuzioni più critiche, cioè distribuzioni asimmetriche con code pesanti, come le distribuzioni leptocurtiche.
Come suggerisce Wilcox [11] il livello di significatività, per restituire delle stime ragionevoli, dovrebbe essere, in linea generale, tale: $0.025 < \alpha < 0.075$. Ma, nel caso di distribuzioni distorte e con code pesanti, Wilcox [11] propone valori di α addirittura superiori a 0.075, come avviene in questa analisi.
Curto [3] ha cercato di risolvere le problematiche relative a una scelta di un valore di α più elevato proponendo questo nuovo metodo che si applica solo se si conosce la relazione che intercorre tra media aritmetica, μ_a, e media geometrica, μ_g. In questo caso la relazione è $\mu_a = \gamma\mu_g$, ed è tale da permettere di individuare intervalli più accurati.
Il problema sorge con la stima di γ che dipende strettamente dalla distribuzione di partenza. Curto [3] suggerisce di stimare γ mediante il metodo bootstrap.
Wilcox [11] però ritiene che questo metodo potrebbe addirittura peggiorare le stime per le distribuzioni non normali, soprattutto all'aumentare della numerosità campionaria. Egli, infatti, sostiene che la proposta di Curto [3] sia interessante, ma la bontà di tale metodo dipende strettamente dalla stima di γ.
In conclusione secondo Wilcox [11] il metodo bootstrap non può garantire delle stime ottime per il parametro γ dato la stretta correlazione che vi è tra il parametro e i dati, rendendo necessario un metodo alternativo all'approccio di Curto [3].
In effetti dai risultati in Tabella 4.1 si nota che, per distribuzioni non noromali, in

particolare leptocurtiche, le stime delle probabilità di copertura restituiscono valori minori all'aumentare della dimensione del campione.
In ogni caso però le stime ottenute risultano migliori delle stime restituite dagli altri intervalli, dimostrando che, in qualche modo, l'intervallo proposto da Curto [3] può migliorare il metodo di stima di intervalli di confidenza per distribuzioni non normali che presentano delle code molto ampie, come nel caso della distribuzione χ^2_1.
Il metodo bootstrap[1] è stato applicato per ottenere le stime dell'intervallo $\gamma\hat{GM}_X$ in Tabella 4.1. In particolare è stato usato per derivare i vettori della media aritmetica e della media geometrica usati per costruire il modello di regressione attraverso cui è stato stimato il parametro γ. A livello pratico è stato costruito un doppio ciclo for, che, ad ogni ciclo esterno, ha generato dei nuovi campioni per costruire i vettori μ_{aX} e μ_{gX}, come si vede nel codice riportato in Appendice A.
Il confronto tra gli intervalli individuati si bassa ancora una volta sulla stima delle probabilità di copertura di ciascun intervallo proposto.
Si simulano 1000 campioni di Monte Carlo di differenti dimensioni, (10, 20, 30, 40, 50, 100), provenienti dalle seguenti distribuzioni teoriche: $Unif(0,1)$ e $Beta(3,3)$ per le distribuzioni platicurtiche, χ^2_1 ed $Exp(1)$ per le distribuzioni leptocurtiche. e $logN(0,1)$.
Il numero di campioni di Monte Carlo, 1000, non è eccessivamente elevato a causa del costo computazionale, sia in termini di tempo, sia in termini di efficienza delle stime dovuto al doppio ciclo for.

[1] I metodi di bootstrap, come suggerisce R. Jhonson [9], sono metodi ad alta intensità computazionale ma più flessibili rispetto ai metodi classici. In genere essi tendono a restituire dei risultati simili ai risultati ottenuti con i metodi classici di stima.
Dai risultati ottenuti da R.Jhonson [9] per l'applicazione della tecnica di bootrstrap non parametrica nella stima degli intervalli di confidenza, dimostra la validità di tale metodo anche nell'ambito di interesse di questo lavoro. In particolare il metodo permette di individuare i campioni bootstrap semplicemente mediante l'estrazione di campioni ripetuti, ognuno dei quali è ottenuto mediante la sostituzione dei dati.

Probabilità di copertura stimate per gli $IC_{1-\alpha}$ per $\hat{\mu}$					
DISTRIBUZIONE	n	NORMALE	BONETT	JHONSON	$\gamma G\hat{M}_X$
$logN(0,1)$	10	0.842	0.838	0.846	0.948
	20	0.860	0.858	0.866	0.950
	30	0.878	0.878	0.884	0.944
	40	0.880	0.876	0.882	0.934
	50	0.894	0.894	0.900	0.964
	100	0.918	0.918	0.926	0.960
$Unif(0,1)$	10	0.951	0.943	0.952	0.924
	20	0.947	0.946	0.950	0.933
	30	0.963	0.960	0.963	0.948
	40	0.950	0.949	0.950	0.953
	50	0.956	0.954	0.957	0.945
	100	0.953	0.952	0.953	0.944
$Beta(3,3)$	10	0.942	0.935	0.945	0.928
	20	0.948	0.944	0.949	0.941
	30	0.960	0.956	0.962	0.946
	40	0.941	0.940	0.942	0.938
	50	0.950	0.949	0.952	0.947
	100	0.948	0.946	0.948	0.941
χ^2_1	10	0.882	0.875	0.885	0.958
	20	0.870	0.866	0.876	0.940
	30	0.909	0.907	0.912	0.945
	40	0.935	0.934	0.939	0.955
	50	0.916	0.916	0.921	0.951
	100	0.935	0.933	0.939	0.948
$Exp(1)$	10	0.907	0.899	0.909	0.960
	20	0.913	0.910	0.914	0.951
	30	0.932	0.926	0.935	0.944
	40	0.939	0.937	0.946	0.951
	50	0.938	0.935	0.943	0.950
	100	0.937	0.936	0.940	0.950

Tabella 4.1 Sono stati simulati 1000 campioni di Monte Carlo di diverse dimensioni provenienti dalle seguenti distribuzioni teoriche: $logN(0,1)$, $Unif(0,1)$, $Beta(3,3)$, χ^2_1 ed $Exp(1)$. Per ciascuna di esse sono state stimate le probabilità di copertura degli intevalli proposti per $\hat{\mu}$.

Dai risultati in Tabella 4.1 risulta chiaro che per le distribuzioni leptocurtiche, χ^2_1 ed $Exp(1)$, il nuovo intervallo proposto da Curto [3] tende a restituire valori migliori degli altri intervalli, risultando così più conservativo.
In realtà però si nota una tendenza che potrebbe sembrare, a primo impatto, illogica. Infatti, come notato in precedenza, le stime ottenute risultano più elevate per una dimensione del campione più piccola e tendono a diminuire all'aumentare della numerosità campionaria n.
Al contrario, gli intervalli «normale», «Bonett» e «Jhonson», risultano essere più liberali in quanto restituiscono valori effettivi minori del livello di confiden-

za, $(1-\alpha) = 0.95$.
Lo stesso ragionamento vale anche per la distribuzione log-normale, non propriamente leptocurtica ma con un'ampia coda destra, in quanto l'intervallo basato sulla trasformazione dei dati, $\gamma \hat{GM}_X$, restituisce i valori migliori tra gli intervalli proposti.
Per le distribuzioni platicurtiche, $Unif(0,1)$ e $Beta(3,3)$, il risultato ottenuto è l'opposto: sembra infatti che tale intervallo non comporti alcun miglioramento alle stime, in alcuni casi tende addirittura a risultare il meno conservativo.
Per quanto riguarda gli intervalli che non si basano sui valori trasformati, quindi «normale», «Bonett» e «Jhonson», i risultati che restituiscono sono abbastanza simili. Rimane da preferire «Jhonson» che, adottando un aggiustamento per il quantile t che corregge l'asimmetria della distribuzione, permette, quasi sempre, nei casi in esame, di ottenere stime leggermente più elevate
Quindi si può dire che il nuovo intervallo proposto da Curto [3] risulta uno strumento di stima intervallare efficace nel caso di distribuzioni leptocurtiche, in quanto restituisce, per ogni valore di n, delle ottime stime della probabilità di copertura. Allo stesso tempo però il suo comportamento risulta leggermente anomalo a causa delle osservazioni sottolineate in precedenza, suggerendo di adottare una certa cautela nel suo impiego.
Per la distribuzione lognormale tale intervallo risulta essere il migliore: restituisce stime della probabilità di copertura più elevate e, all'aumentare della dimensione dei campioni, tende ad aumentare, così come era auspicabile.
Per le distribuzioni platicurtiche invece si suggerisce l'impiego di uno degli altri tre intervalli proposti in quanto i valori resituiti dalle stime risultano più conservativi rispetto all'intervallo $\gamma \hat{GM}_X$.

Capitolo 5

Intervalli di confidenza per la differenza di medie

L'ultimo obiettivo è individuare l'intervallo di confidenza migliore per la differenza di medie aritmetiche di due popolazioni aventi stessa distribuzione.
Anche in questo caso si confrontano quattro intervalli, alcuni dei quali sono costruiti sui dati originali. Il quarto intervallo, $\gamma\hat{GM}_X$, suggerito da Curto [3], invece, è basato, come nella sezione 4, sulla trasformazione logaritmica dei dati originali. Ovviamente questi intervalli dovranno essere adattati al caso di due popolazioni a confronto.
Siano date due variabili, $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$.
L'intervallo «normale» si ricava semplicemente adattando la formula della sezione precedente al nuovo parametro, $(\hat{\mu}_1 - \hat{\mu}_2)$ e risulterà il seguente:

$$[IC]_{1-\alpha} = \left[(\hat{\mu}_1 - \hat{\mu}_2) - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \; ; \; (\hat{\mu}_1 - \hat{\mu}_2) + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \right],$$

dove $z_{1-\alpha/2}$ è il quantile della distribuzione normale standardizzata.
Lo stesso ragionamento si può estendere all'intervallo di Bonett in quanto la differenza consiste semplicemente nel considerare lo standard error corretto da Bonett [2], $\hat{\sigma}_B^2$; l'intervallo sarà:

$$[IC]_{1-\alpha} = \left[(\hat{\mu}_1 - \hat{\mu}_2) - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_{B1}^2}{n_1} + \frac{\hat{\sigma}_{B2}^2}{n_2}}; (\hat{\mu}_1 - \hat{\mu}_2) + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_{B1}^2}{n_1} + \frac{\hat{\sigma}_{B2}^2}{n_2}} \right],$$

ricordando che $\hat{\sigma}_{Bi}^2$ contiene un fattore di aggiustamento, c_i, tale da apportare correzioni in merito alla numerosità campionaria, e sarà il seguente:

$$\hat{\sigma}_{Bi}^2 = exp\left[\ln(c_i\hat{\sigma}_i^2)\right], \qquad c_i = \frac{n_i}{n_i - z_{1-\alpha/2}}.$$

L'intervallo di Jhonson [8], come nel caso univariato, ha l'obiettivo di migliorare l'accuratezza del quantile t. Anche in questo caso sarà necessario apportare delle modifiche all'intervallo di partenza in modo tale da adattarlo al caso bivariato.
L'intervallo considera l'approssimazione di t ottenuta applicando la procedura jack-knife, suggerita da Jhonson [8], una generalizzazione del metodo applicato nella

sezione 4.
In sintesi la procedura consiste nel creare dei sottoinsime del campione originale, ottenuti rimuovendo, in ogni sottoinsieme, una delle osservazioni del campione iniziale. A questo punto per ciascun sottoinsieme si stima il parametro di interesse. La stima finale, detta «stima jackknife», consiste nel calcolare una media dei valori ottenuti dalle stime precedenti, con l'obiettivo di ridurre la distorsione dei campioni.
Jhonson [8] però afferma che, in molti casi, i valori ottenuti mediante le stime possono essere considerati delle osservazioni indipendenti e quindi si può sfruttare la distribuzione t-Student per determinare gli intervalli di confidenza per il parametro incognito.
Egli inoltre afferma che si può dimostrare che per variabili come la media, questa procedura elimina la distorsione per almeno una espansione di Cornish-Fisher, obiettivo che, appunto, questa sezione vuole raggiungere.
Applicando i risultati di Jhonson [8], il professor Curto [3] propone l'intervallo basato sulla correzione del quantile t avente i seguenti estremi:

$$LB = \left(\hat{\mu}_1 + \frac{\hat{\mu}_{3,1}}{6\hat{\sigma}_1^2 n_1}\right) - \left(\hat{\mu}_2 + \frac{\hat{\mu}_{3,2}}{6\hat{\sigma}_2^2 n_1}\right) - t_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

$$UB = \left(\hat{\mu}_1 + \frac{\hat{\mu}_{3,1}}{6\hat{\sigma}_1^2 n_1}\right) - \left(\hat{\mu}_2 + \frac{\hat{\mu}_{3,2}}{6\hat{\sigma}_2^2 n_1}\right) + t_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

dove LB è l'estremo inferiore, mentre UB è l'estremo superiore.
Inoltrer $t_{1-\alpha/2} = t_{n_1+n_2-2,1-\alpha/2}$ è il quantile della distribuzione t-Student con $(n_1 + n_2 - 2)$ gradi di libertà.
Infatti la differenza delle due variabili inziali, aventi distribuzioni t-Student, rispettivamente con $(n_1 - 1)$ ed $(n_2 - 1)$ gradi di libertà, la differenza dei due parametri avrà ancora distribuzione t-Student ma con $(n_1 + n_2 - 2)$ gradi di libertà.
L'intervallo $\gamma G\hat{M}_X$ è un po' più complicato da ricavare. Per la sua costruzione si estendono i risultati della sezione 4 al caso bivariato.
Ricordando l'equivalenza: $\mu_{gX} = exp(\mu_{aY})$, il rapporto tra le medie di due popolazioni aventi stessa distribuzione sarà:

$$\frac{\mu_{X_1}}{\mu_{X_2}} = \frac{\gamma_1 exp(\mu_{Y_1})}{\gamma_2 exp(\mu_{Y_2}} = \frac{\gamma_1}{\gamma_2} exp(\mu_{Y_1} - \mu_{Y_2}).$$

In questo caso il confronto tra le distribuzioni avviene mediante il rapporto delle medie geometriche dei dati originali, quindi, per la definizione data a μ_{gX} precedentemente, mediante l'esponenziale della differenza delle medie dei dati trasformati.
Allora assumendo la normalità della distribuzione Y_i, quindi la lognormalità di X_i, si costruisce l'intervallo finale.
Il procedimento è abbastanza semplice perchè basta considerare l'esponenziale del classico intervallo di confidenza per la differenza di medie, che, in questo caso, sono relative ai dati trasformati, cioè alle Y_i. L'intervallo è:

$$\left]\frac{\hat{\gamma}_1}{\hat{\gamma}_2} exp\left(\hat{\mu}_{Y_1} - \hat{\mu_{Y_2}} - t\sqrt{\frac{\hat{\sigma}_{Y_1}^2}{n_1} + \frac{\hat{\sigma}_{Y_2}^2}{n_2}}\right); \frac{\hat{\gamma}_1}{\hat{\gamma}_2} exp\left(\hat{\mu}_{Y_1} - \hat{\mu_{Y_2}} + t\sqrt{\frac{\hat{\sigma}_{Y_1}^2}{n_1} + \frac{\hat{\sigma}_{Y_2}^2}{n_2}}\right)\right[,$$

dove $t = t_{n_1+n_2-2,1-\alpha/2}$ è il quantile della distribuzione t-Student con $(n_1 + n_2 - 2)$ gradi di libertà.
Il quantile potrebbe essere sostituito dal quantile della distribuzione normale standard, quindi $z_{1-\alpha/2}$ se l'ipotesi di approssimazione normale fosse valida.
Come nel caso univariato, anche nel caso bivariato sarà necessario tener conto di un fattore di aggiustamento rappresentato da γ.
Poichè l'intervallo è costruito sul parametro che corrisponde al rapporto di medie aritmetiche, si considera come fattore di correzione il rapporto tra i singoli fattori, quindi $\frac{\hat{\gamma}_1}{\hat{\gamma}_2}$.
Dal punto di vista pratico bisogna considerare due modelli di regressione semplice, uno per ognuno dei due campioni, quindi ogni γ_i sarà relativo a un coefficiente di regressione differente.
Il confronto, come nelle analisi precedenti, si basa sulla stima delle probabilità di copertura di ciascun intervallo considerato.
Si simulano 1000 campioni di Monte Carlo di dimensioni diverse provenienti dalle seguenti distribuzioni teoriche: $Unif(0, 1)$ e $Beta(3, 3)$ per le distribuzioni platicurtiche, e χ^2_1 ed $Exp(1)$ per le distribuzioni leptocurtiche e $logN(0, 1)$.
Poichè si è interessati al confronto di una serie distribuzioni, si possono scegliere numerosità differenti per ciascun campione. Le coppie di numerosità per i due campioni sono state scelte in modo da non essere troppo differenti, sono le seguenti: $(10, 15), (20, 25), (30, 35), (40, 45), (50, 55), (100, 110)$.
Anche in questo caso il numero di campioni di Monte Carlo, 1000, non è eccessivamente elevato per rendere più efficiente la stima, soprattutto in termini di tempo, a causa del doppio ciclo for.

Probabilità di copertura stimate per gli $IC_{1-\alpha}$ per $(\hat{\mu}_1 - \hat{\mu}_2)$					
DISTRIBUZIONE	(n_1, n_2)	NORMALE	BONETT	JHONSON	$\gamma G\hat{M}_X$
$logN(0,1)$	(10, 15)	0.961	0.970	0.970	0.937
	(20, 25)	0.949	0.960	0.950	0.918
	(30, 35)	0.960	0.966	0.960	0.908
	(40, 45)	0.961	0.965	0.960	0.896
	(50, 55)	0.952	0.955	0.953	0.896
	(100, 110)	0.953	0.955	0.950	0.875
$Unif(0,1)$	(10, 15)	0.933	0.950	0.947	0.899
	(20, 25)	0.942	0.951	0.951	0.894
	(30, 35)	0.957	0.968	0.966	0.907
	(40, 45)	0.941	0.947	0.947	0.876
	(50, 55)	0.945	0.950	0.949	0.877
	(100, 110)	0.955	0.955	0.955	0.886
$Beta(3,3)$	(10, 15)	0.930	0.949	0.943	0.873
	(20, 25)	0.943	0.955	0.950	0.867
	(30, 35)	0.949	0.954	0.952	0.863
	(40, 45)	0.945	0.946	0.946	0.877
	(50, 55)	0.946	0.949	0.948	0.857
	(100, 110)	0.944	0.947	0.946	0.836
χ^2_1	(10, 15)	0.946	0.962	0.954	0.970
	(20, 25)	0.948	0.962	0.957	0.950
	(30, 35)	0.949	0.956	0.950	0.952
	(40, 45)	0.946	0.949	0.947	0.935
	(50, 55)	0.953	0.595	0.954	0.936
	(100, 110)	0.951	0.953	0.950	0.906
$Exp(1)$	(10, 15)	0.951	0.963	0.960	0.940
	(20, 25)	0.940	0.946	0.942	0.933
	(30, 35)	0.948	0.955	0.950	0.934
	(40, 45)	0.960	0.962	0.962	0.917
	(50, 55)	0.942	0.946	0.945	0.906
	(100, 110)	0.949	0.952	0.951	0.893

Tabella 5.1 Sono stati simulati 1000 campioni di Monte Carlo di diverse dimensioni provenienti dalle seguenti distribuzioni teoriche: $logN(0,1)$, $Unif(0,1)$, $Beta(3,3)$, χ^2_1 ed $Exp(1)$. Per ciascuna di esse sono state stimate le probabilità di copertura degli intervalli proposti per $(\hat{\mu}_1 - \hat{\mu}_2)$.

Dai risultati in Tabella 5.1, è evidente che, in linea generale, gli intervalli individuati si equivalgano, in quanto risultano piuttosto conservativi. In particolare il valore effettivo della probabilità di copertura tende a coincidere, anche approssimativamente, con il valore nominale scelto, $(1 - \alpha) = 0.95$, nella maggior parte dei casi. L'intervallo più critico è sicuramente quello proposto dal professor Curto [3], in quanto i valori restituiti tendono a diminiuire con l'aumentare della numerosità campionaria, rendendo così le stime meno accurate.
I valori ottenuti risultano più coerenti con le osservazioni di Wilcox [11] in merito all'accuretezza delle stime del nuovo intervallo, nonostante egli abbia evidenziato

le criticità di cui si è parlato in precedenza nel caso univariato. Tali osservazioni risultano però più evidenti nel caso bivariato per cui, stando ai risultati ottenuti, non si può definire ottimo in senso assoluto il nuovo intervallo. Infatti basta notare che le stime delle probabilità di copertura dell'intervallo $\gamma G\hat{M}_X$ tendono a diminuire all'aumentare della dimensione del campione, proprio come aveva evidenziato Wilcox [11].
In questo caso, al contrario dei casi analizzati nelle sezioni precedenti, non si riscontrano grandi differenza tra i tipi di distribuzioni considerati. La distinzione dovuta alla forma della distribuzione sembra non essere influente ai fini dell'analisi effettuata.
Quindi si può dire che, per intervalli con numerosità campionarie non particolarmente elevate, i quattro intervalli si equivalgono e possono essere usati indistintamente.
Al contrario, per numerosità campionarie più elevate, è preferibile adoperare intervalli di confidenza non basati su alcuna trasformazione. Gli intervalli migliori saranno «normale», «Bonett» e «Jhonson».
In virtù di quanto detto nell'Introduzione, le caratteristiche fondamentali degli intervalli di confidenza sono la probabilità di copertura e l'ampiezza degli intervalli. Quindi sapendo che se una migliora l'altra peggiora, per approfondire il confronto, si può studiare anche l'ampiezza degli intervalli proposti.
Per questo motivo, poichè i risultati delle stime delle probabilità di copertura in Tabella 5.1 non sono quelli sperati, il professor Curto [3] suggerisce di confrontare gli intervalli dal punto di vista della loro ampiezza. Egli cerca di individuare gli intervalli con ampiezza minore, in modo da garantire una maggiore efficienza.
Dai risultati in Tabella 5.2, risulta necessario ragionare in modo differente per le distribuzioni analizzate.
Per le distribuzioni platicurtiche l'ampiezza degli intervalli risulta, nella maggior parte dei casi, maggiore per l'intervallo di Curto [3]. Quindi, anche in termini di ampiezza, l'intervallo $\gamma G\hat{M}_X$ è il peggiore tra gli intervalli proposti, non garantendo alcun miglioramento nell'efficienza delle stime.
Per le distribuzioni leptocurtiche invece, all'aumentare della numerosità campionaria, le stime dell'ampiezza degli intervalli tendono a migliorare, in alcuni casi risultano addirittura più piccole degli altri intervalli.
Questo lieve miglioramento in realtà non necessariamente giustifica una scelta decisa su questo intervallo. Infatti l'ampiezza risulta di poco più piccola rispetto agli altri intervalli. Al contrario, dalle stime in Tabella 5.1, le differenze tra le probabilità di copertura sono più significative. La scelta deve essere effettuata considerando un giusto compromesso tra valore effettivo della probabilità di copertura e ampiezza dell'intervallo.
Quindi non sempre usare l'intervallo $\gamma G\hat{M}_X$ può risultare la scelta migliore nella stima degli intervalli di confidenza per il parametro $(\hat{\mu}_1 - \hat{\mu}_2)$, concludendo che è possibile ammettere intervalli alternativi per rendere le stime più accurate.

Ampiezza degli $IC_{1-\alpha}$ per $\hat{\mu}_1 - \hat{\mu}_2$					
DISTRIBUZIONE	(n_1, n_2)	NORMALE	BONETT	JHONSON	$\gamma \hat{GM}_X$
$logN(0,1)$	(10, 15)	3.292	3.635	3.471	1.990
	(20, 25)	2.732	2.847	2.811	0.515
	(30, 35)	1.379	1.423	1.407	0.833
	(40, 45)	1.487	1.523	1.509	0.663
	(50, 55)	1.462	1.490	1.479	0.670
	(100, 110)	0.883	0.891	0.888	0.476
$Unif(0,1)$	(10, 15)	0.457	0.501	0.482	1.250
	(20, 25)	0.325	0.340	0.334	1.450
	(30, 35)	0.282	0.291	0.287	1.420
	(40, 45)	0.242	0.248	0.245	0.656
	(50, 55)	0.234	0.238	0.237	0.673
	(100, 110)	0.157	0.158	0.158	0.421
$Beta(3,3)$	(10, 15)	0.312	0.343	0.329	0.584
	(20, 25)	0.246	0.258	0.254	0.595
	(30, 35)	0.187	0.193	0.150	0.319
	(40, 45)	0.156	0.161	0.159	0.293
	(50, 55)	0.167	0.170	0.169	0.275
	(100, 110)	0.101	0.102	0.101	0.183
χ^2_1	(10, 15)	1.241	1.377	1.310	8.920
	(20, 25)	2.491	2.621	2.563	3.560
	(30, 35)	1.558	1.608	1.589	1.781
	(40, 45)	1.010	1.035	1.025	1.615
	(50, 55)	1.060	1.080	1.073	1.311
	(100, 110)	0.775	0.782	0.779	0.785
$Exp(1)$	(10, 15)	1.517	1.676	1.601	2.440
	(20, 25)	1.156	1.247	1.220	1.246
	(30, 35)	0.899	0.927	0.916	0.942
	(40, 45)	0.785	0.804	0.796	0.728
	(50, 55)	0.701	0.714	0.709	0.688
	(100, 110)	0.627	0.634	0.631	0.566

Tabella 5.2 Sono stati simulati 10000 campioni di Monte Carlo di diverse dimensioni provenienti dalle seguenti distribuzioni teoriche: $logN(0,1)$, $Unif(0,1)$, $Beta(3,3)$, χ^2_1 ed $Exp(1)$. Sono state stimate le ampiezze degli intevalli di confidenza per il parametro $(\hat{\mu}_1 - \hat{\mu}_2)$.

Capitolo 6

Conclusioni

La presente tesi aveva l'obiettivo di studiare i risultati del professor Curto [3] su una serie di intervalli di confidenza proposti per i quattro parametri: varianza $\hat{\sigma}^2$, rapporto di varianze $\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$, media $\hat{\mu}$ e differenza di medie $(\hat{\mu}_1 - \hat{\mu}_2)$.
Nella prima parte, sulla varianza e sul rapporto di varianze, sono stati simulati 10000 campioni di Monte Carlo di dimensioni diverse provenienti da una serie di distribuzioni teoriche. Per la seconda parte, sulla media e sulla differenza di medie, sono stati simulati, con la stessa logica, 1000 campioni di Monte Carlo.
Il numero di campioni simulati, in entrambi i casi, è stato scelto in modo da rendere più veloce l'esecuzione dei programmi in R dal punto di vista computazionale, nonostante, soprattutto nel secondo caso, tale numero risulti relativamente basso per le stime.
Come evidenziato dal professor Curto [3], le distribuzioni più interessanti sono state le distribuzioni leptocurtiche, χ_1^2, t_5 ed $Exp(1)$, in quanto presentano più criticità nelle stime.
Per prima cosa è stato individuato un nuovo stimatore della curtosi, alternativo allo stimatore di Pearson, in grado di restituire stime più precise, in quanto meno distorte, per l'indice di curtosi. L'indice in questione è costruito sulla mediana, più robusta rispetto alla media. In quanto tale esso risente meno dell'influenza di valori anomali, particolarmente presenti nelle distribuzioni con code più ampie.
Grazie a questo primo risultato è stato possibile costruire dei nuovi intervalli di confidenza per la varianza della popolazione. È stato necessario ricorrere alla trasformazione logaritmica della varianza per poter ridurre l'asimmetria della distribuzione di partenza.
L'intervallo classico è stato confrontato con i nuovi intervalli costruiti mediante i tre nuovi stimatori della curtosi.
Anche in questo caso, coerentemente con i risultati ottenuti in precedenza, lo stimatore migliore per le distribuioni leptocurtiche è risultato l'intervallo $\hat{\gamma}_4(3)$, appunto basato sulla mediana.
In seguito si è cercato di estendere i risultati ottenuti nel caso univariato al caso bivariato, basato sul confronto tra varianze di popolazioni provenienti dalla stessa distribuzione. Sono stati individuati quattro intervalli da confrontare con il classico intervallo teorico.
Si giunge alle medesime conslusioni: i nuovi intervalli proposti dal professor Curto

[3] risultano i migliori, restituendo stime di probabilità di copertura degli intervalli più consevrative.
Nella seconda parte il parametro di interesse è stata la media aritmetica e il ragionamento per derivare gli intervalli è stato leggermente diverso.
L'intervallo ricavato con metodo classico è stato confrontato con altri tre intervalli, uno dei quali, proposto dal professor Curto [3], è basato sulla trafsormazione logaritmica dei dati originali per ridurre distorisone e asimmetria.
In questo caso è stato più difficile determinare un intervallo migliore in senso assoluto. In linea generale però l'intervallo proposto dal professor Curto [3] ha restituito stime conservative.
L'ultimo obiettivo era individuare l'intervallo di confidenza migliore per la differenza di medie. Anche in questa sezione sono stati individuati quattro intervalli, gli stessi del caso univariato.
I risultati ottenuti però non hanno dimostrato un netto vantaggio nell'impiego del nuovo intervallo. Per poter verificare l'efficienza dell'intervallo basato sulla trsformazione logaritmica, sono state stimate le ampiezze degli intervalli individuati, ma, ancora una volta i risultati non sono stati quelli sperati.
Quindi non è stato possibile stabilire che un intervallo sia migliore degli altri in senso assoluto: tutti i quattro intervalli proposti sono dei buoni strumenti di stima intervallare.
In conclusione questo studio, sulla base delle simulazioni eseguite, permette di definire gli intervalli proposti dal professor Curto [3] dei buoni strumenti di stima, ottimi nel caso di distribuzioni leptocurtiche e per i parametri dipendenti dalla varianza.

Appendice A

Codici R

Per ciascuna tabella è riportato il codice per derivare i risultati nel caso di una sola distribuzione con una numerosità campionaria pari a 10.
Per ottenere tutte le restanti stime sarà sufficiente sostituire i seguenti valori:

- n: 10, 20, 30, 40, 50, 100.
- camp, camp1, camp2: rnorm(n, 0, 1), runif(n, 0, 1), rbeta(n, 3, 3), rchisq(n, 1), rt(n, 5), rexp(n, 1).

A.1 Codice per le stime in Tabella 2.1

Il valore della curtosi cambia per ciascuna distribuzione; i valori saranno:

- 3 per N(0,1), 1.8 per Unif(0,1), 2.333 per Beta(3,3), 15 per Chi-quadro(1), 9 per t(5), 9 per Exp(1).

```
# N(0,1)
n = 10
M = 10000
sim1 = matrix(NA, M)
sim2 = matrix(NA, M)
sim3 = matrix(NA, M)
for(i in 1:M){
  camp = rnorm(n, 0, 1)
  perc = 1/(2*(n-4)^(1/2))
  gamma41 = n*sum((camp-mean(camp))^4)/(sum((camp-mean(camp))^2))^2
  gamma42 = n*sum((camp-mean(camp,perc))^4)/
  (sum((camp-mean(camp))^2))^2
  gamma43=n*sum((camp-median(camp))^4)/(sum((camp-mean(camp))^2))^2
  sim1[i] = gamma41 - curtosi
  sim2[i] = gamma42 - curtosi
  sim3[i] = gamma43 - curtosi}
out1 = mean(sim1)
out2 = mean(sim2)
out3 = mean(sim3)
```

A.2 Codice per le stime in Tabella 2.2

Il valore vero della varianza cambia per ciascuna distribuzione; i valori saranno:

- var: 1 per N(0,1), 0.083 per Unif(0,1), 0.036 per Beta(3,3), 2 per Chi-quadro(1), 1.667 per t(5), 1 per Exp(1).

```
# N(0,1)
n = 10
M = 10000
var = 1
sm0 = matrix(NA, M)
sm1 = matrix(NA, M)
sm2 = matrix(NA, M)
sm3 = matrix(NA, M)
for(i in 1:M){
  camp = rnorm(n, 0, 1)
  var = 1
  perc = 1/(2*(n-4)^(1/2))
  gamma41 = n*sum((camp-mean(camp))^4)/(sum((camp-mean(camp))^2))^2
  gamma42 = n*sum((camp-mean(camp,perc/2))^4)/
  (sum((camp-mean(camp))^2))^2
  gamma43 = n*sum((camp-median(camp))^4)/(sum((camp-mean(camp))^2))^2
  alpha = 0.05
  z = qnorm(1-alpha/2, 0, 1)
  se11 = ((gamma41-(n-3)/n)/(n-1))^(1/2)
  se12 = ((gamma42-(n-3)/n)/(n-1))^(1/2)
  se13 = ((gamma43-(n-3)/n)/(n-1))^(1/2)
  LL0 = (n-1)*var(camp)/qchisq((1-alpha/2), n-1)
  UL0 = (n-1)*var(camp)/qchisq((alpha/2), n-1)
  if(LL0 < var & UL0 > var) { sm0[i] = 1 } else {sm0[i] = 0}
  LL1 = exp(log(var(camp)) - z * se11)
  UL1 = exp(log(var(camp)) + z * se11)
  if(LL1 < var & UL1 > var) {sm1[i] = 1 } else{sm1[i] = 0}
  LL2 = exp(log(var(camp)) - z * se12)
  UL2 = exp(log(var(camp)) + z * se12)
  if(LL2 < var & UL2 > var) {sm2[i] = 1} else {sm2[i] = 0}
  LL3 = exp(log(var(camp)) - z * se13)
  UL3 = exp(log(var(camp)) + z * se13)
  if(LL3 < var & UL3 > var) {sm3[i] = 1} else {sm3[i] = 0}}
out0 = mean(sm0)
out1 = mean(sm1)
out2 = mean(sm2)
out3 = mean(sm3)
```

A.3 Codice per le stime in Tabella 2.3

Il valore vero della varianza cambia per ciascuna distribuzione; i valori saranno:

- var: 1 per N(0,1), 0.083 per Unif(0,1), 0.036 per Beta(3,3), 2 per Chi-quadro(1), 1.667 per t(5), 1 per Exp(1).

```
# N(0,1)
n = 10
M = 10000
var = 1
sm1 = matrix(NA, M)
sm2 = matrix(NA, M)
sm3 = matrix(NA, M)
for(i in 1:M){
  camp = rnorm(n, 0, 1)
  var = 1
  perc = 1/(2*(n-4)^(1/2))
  gamma41 = n*sum((camp-mean(camp))^4)/(sum((camp-mean(camp))^2))^2
  gamma42 = n*sum((camp-mean(camp,perc/2))^4)/
  (sum((camp-mean(camp))^2))^2
  gamma43 = n*sum((camp-median(camp))^4)/(sum((camp-mean(camp))^2))^2
  alpha = 0.05
  z = qnorm(1-alpha/2, 0, 1)
  c = n/(n-z)
  se21 = c*((gamma41-(n-3)/n)/(n-1))^(1/2)
  se22 = c*((gamma42-(n-3)/n)/(n-1))^(1/2)
  se23 = c*((gamma43-(n-3)/n)/(n-1))^(1/2)
  LL1 = exp(log(var(camp)) - z * se21)
  UL1 = exp(log(var(camp)) + z * se21)
  if(LL1 < var & UL1 > var) {sm1[i] = 1 } else{sm1[i] = 0}
  LL2 = exp(log(var(camp)) - z * se22)
  UL2 = exp(log(var(camp)) + z * se22)
  if(LL2 < var & UL2 > var) {sm2[i] = 1} else {sm2[i] = 0}
  LL3 = exp(log(var(camp)) - z * se23)
  UL3 = exp(log(var(camp)) + z * se23)
  if(LL3 < var & UL3 > var) {sm3[i] = 1} else {sm3[i] = 0}}
out1 = mean(sm1)
out2 = mean(sm2)
out3 = mean(sm3)
```

A.4 Codice per le stime in Tabella 2.4

Il valore vero della varianza cambia per ciascuna distribuzione; i valori saranno:

- var: 1 per N(0, 1), 1 per Unif(0, 1), 1 per Beta(3, 3), 2 per Chi-quadro(1), 1 per t(5), 1 per Exp(1).

```
# N(0,1)
n = 10
M = 10000
```

```
var = 1
smf1 = matrix(NA, M)
smn = matrix(NA, M)
sm1 = matrix(NA, M)
sm2 = matrix(NA, M)
sm3 = matrix(NA, M)
for(i in 1:M){
  c1 = rnorm(n, 0, 1)
  c2 = rnorm(n, 0, 1)
  perc = 1/(2*(n-4)^(1/2))
  gamma41_1 = n*sum((camp1-mean(camp1))^4)/
  (sum((camp1-mean(camp1))^2))^2
  gamma41_2 = n*sum((camp2-mean(camp2))^4)/
  (sum((camp2-mean(camp2))^2))^2
  gamma42_1 = n*sum((camp1-mean(camp1,perc/2))^4)/
  (sum((camp1-mean(camp1))^2))^2
  gamma42_2 = n*sum((camp2-mean(camp2,perc/2))^4)/
  (sum((camp2-mean(camp2))^2))^2
  gamma43_1 = n*sum((camp1-median(camp1))^4)/
  (sum((camp1-mean(camp1))^2))^2
  gamma43_2 = n*sum((camp2-median(camp2))^4)/
  (sum((camp2-mean(camp2))^2))^2
  alpha = 0.05
  z = qnorm(1-alpha/2, 0, 1)
  c = n/(n-z)
  se11_1 = c * ((gamma41_1 - (n-3)/n)/(n-1))^(1/2)
  se11_2 = c * ((gamma41_2 - (n-3)/n)/(n-1))^(1/2)
  se12_1 = c * ((gamma42_1 - (n-3)/n)/(n-1))^(1/2)
  se12_2 = c * ((gamma42_2 - (n-3)/n)/(n-1))^(1/2)
  se13_1 = c * ((gamma43_1 - (n-3)/n)/(n-1))^(1/2)
  se13_2 = c * ((gamma43_2 - (n-3)/n)/(n-1))^(1/2)
  LLn = (var(camp2)/var(camp1)) * qf((alpha/2), n-1, n-1)
  ULn = (var(camp2)/var(camp1)) * qf((1-alpha/2), n-1, n-1)
  if(LLn < var & ULn > var) {smn[i] = 1} else{smn[i] = 0}
  f11 = sum((c1 - mean(camp1))^4)
  f12 = sum((c2 - mean(camp2))^4)
  mu4f1 = (f11 + f12)/(n+n)
  s2f1 = ((n-1) * var(camp1) + (n-1) * var(camp2))/(n+n)
  gamma4f1 = mu4f1/(s2f1^2)
  r1 = (2*n) / (gamma4f1 - ((n-3)/(n-1)))
  r2 = (2*n) / (gamma4f1 - ((n-3)/(n-1)))
  LLf1 = (var(camp2)/var(camp1)) * qf((alpha/2), r1, r2)
  ULf1 = (var(camp2)/var(camp1)) * qf((1-alpha/2), r1, r2)
  if(LLf1 < var & ULf1 > var) {smf1[i] = 1} else {smf1[i] = 0}
  LL1 = exp(log(c*var(camp2))-log(c*var(camp1))-z*(se11_1+se11_2))
  UL1 = exp(log(c*var(camp2))-log(c*var(camp1))+z*(se11_1+se11_2))
  if(LL1 < var & UL1 > var) {sm1[i] = 1 } else {sm1[i] = 0}
```

```
  LL2 = exp(log(c*var(camp2))-log(c*var(camp1))-z*(se12_1+se12_2))
  UL2 = exp(log(c*var(camp2))-log(c*var(camp1))+z*(se12_1+se12_2))
  if(LL2 < var & UL2 > var) {sm2[i] = 1 } else {sm2[i] = 0}
  LL3 = exp(log(c*var(camp2))-log(c*var(camp1))-z*(se13_1+se13_2))
  UL3 = exp(log(c*var(camp2))-log(c*var(camp1))+z*(se13_1+se13_2))
  if(LL3 < var & UL3 > var) {sm3[i] = 1 } else {sm3[i] = 0}}
outF1 = mean(smf1)
outN = mean(smn)
out1 = mean(sm1)
out2 = mean(sm2)
out3 = mean(sm3)
```

A.5 Codice per le stime in Tabella 3.1

Il valore vero della media dipende da ciascuna distribuzione; i valori saranno:

- mu: 1.649 per logN(0, 1), 0.5 per Unif(0, 1), 0.5 per Beta(3, 3), 1 per Chi-quadro(1), 1 per Exp(1).

```
# LOGN(0,1)
n = 10
M = 1000
mu = exp(1/2)
mug_x = function(x){
  out = exp(1/n * sum(log(x)))
  return(out)}
sim1 = matrix(NA, M)
sim2 = matrix(NA, M)
sim3 = matrix(NA, M)
sim4 = matrix(NA, M)
mux = matrix(NA, M)
mug = matrix(NA, M)
for(i in 1:M){
alpha = 0.05
  z = qnorm(1-alpha/2, 0, 1)
  c = n/(n-z)
  t = qt(1-alpha/2, n-1)
  for(j in 1:M){
    camp = rlnorm(n, 0, 1)
    mux[j] = mean(camp)
    mug[j] = mug_x(camp)}
  mod1 = lm(mux ~ 0 + mug)
  gamma = coef(mod1)
  LL1 = mean(camp) - t * sd(camp)/(sqrt(n))
  UL1 = mean(camp) + t * sd(camp)/(sqrt(n))
  if(LL1 < mu & UL1 > mu) {sim1[i] = 1 } else {sim1[i] = 0}
  sigmab = sqrt(exp(log(c * var(camp))))
```

```
  LL2 = mean(camp) - z * sigmab/(sqrt(n))
  UL2 = mean(camp) + z * sigmab/(sqrt(n))
  if(LL2 < mu & UL2 > mu) {sim2[i] = 1} else{sim2[i] = 0}
  mu3 = mean((camp - mean(camp))^3)
  LL3 = (mean(camp) + mu3/(6*var(camp)*n)) - t * sd(camp)/(sqrt(n))
  UL3 = (mean(camp) + mu3/(6*var(camp)*n)) + t * sd(camp)/(sqrt(n))
  if(LL3 < mu & UL3 > mu) {sim3[i] = 1} else{sim3[i] = 0}
  LL4 = gamma * exp(mean(log(camp)) - t * sd(log(camp))/(sqrt(n)))
  UL4 = gamma * exp(mean(log(camp)) + t * sd(log(camp))/(sqrt(n)))
  if(LL4 < mu & UL4 > mu){sim4[i] = 1} else{sim4[i] = 0}}
out1 = mean(sim1)
out2 = mean(sim2)
out3 = mean(sim3)
out4 = mean(sim4)
```

A.6 Codice per le stime in Tabella 4.1 e 5.1

Il vero valore della media cambia per ciascuna distribuzione e cambia anche il valore della numerosità campionaria; i valori saranno:

- mu: 0 per logN(0,1), 0.5 per Unif(0, 1), 0.5 per Beta(3, 3), 1 per Chi-quadro(1), 1 per Exp(1).
- n1: 10, 20, 30, 40, 50, 100.
- n2: 15, 25, 35, 45, 55, 110.

```
# LOGN(0,1)
n1 = 10
n2 = 15
M = 1000
mu = 0
sim1 = matrix(NA, M)
sim2 = matrix(NA, M)
sim3 = matrix(NA, M)
sim4 = matrix(NA, M)
mux1 = matrix(NA, M)
mux2 = matrix(NA, M)
mug1 = matrix(NA, M)
mug2 = matrix(NA, M)
for(i in 1:M){
alpha = 0.05
  z = qnorm(1-alpha/2, 0, 1)
  c1 = n1/(n1-z)
  c2 = n2/(n2-z)
  t = qt(1 - alpha/2, (n1+n2-2))
  for(j in 1:M){
    camp1 = rlnorm(n1, 0, 1)
```

```
    mux1[j] = mean(camp1)
    mug1[j] = mug_x(camp1, n1)}
  for(l in 1:M){
    camp2 = rlnorm(n2, 0, 1)
    mux2[l] = mean(camp2)
    mug2[l] = mug_x(camp2, n2)}
  mod1 = lm(mux1 ~ 0 + mug1)
  gamma1 = coef(mod1)
  mod2 = lm(mux2 ~ 0 + mug2)
  gamma2 = coef(mod2)
  LL1 = (mean(camp1)-mean(camp2)) - z * (sqrt((var(camp1)/n1)
  + (var(camp2)/n2)))
  UL1 = (mean(camp1)-mean(camp2)) + z * (sqrt((var(camp1)/n1)
  + (var(camp2)/n2)))
  if(LL1 < mu & UL1 > mu){sim1[i] = 1} else{sim1[i] = 0}
  amp1 = UL1-LL1
  sigmab1 = (exp(log(c1 * var(camp1))))
  sigmab2 = (exp(log(c2 * var(camp2))))
  LL2 = (mean(camp1)-mean(camp2)) - z * (sqrt((sigmab1/n1)
  + (sigmab2/n2)))
  UL2 = (mean(camp1)-mean(camp2)) + z * (sqrt((sigmab1/n1)
  + (sigmab2/n2)))
  if(LL2 < mu & UL2 > mu){sim2[i] = 1} else{sim2[i] = 0}
  amp2=UL2-LL2
  mu13 = mean((camp1 - mean(camp1))^3)
  mu23 = mean((camp2 - mean(camp2))^3)
  a1 = mean(camp1) + (mu13/(6 * n1 * var(camp1)))
  a2 = mean(camp2) + (mu23/(6 * n2 * var(camp2)))
  LL3 = (a1 - a2) - t * sqrt((var(camp1)/n1 + var(camp2)/n2))
  UL3 = (a1 - a2) + t * sqrt((var(camp1)/n1 + var(camp2)/n2))
  if(LL3 < mu & UL3 > mu){sim3[i] = 1} else{sim3[i] = 0}
  amp3=UL3-LL3
  LL = exp(mean(log(camp1)) - mean(log(camp2))
  - t * sqrt((var(log(camp1))/n1)) + (var(log(camp2))/n2)))
  UL = exp(mean(log(camp1)) - mean(log(camp2))
  + t * (sqrt((var(log(camp1))/n1)) + (var(log(camp2))/n2)))
  LL4 = (gamma1/gamma2) * LL
  UL4 = (gamma1/gamma2) * UL
  if(LL4 < 1 & UL4 > 1){sim4[i] = 1} else{sim4[i] = 0}
  amp4=UL4-LL4}
out1 = mean(sim1)
out2 = mean(sim2)
out3 = mean(sim3)
out4 = mean(sim4)
```

A.7 Codice per il grafico in Figura 2.1

```
n = 50
M = 10000
g = n - 1
var_camp = numeric(M)
for (i in 1:M) {
  camp = rnorm(n, 0, 1)
  var_camp[i] = var(camp) * (n - 1) / n }
media_var = mean(var_camp)
varianza_var = var(var_camp)
hist(var_camp, breaks = 30, probability = TRUE,
     main = paste("Distribuzione Campionaria della Varianza"),
     xlab = "Varianza Campionaria", col = "lightblue",
     border = "black",
     ylab = "Densita")
# Curva teorica chi-quadro
curve(dchisq(x * (n - 1), g) * (n - 1),
    add = TRUE, col = "blue", lwd = 2)
text(1.4,1.5, "Curva teorica Chi-Quadro", col = "blue")
```

A.8 Codice per il grafico in Figura 2.2

```
n = 50
M = 10000
g = n - 1
log_var = numeric(M)
for (i in 1:M) {
  camp = rnorm(n, 0, 1)
  var_camp = var(camp) * (n - 1) / n
  log_var[i] = log(var_camp)}
media_log = mean(log_var)
sd_log = sd(log_var)
hist(log_var, breaks = 30, probability = TRUE,
     main = paste("Distribuzione Campionaria del logaritmo
     della varianza"),
     xlab = "Log(Varianza Campionaria)", col = "lightblue",
     border = "black", ylab = "Densita")
# Curva normale teorica
curve(dnorm(x, media_log, sd_log),
    add = TRUE, col = "blue", lwd = 2)
text(0.4,1.5, "Curva teorica Normale", col = "blue")
```

A.9 Codice per il grafico in Figura 3.1 e 3.2

```
n = 100
M = 10000
g = 1
var = 2 * g
smf = matrix(NA, M)
smn = matrix(NA, M)
lb0 = numeric(M)
ub0 = numeric(M)
lb1 = numeric(M)
ub1 = numeric(M)
for(i in 1:M){
  camp1 = rchisq(n, g)
  camp2 = rchisq(n, g)
  perc = 1/(2*(n-4)^(1/2))
  gamma41_1 = n*sum((camp1-mean(camp1))^4)/
  (sum((camp1-mean(camp1))^2))^2
  gamma41_2 = n*sum((camp2-mean(camp2))^4)/
  (sum((camp2-mean(camp2))^2))^2
  gamma42_1 = n*sum((camp1-mean(camp1,perc/2))^4)/
  (sum((camp1-mean(camp1))^2))^2
  gamma42_2 = n*sum((camp2-mean(camp2,perc/2))^4)/
  (sum((camp2-mean(camp2))^2))^2
  gamma43_1 = n*sum((camp1-median(camp1))^4)/
  (sum((camp1-mean(camp1))^2))^2
  gamma43_2 = n*sum((camp2-median(camp2))^4)/
  (sum((camp2-mean(camp2))^2))^2
  alpha = 0.05
  z = qnorm(1-alpha/2, 0, 1)
  c = n/(n-z)
  se11_1 = c * ((gamma41_1 - (n-3)/n)/(n-1))^(1/2)
  se11_2 = c * ((gamma41_2 - (n-3)/n)/(n-1))^(1/2)
  se12_1 = c * ((gamma42_1 - (n-3)/n)/(n-1))^(1/2)
  se12_2 = c * ((gamma42_2 - (n-3)/n)/(n-1))^(1/2)
  se13_1 = c * ((gamma43_1 - (n-3)/n)/(n-1))^(1/2)
  se13_2 = c * ((gamma43_2 - (n-3)/n)/(n-1))^(1/2)
  LLn = (var(camp2)/var(camp1)) * qf((alpha/2), n-1, n-1)
  lb0[i] = LLn
  ULn = (var(camp2)/var(camp1)) * qf((1-alpha/2), n-1, n-1)
  ub0[i] = ULn
  if(LLn < var & ULn > var) {smn[i] = 1} else{smn[i] = 0}
  f11 = sum((camp1 - mean(camp1))^4)
  f12 = sum((camp2 - mean(camp2))^4)
  mu4f1 = (f11 + f12)/(2*n)
  s2f1 = ((n-1) * var(camp1) + (n-1) * var(camp2))/(n+n)
  gamma4f1 = mu4f1/(s2f1^2)
```

```
  r1 = (2*n) / (gamma4f1 - ((n-3)/(n-1)))
  r2 = (2*n) / (gamma4f1 - ((n-3)/(n-1)))
  LLf1 = (var(camp2)/var(camp1)) * qf((alpha/2), r1, r2)
  lb1[i] = LLf1
  ULf1 = (var(camp2)/var(camp1)) * qf((1-alpha/2), r1, r2)
  ub1[i] = ULf1
  if(LLf1 < var & ULf1 > var) {smf[i] = 1} else {smf[i] = 0}}
outF1 = mean(smf)
outN = mean(smn)
# Grafico 3.1
data1 = data.frame(
  simu = 1:M,
  lower = lb0,
  upper = ub0,
  met = "NORMALE",
  cop = smn)
ggplot(data1, aes(x = simu, ymin = lower, ymax = upper)) +
  geom_linerange(aes(color = factor(cop))) +
  facet_wrap(~ met, ncol = 1) +
  geom_hline(yintercept = var, linetype = "dashed", color = "red") +
  labs(title = "Intervalli di confidenza NORMALE per la varianza",
       x = "Simulazioni", y = "Varianza") +
  scale_color_manual(values = c("1" = "blue", "0" = "lightblue"),
  labels = c("Non Copre", "Copre")) +
  theme_minimal() +
  theme(legend.title = element_blank())
# Grafico 3.2
data2 = data.frame(
  simu = 1:M,
  lower = lb1,
  upper = ub1,
  met = "F",
  cop = smf)
ggplot(data2, aes(x = simu, ymin = lower, ymax = upper)) +
  geom_linerange(aes(color = factor(cop))) +
  facet_wrap(~ met, ncol = 1) +
  geom_hline(yintercept = var, linetype = "dashed", color = "red") +
  labs(title = "Intervalli di confidenza F per la varianza",
       x = "Simulazioni", y = "Varianza") +
  scale_color_manual(values = c("1" = "blue", "0" = "lightblue"),
  labels = c("Non Copre", "Copre")) +
  theme_minimal() +
  theme(legend.title = element_blank())
```

Bibliografia

[1] M. S. Bartlett e D. G. Kendall. «The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation». In: *Supplement to the Journal of the Royal Statistical Society* 8.1 (dic. 2018), pp. 128–138. ISSN: 1466-6162. DOI: 10.2307/2983618. eprint: https://academic.oup.com/jrsssb/article-pdf/8/1/128/49093455/jrsssb_8_1_128.pdf. URL: https://doi.org/10.2307/2983618.

[2] Douglas G. Bonett. «Approximate confidence interval for standard deviation of nonnormal distributions». In: *Computational Statistics Data Analysis* 50.3 (2006), pp. 775–782. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2004.10.003. URL: https://www.sciencedirect.com/science/article/pii/S0167947304002993.

[3] José Dias Curto. «Confidence intervals for means and variances of nonnormal distributions». In: *Communications in Statistics - Simulation and Computation* 52.9 (2023), pp. 4414–4430. DOI: 10.1080/03610918.2021.1963448. eprint: https://doi.org/10.1080/03610918.2021.1963448. URL: https://doi.org/10.1080/03610918.2021.1963448.

[4] *Curtosi*. URL: https://www.treccani.it/enciclopedia/curtosi_%28Dizionario-di-Economia-e-Finanza%29/.

[5] Changyong Feng et al. «Log transformation: application and interpretation in biomedical research». In: *Statistics in Medicine* 32.2 (2013), pp. 230–239. DOI: https://doi.org/10.1002/sim.5486. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5486. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5486.

[6] *Fisher, distribuzione F*. URL: https://www.treccani.it/enciclopedia/distribuzione-f-di-fisher_(Enciclopedia-della-Matematica)/.

[7] Marco Minozzo Giuseppe Cicchitelli Pierpaolo D'Urso. *Statistica: Principi e Metodi*. Pearson, 2018.

[8] Norman J. Johnson. «Modified t Tests and Confidence Intervals for Asymmetrical Populations». In: *Journal of the American Statistical Association* 73.363 (1978), pp. 536–544. DOI: 10.1080/01621459.1978.10480051. eprint: https://doi.org/10.1080/01621459.1978.10480051. URL: https://doi.org/10.1080/01621459.1978.10480051.

[9] Roger W. Johnson. «An Introduction to the Bootstrap». In: *Teaching Statistics* 23.2 (2001), pp. 49–54. DOI: https://doi.org/10.1111/1467-9639.00050. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9639.00050. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9639.00050.

[10] Lewis H Shoemaker. «Fixing the F Test for Equal Variances». In: *The American Statistician* 57.2 (2003), pp. 105–114. DOI: 10.1198/0003130031441. eprint: https://doi.org/10.1198/0003130031441. URL: https://doi.org/10.1198/0003130031441.

[11] Rand R. Wilcox. «A note on computing a confidence interval for the mean». In: *Communications in Statistics - Simulation and Computation* 53.1 (2024), pp. 164–166. DOI: 10.1080/03610918.2021.2011926. eprint: https://doi.org/10.1080/03610918.2021.2011926. URL: https://doi.org/10.1080/03610918.2021.2011926.

[12] Jeffrey M. Woolridge. *Introductory Econometrics: A modern approch*. South-Western, Mason, 2020.