

# **Statistica Matematica**

## Dispensa del corso

Pier Luigi Novi Inverardi  
*a.a. 2020-2021*

6 novembre 2020



# Premessa

La realizzazione di questa dispensa è stata resa possibile grazie al lavoro di riscrittura di appunti e note relative alle lezioni del corso di Statistica Matematica nell'a.a.2019-20 da parte di **Chiara Avigo** e di **Andrea Gadotti, Michele Nardin e Marco Peruzzetto** in anni accademici precedenti. La nascita di questo strumento didattico si deve massimamente al loro impegno e alla loro disponibilità. Recentemente ho ripreso e omogeneizzato i loro contributi, rielaborandoli e aggiungendo quanto mancava a completamento dell'opera: questo ne è il risultato.

A Chiara, Andrea, Michele e Marco va fin d'ora il mio più sincero ringraziamento per il prezioso lavoro svolto e messo a disposizione della comunità degli studenti del corso di Statistica Matematica. Va da sé che mi assumo tutta la responsabilità di errori, imprecisioni o refusi presenti nel testo. Ogni loro segnalazione sarà gradita e anche di questo, ringrazio fin d'ora.

Buona lettura!



**Attribuzione - Non commerciale - Non opere derivate**

**CC BY-NC-ND**

Quest'opera è distribuita con Licenza Pubblica Creative Commons Attribuzione-NonCommerciale-NonOpereDerivate 4.0 Internazionale - Condividi allo stesso modo 4.0 Italia.

Per leggere una copia della licenza visita il sito web

<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode.it>

# Introduzione

La statistica, come avremo modo di vedere, è essenzialmente un *processo induttivo* che, partendo da una quantità di *informazione limitata*, contenuta in un insieme di osservazioni disponibili in merito a un certo fenomeno (usualmente stocastico) arriva a conclusioni di carattere generale relativamente al processo generante i dati osservati e, di conseguenza, alla *descrizione/previsione* del fenomeno stesso. In questo senso, il metodo statistico è considerato un metodo *logico-inferenziale* e, in virtù della condizione cronica di informazione limitata che deriva dal non poter osservare il fenomeno in tutte le sue manifestazioni, porta a conclusioni *vere in probabilità*. Anzi, potremmo dire che un metodo statistico è tanto più buono quanto aumenta la probabilità della verità delle conclusioni raggiunte. Avremo modo di tornare sull'argomento nel corso delle lezioni, con dovizia di particolari.

Esistono due grandi approcci alla *statistica classica*: quello *parametrico* e quello *non parametrico*. Il primo si distingue dal secondo per il fatto di assumere che le osservazioni disponibili siano state generate da un processo che è possibile rappresentare in termini di un modello matematico di forma funzionale nota e indicizzata da uno o più parametri (tutti o in parte di valore incognito), al variare dei quali cambiano le caratteristiche, espresse in termini di *funzioni dei parametri*, del modello stesso. Divesamente, in ambito *non parametrico*, non si assume alcuna forma funzionale tantomeno nota per il modello matematico (ossia siamo in un contesto *distribution-free*) a meno di qualche caratteristica molto generale quale, per esempio, la simmetria giustificabile alla luce delle caratteristiche del fenomeno cui i dati si riferiscono (per esempio, se interessati a misure antropometriche, la distribuzione dei dati a loro riferiti presenta una spiccata simmetria rispetto al valor medio). In questo ambito ci si serve di particolari strumenti statistici particolarmente *flessibili* per arrivare a una caratterizzazione del problema studiato *esclusivamente sulla base dei dati*: sono i dati a guidare in modo determinante l'inferenza e la produzione dei risultati (secondo il principio: *let the data speak for themselves* da considerare con attenzione in tempi di BigData).

Viene da sè che mentre un modello parametrico è caratterizzato da un numero *finito* (e sperabilmente *piccolo*) di parametri, un modello non parametrico può ammettere un *gran numero* di parametri (potenzialmente infinito) pari al numero delle osservazioni di cui si dispone. L'ambito in cui opereremo (parametrico o non parametrico) sarà evidente dal contesto e dalla notazione che conseguentemente adotteremo ma possiamo fin d'ora anticipare che molta parte del corso vivrà in ambito parametrico. In quest'ambito, scelto il modello opportuno per il fenomeno analizzato, essotipicamente dipende da uno o più parametri incogniti. Solo se si riesce a *dare un valore* a tali parametri allora la distribuzione da cui si ritiene provengano i dati è completamente nota e, a quel punto, si sa tutto del fenomeno analizzato. Ma, in generale, di tali parametri si sa solo che assumono valori in un certo insieme, detto spazio parametrico, e non di più. Come fare allora per determinare i parametri?

Anzi, più in generale, come rispondere ai tre problemi alla base dell'inferenza statistica sui parametri d'interesse, ovvero:

1. **Stima puntuale.** Come tentare di indovinare il valore dei parametri?
2. **Stima per intervallo.** Come determinare, per ciascun parametro d'interesse, un intervallo che ne contiene il vero valore "con alta probabilità"?

**3. Test delle ipotesi.** Come procedere all'accettazione o al rifiuto di una particolare ipotesi formulata sui parametri?

Per rispondere alle domande di cui sopra occorre sfruttare qualche informazione. Come informazione di base l'inferenza statistica utilizza un certo insieme di osservazioni del fenomeno oggetto di studio: tale insieme è detto campione. In sintesi: sulla base dell'*informazione* contenuta *nel campione* si vuol fare *inferenza sui parametri* del modello scelto e, di conseguenza, sulla legge probabilistica che esprime formalmente il *meccanismo generatore* dei dati del fenomeno analizzato.

Poco fa, trattando di approccio parametrico e non parametrico alla statistica, abbiamo usato per quest'ultima l'aggettivo *classica*. Vi è anche una statistica *non classica*? La risposta è sì e in questo caso si fa riferimento all'approccio bayesiano all'inferenza statistica. Semplificando molto, la *statistica bayesiana* poggia su una diversa concezione di che *cos'è la probabilità*.

La statistica classica adotta una concezione *frequentista* di probabilità: la probabilità di un evento è concepita quale limite a cui tende la *frequenza relativa* di accadimento dell'evento stesso. La concezione bayesiana di probabilità è sostanzialmente diversa: la probabilità di un evento viene interpretata come *aspettazione razionale* rappresentante uno stato di conoscenza o, in altre parole, come *quantificazione di una convinzione personale* (o *degree of belief*) che il ricercatore ha in merito all'accadimento dell'evento di cui interessa la probabilità, credenza che poggia su informazioni *a priori* che, generalmente, hanno natura soggettiva oltre che sull'evidenza sperimentale. Non ci addentreremo in discussioni filosofiche, perlomeno estremamente interessanti sui fondamento della probabilità, ma preme osservare che il tema del contendere tra ifrequentisti e bayesiani è relativo alla *sorgente* della probabilità: è *insita nel fenomeno* stocastico per i frequentisti che sperano di ricavare informazione dall'osservazione ripetuta dello stesso (*oggettività*) mentre è *insita nel ricercatore* per i bayesiani che non disdegnano di far interagire livelli di conoscenza personali e evidenza sperimentale (*soggettività*). Il nome di *statistica bayesiana*, perlomeno, viene dal fatto di utilizzare il *teorema di Bayes* come strumento fondamentale di inferenza. Dedicheremo un po' di tempo a quest'ultimo approccio all'inferenza statistica alla fine del corso.

# Indice

<b>1 Un veloce ripasso di quello che serve ricordare...</b>	<b>6</b>
1.1 Variabili casuali e loro distribuzione . . . . .	6
1.1.1 Alcune distribuzioni comunemente usate . . . . .	7
1.2 Teoremi di trasformazione di variabili casuali . . . . .	9
1.3 Valori attesi . . . . .	13
1.3.1 Funzione generatrice dei momenti e momenti di una distribuzione . . . . .	18
1.4 Famiglie esponenziali . . . . .	23
1.5 Distribuzioni multivariete . . . . .	26
1.5.1 Distribuzione Multinomiale . . . . .	26
1.5.2 Distribuzione Normale multivariata . . . . .	27
<b>2 Popolazioni, campioni e statistiche</b>	<b>31</b>
2.1 Funzione di distribuzione empirica e principio del plug-in . . . . .	34
2.1.1 Uso della funzione di distribuzione empirica . . . . .	36
2.1.2 Il metodo dei momenti . . . . .	41
2.2 Statistiche ordinate . . . . .	43
2.2.1 I cinque numeri magici e l'analisi esplorativa dei dati . . . . .	45
<b>3 Non distorsione, consistenza e distribuzioni limite</b>	<b>50</b>
3.1 Convergenza in probabilità . . . . .	51
3.1.1 Alcuni utili risultati relativi alla convergenza in probabilità . . . . .	56
3.2 Convergenza in distribuzione . . . . .	59
3.2.1 Alcuni utili risultati relativi alla convergenza in distribuzione . . . . .	61
3.2.2 Delta method . . . . .	64
3.2.3 Funzione generatrice dei momenti e convergenza in distribuzione . . . . .	69
3.3 Convergenza in media quadratica . . . . .	70
<b>4 Un piccolo assaggio di teoria delle decisioni</b>	<b>75</b>
4.1 Introduzione . . . . .	75
4.2 Regole di decisione . . . . .	75
4.3 Perdita e rischio . . . . .	76
<b>5 Metodi elementari di inferenza statistica</b>	<b>79</b>
5.1 Statistiche pivot . . . . .	79
5.2 Campionamento da popolazione Normale . . . . .	80
5.2.1 Distribuzioni campionarie . . . . .	83
5.2.1.1 Distribuzione chi-quadrato . . . . .	83
5.2.1.2 Distribuzione $t$ di Student . . . . .	84

5.2.1.3 Distribuzione di Fisher-Snedecor . . . . .	86
5.3 Intervalli di confidenza . . . . .	88
5.3.1 Intervalli di confidenza esatti . . . . .	89
5.3.2 Intervalli di confidenza approssimati . . . . .	92
5.3.3 Intervalli di confidenza per differenze . . . . .	94
5.3.3.1 Intervalli di confidenza per la differenza di medie . . . . .	94
5.3.3.2 Intervallo di confidenza per la differenza di proporzioni . . . . .	96
5.3.3.3 Intervallo di confidenza per il rapporto di varianze . . . . .	97
5.4 Test di ipotesi . . . . .	101
5.4.1 Regola di decisione e potenza del test . . . . .	102
5.4.1.1 Test per la media (campionamento da $N(\mu, \sigma^2)$ con $\sigma^2$ noto) . . . . .	104
5.4.1.2 Test per la media (campionamento da $N(\mu, \sigma^2)$ con $\sigma^2$ non noto) . . . . .	106
5.4.1.3 Test per la proporzione (campionamento da $b(1, p)$ ) . . . . .	108
5.4.1.4 Test unilaterale sulla media per grandi campioni . . . . .	110
5.4.1.5 Test sulla varianza . . . . .	112
5.4.1.6 Test bilaterali . . . . .	113
5.4.1.7 Test bilaterale sulla media (esatto) . . . . .	113
5.4.1.8 Test bilaterale per differenza di medie (esatto) . . . . .	115
5.4.1.9 Test per differenza di proporzioni . . . . .	116
5.4.1.10 Test per il confronto di varianze . . . . .	116
5.5 Test di Kolmogorov-Smirnov . . . . .	118
5.5.0.1 Test di Kolmogorov-Smirnov per un campione . . . . .	121
5.5.0.2 Test di Kolmogorov-Smirnov per due campioni . . . . .	122
5.6 <i>p</i> -value . . . . .	124
<b>6 Bootstrap</b>	<b>127</b>
6.1 Ricampionamento . . . . .	127
6.2 Il bootstrap . . . . .	128
6.2.1 La tecnica bootstrap . . . . .	129
6.2.1.1 Distribuzione e inferenza bootstrap . . . . .	130
6.3 Appendice . . . . .	132
<b>7 Regressione lineare</b>	<b>133</b>
7.1 Il modello . . . . .	133
7.1.1 Le ipotesi "classiche" . . . . .	134
7.1.2 Il metodo di stima dei minimi quadrati . . . . .	134
7.2 Inferenza sul modello di regressione . . . . .	137
7.2.1 Test di ipotesi e intervalli di confidenza per i parametri del modello di regressione . . . . .	140
7.2.2 Test per la significatività del modello . . . . .	141
7.3 Un breve cenno al modello di regressione lineare multipla . . . . .	145
<b>8 Principles of Data Reduction</b>	<b>147</b>
8.1 Verosimiglianza e principio di verosimiglianza . . . . .	147
8.1.1 Principio formale di verosimiglianza . . . . .	151
8.2 Sufficienza . . . . .	152

<b>9 Methods of Evaluating Estimators</b>	<b>163</b>
9.1 Viaggio alla ricerca dello stimatore migliore . . . . .	163
9.2 Efficienza e estimatori efficienti . . . . .	165
9.2.0.1 Efficienza (assoluta e relativa) . . . . .	170
9.2.0.2 Estensioni della disuguaglianza di Rao-Cramér . .	171
9.3 Score function ed efficienza . . . . .	174
9.4 Stimatori UMVU . . . . .	177
<b>10 Stimatori di massima verosimiglianza</b>	<b>188</b>
10.1 Proprietà degli estimatori di massima verosimiglianza . . . . .	188
10.2 Metodi numerici per la massima verosimiglianza . . . . .	200
10.2.1 Tecnica di cattura-ricattura . . . . .	200
10.2.2 Metodo di Newton-Raphson . . . . .	201
10.2.3 Modello di regressione logistica . . . . .	205
<b>11 Likelihood Ratio Tests</b>	<b>211</b>
11.1 Test per dati appaiati . . . . .	218
11.2 Test sulle medie di M popolazioni . . . . .	218
11.3 Ancillarità . . . . .	220
<b>12 Statistica bayesiana</b>	<b>221</b>
12.1 Introduzione . . . . .	221
12.2 Stimatori di Bayes . . . . .	222
<b>Bibliography</b>	<b>224</b>

# 1 Un veloce ripasso di quello che serve ricordare...

Vale la pena spendere qualche minuto per richiamare alcuni concetti di fondamentale importanza in statistica matematica.

## 1.1 Variabili casuali e loro distribuzione

Nella teoria della probabilità, una *variabile casuale* (detta anche *variabile aleatoria*) è un'opportuna funzione che può assumere valori diversi (e, in genere, con diversa probabilità) in dipendenza da qualche fenomeno aleatorio. Più formalmente,

**Definizione 1.1.1** (Variabile casuale). Dato uno spazio di probabilità  $(\Omega, \mathcal{A}, \nu)$  e dato uno spazio misurabile  $(E, \mathcal{E})$ , una  $(E, \mathcal{E})$ -variabile casuale è una funzione misurabile  $X : \Omega \rightarrow E$ .

In questa definizione si intende che una funzione  $X$  è *misurabile* se per ogni  $A \in \mathcal{E}$  si ha che  $X^{-1}(A) \in \mathcal{A}$ . Se  $E$  è uno spazio topologico e  $\mathcal{E}$  è la sigma-algebra di Borel allora  $X$  è detta anche  $E$ -variabile casuale. Inoltre se  $E = \mathbb{R}^n$  allora  $X$  è detta semplicemente variabile casuale.

La misura di probabilità indotta sullo spazio misurabile di arrivo  $(E, \mathcal{E})$  da una variabile aleatoria  $X$  a partire dalla misura di probabilità  $\nu$  su  $(\Omega, \mathcal{A})$  è detta *distribuzione, o legge, di probabilità*, di  $X$  ed è indicata con  $P_X$  definita nel seguente modo

$$P_X(A) := \nu(X^{-1}(A)), \quad \forall A \in \mathcal{E}.$$

Essa è ben definita proprio perché  $X^{-1}(A) \in \mathcal{A}$  per ogni  $A \in \mathcal{E}$ . Quando la variabile aleatoria è chiara dal contesto spesso si omette il pedice  $X$ . Per brevità, invece di scrivere  $\nu(X^{-1}(A))$  o  $\nu(\omega \in \Omega : X(\omega) = x)$  per ogni  $x \in A$  spesso si usa la notazione  $P_X(A) = P(X \in A)$ .

Nel corso delle lezioni considereremo essenzialmente il caso  $E = \mathbb{R}^n$ , ovvero variabili aleatorie a valori reali. Vale la pena qui richiamare alcune utili definizioni.

**Definizione 1.1.2** (Variabile casuale discreta). Una variabile casuale è detta *discreta* se l'insieme dei suoi possibili valori  $x_i$  (ovvero il suo supporto  $\mathcal{S}_X$ ) ha cardinalità finita o al più numerabile. Si definisce inoltre *funzione di massa probabilistica* la funzione che ad ogni elemento  $x_i$  associa la probabilità che la variabile casuale discreta assuma tale valore ossia,  $f_X(x_i) = P(X = x_i)$ .

**Definizione 1.1.3** (Variabile casuale continua). Una variabile casuale è detta *continua* se può assumere tutti gli infiniti valori compresi in un intervallo  $I$ , limitato o

illimitato, di numeri reali (nel qual caso il supporto di  $X$  è restituito da  $S_X = I$ ). A essa rimane associata una *funzione densità di probabilità*  $f_X(x)$ , anch'essa necessariamente continua e in stretta relazione con la funzione di ripartizione.

Nel caso di variabili aleatorie a valori reali, la legge di probabilità della variabile casuale  $X$  è *univocamente* individuata dalla sua *funzione di ripartizione*.

**Definizione 1.1.4** (Funzione di ripartizione). Definiamo la funzione

$$F_X(x) = P(X \leq x).$$

con  $F_X(x)$  funzione *non negativa, non decrescente, continua a destra* e tale che

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \text{e} \quad \lim_{x \rightarrow +\infty} F_X(x) = 1.$$

*funzione di ripartizione* della variabile aleatoria  $X$ .

In sintesi, la funzione di ripartizione è una funzione di variabile reale che racchiude le informazioni su un fenomeno (un insieme di dati, un evento casuale) riguardanti la sua presenza o la sua distribuzione prima o dopo un certo punto; in altri termini, ne descrive il comportamento (sia esso aleatorio o meno).

Data una variabile casuale (assolutamente) continua  $X$ , la sua funzione di ripartizione  $F_x(x)$  è legata alla sua densità di probabilità  $f_X(x)$  dalla relazione

$$F_X(x) = P_X(X \leq x) = \int_{-\infty}^x f_X(u) du$$

e di conseguenza,

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Laddove, invece,  $X$  sia una variabile aleatoria discreta, tra funzione di ripartizione e funzione di massa probabilistica vale la seguente relazione:

$$F_X(x) = P_X(X \leq x) = \sum_{u \leq x_i} P_X(X = u)$$

e quindi,

$$P_X(X = x) = F_X(x) - F_X(x^-)$$

### 1.1.1 Alcune distribuzioni comunemente usate

Le distribuzioni che seguono sono alcune di quelle più utilizzate in ambito statistico e, come si può facilmente notare, sono parametrizzate da uno o più parametri; al variare del valore assunto da queste quantità numeriche, varia la forma della distribuzione. Spesso queste distribuzioni costituiscono dei mattoncini elementari a partire dai quali costruire distribuzioni più articolate e adatte a descrivere la coplessità dei fenomeni del mondo reale.

Diremo *bernelliana* o *di Bernoulli* una v.c.  $X$  che può assumere esclusivamente due valori, usualmente 0 e 1 e scriveremo  $X \sim b(1, p)$  con  $p \in [0, 1]$ . Essa è generalmente associata all'esito di un esperimento dicotomico, ovvero che può

avere come risultato un successo (codificato con 1) o un fallimento (codificato con 0). Tale variabile dipende da un parametro,  $p \in [0, 1]$ , che rappresenta la probabilità di successo nell'esperimento di Bernoulli, ovvero la probabilità che la variabile di Bernoulli assuma il valore 1; di conseguenza, la probabilità che la medesima variabile assuma valore zero sarà pari a  $P_X(X = 0) = 1 - p$ . Il comportamento aleatorio di questa v.c. è descritto dalla seguente funzione di massa:

$$P_X(X = x) = p^x(1 - p)^{1-x} \mathbb{1}_{\{0,1\}}(x)$$

dove, in questo caso, il supporto di  $X$  è  $\mathcal{S}_X = \{0, 1\}$ .

Si definisce inoltre *processo di Bernoulli* una successione di  $n$  variabili casuali indipendenti e identicamente distribuite secondo  $b(1, p)$ . Si definisce *distribuzione binomiale*  $b(n, p)$  la distribuzione che descrive la probabilità associata al numero di successi in un processo di Bernoulli con probabilità di successo pari a  $p \in [0, 1]$ . La sua funzione di massa è definita come segue:

$$P_X(X = x) = \binom{n}{x} p^x(1 - p)^{n-x} \mathbb{1}_{\{0,\dots,n\}}(x)$$

dove, in questo caso, il supporto di  $X$  è  $\mathcal{S}_X = \{0, 1, \dots, n\}$ .

In un processo di Bernoulli, sia inoltre  $T$  il numero aleatorio associato al primo successo. La distribuzione di probabilità ad esso associata è nota come *distribuzione geometrica*  $G(p)$ , ed è definita come segue:

$$P_X(X = x) = p(1 - p)^{x-1} \mathbb{1}_{\{0,\dots,n\}}(x) \quad (1.1)$$

con  $\mathcal{S}_X = \{0, \dots, n\}$  dato dai primi  $n$  numeri naturali, zero incluso.

Si definisce *distribuzione di Poisson*  $P(\lambda)$  la distribuzione

$$P_X(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \mathbb{1}_{\mathbb{N}}^*(x)$$

con  $\lambda \in \mathbb{R}_+$ . La distribuzione di Poisson viene spesso utilizzata per descrivere il numero di eventi che si verifica in un fissato intervallo temporale  $\Delta t$ , sapendo che mediamente in tale intervallo se ne verifica un numero  $\bar{x}$ . In tal caso, si sceglie  $\lambda = \bar{x}$ .

Una variabile casuale continua ha *distribuzione Normale*  $N(\mu, \sigma^2)$ , con  $\mu \in \mathbb{R}$  e  $\sigma^2 \geq 0$ , se la sua densità di probabilità è definita come segue:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} \mathbb{1}_{\mathbb{R}}(x)$$

Spesso, ai fini del calcolo di probabilità utilizzando tale distribuzione, si effettua la *trasformazione standardizzante*  $Z = (X - \mu)/\sigma$ . Dimostreremo in seguito che la variabile che si ottiene mediante tale trasformazione segue la distribuzione  $N(0, 1)$ .

Una variabile casuale continua ha *distribuzione gamma*  $\mathcal{G}(\alpha, \beta)$ , con  $\alpha, \beta > 0$  se la sua densità di probabilità è definita come segue:

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left\{-\frac{x}{\beta}\right\} \mathbb{1}_{\mathbb{R}^+}(x)$$

dove  $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$  è una funzione i cui valori sono tabulati. Ricordiamo inoltre che per  $\alpha \in \mathbb{N}$ , vale che  $\Gamma(\alpha) = (\alpha - 1)!$ . Inoltre, se  $\alpha = 1$ , la distribuzione  $\mathcal{G}(1, \beta)$  è nota come *distribuzione esponenziale*, con densità di probabilità

$$f(x, \beta) = \frac{1}{\beta} \exp\left\{-\frac{x}{\beta}\right\} \mathbb{1}_{\mathbb{R}^+}(x)$$

## 1.2 Teoremi di trasformazione di variabili casuali

Spesso in statistica si pone il seguente problema: si dispone di una variabile casuale  $X$  di cui è nota la distribuzione  $F_X(x; \theta)$  o la funzione di densità  $f_X(x; \theta)$  o quella di massa  $f_X(x; \theta) = P_\theta(X = x)$  a meno del valore del parametro  $\theta$  (scalare o vettore che sia) e, però, si è interessati a conoscere la distribuzione di una sua trasformazione  $Y = h(X)$ . Al fine di non complicare troppo la notazione, omettiamo la dipendenza della distribuzione sul parametro  $\theta$  limitandoci a scrivere  $F_X(x)$  in luogo di  $F_X(x; \theta)$  o  $f_X(x)$  in luogo di  $f_X(x; \theta)$ .

Partiamo dalla considerazione che se  $X$  è una v.c. (scalare o vettore che sia) allora ogni sua funzione è, a sua volta, una v.c. ossia  $Y = h(X) \sim F_Y(y)$ . Da ciò segue che possiamo descrivere il comportamento probabilistico di  $Y$  in termini di quello di  $X$ ; ovvero per un insieme  $A$  qualunque,

$$P(Y \in A) = P(h(X) \in A)$$

notando, appunto, che la distribuzione di  $Y$  è completamente determinata dalle funzioni  $F_X$  e  $h$ . Inoltre, dipendendo dalla scelta di  $h$  talvolta è possibile ottenere espressioni (matematicamente) trattabili per  $P(Y \in A)$ .

Affrontiamo inizialmente il problema per variabili casuali *discrete*, assumendo che la trasformazione  $h$  sia biunivoca. In tal caso, la variabile  $Y$ , anch'essa discreta, avrà al supporto  $\mathcal{S}_Y = \{y = h(x) : x \in \mathcal{S}_X\}$ . Inoltre, la sua funzione di massa sarà data da

$$P(Y = y) = P(h(X) = y) = P(X = h^{-1}(y))$$

**Esempio 1.2.1.** Consideriamo i due seguenti esempi in cui ci si ripromette di trovare la distribuzione di una trasformata  $W = h(X)$  della originaria v.c.  $X$ .

- a) Sia  $X \sim b(n, p)$  con relativa funzione di massa  $f_X(x, p) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{1}_{0,1,\dots,n}(x)$ ,  $n$  noto e  $p \in (0, 1)$ . Consideriamo ora  $Y = h(X) = n - X$ . Allora,

$$P(Y = y) = P(X = n - y) = \binom{n}{n-y} p^{n-y} (1-p)^y \mathbb{1}_{0,1,\dots,n}(y) \quad (1.2)$$

- b) Sia  $X$  una vc tale che  $f_X(x) = P(X = x) = \left(\frac{1}{2}\right)^x \mathbb{1}_{\mathbb{N}}(x)$ ,  $W = X^3$ . Allora,

$$P(Y = y) = P(X^3 = y) = P(X = \sqrt[3]{y}) = \left(\frac{1}{2}\right)^{\sqrt[3]{y}} \mathbb{1}_{1,8,27,64,\dots}(y) \quad (1.3)$$

In generale, il problema delle trasformazioni di variabili casuali, discrete o continue, può essere leggermente più complicato in relazione alla scelta della legge di trasformazione  $h$  e necessita di opportuni teoremi di trasformazione.

**Teorema 1.2.1** (Teorema di trasformazione di variabili casuali discrete). Sia  $X$  una v.c. **discreta** con funzione di massa  $f_X(x) = P(X = x)$  e sia  $\mathcal{S}_X$  il suo supporto. Sia  $Y = h(X)$  una nuova variabile casuale. Allora la funzione di massa della v.c.  $Y$  è

$$f_Y(y) = P(Y = y) = \sum_{\{x \in \mathcal{S}_X : h(x) = y\}} P(X = x), \quad \forall y \in \mathcal{S}_Y$$

mentre  $f_Y(y) = 0, \forall y \notin \mathcal{S}_Y$ .

*Dimostrazione.* Per esercizio. □

**Esempio 1.2.2.** Consideriamo la variabile casuale  $X$ , avente funzione di massa

$$P(X = x) = \begin{cases} \frac{3!}{x!(3-x)!} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{3-x} & \text{per } x = 0, 1, 2, 3 \\ 0 & \text{altrimenti} \end{cases} \quad (1.4)$$

Definiamo inoltre la variabile casuale  $Y = X^2$  e stabiliamone la funzione di massa mediante la regola appena stabilita. Consideriamo quindi la funzione inversa  $x = g^{-1}(y) = \sqrt{y}$  e applichiamola alla funzione di massa, ricordandoci di trasformare anche il supporto della funzione. Otteniamo

$$P(Y = y) = \frac{3!}{(\sqrt{y})!(3 - \sqrt{y})!} \left(\frac{2}{3}\right)^{\sqrt{y}} \left(\frac{1}{3}\right)^{3-\sqrt{y}} \quad \text{per } y = 0, 1, 4, 9 \quad (1.5)$$

**Teorema 1.2.2** (Teorema di trasformazione di variabili casuali continue). Sia  $X$  una v.c. (assolutamente) **continua** con funzione di densità  $f_X(x)$  e sia  $Y = h(X)$ , e sia  $h$  una funzione monotona di  $X$ . Supponiamo inoltre che  $f_X(x)$  sia continua su  $\mathcal{S}_X$ , supporto di  $X$  e che  $h^{-1}(y)$  abbia derivata continua su  $\mathcal{S}_Y$ , supporto di  $Y$ . Allora la funzione di densità di  $Y$  è data da

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right| \mathbb{1}_{\mathcal{S}_Y}(y). \quad (1.6)$$

*Dimostrazione.* Consideriamo  $X \sim F_X(x)$  e dunque

$$F_X(x) = P_\theta(X \leq x) \quad \text{e} \quad f_X(x) = \frac{d}{dx} F_X(x)$$

Allora

$$F_Y(y) = P_\theta(h(X) \leq y) = P_\theta(X \leq h^{-1}(y)) = F_X(h^{-1}(y))$$

e per la *chain rule*,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(h^{-1}(y)) = f_X(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right| \mathbb{1}_{\mathcal{S}_Y}(y)$$

□

**Esempio 1.2.3** (v.c. Normale standard). Sia  $X$  una variabile casuale con distribuzione Normale  $N(\mu, \sigma^2)$ . Sappiamo che la funzione densità associata a tale distribuzione è la seguente:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathbb{1}_{\mathbb{R}}(x)$$

Definiamo la funzione  $Z = g(X) = (X - \mu)/\sigma$ , con relativa funzione inversa  $g^{-1}(Z) = \sigma Z + \mu$ . Tale funzione ha derivata pari a  $\sigma$  sul supporto di  $Z$ . Poiché la sua derivata è continua, possiamo applicare il Teorema 1.2.2 ottenendo la seguente funzione di densità di  $Z$

$$f_Z(z) = f_X(\sigma z + \mu) \cdot |\sigma| \mathbb{1}_{\mathbb{R}}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathbb{1}_{\mathbb{R}}(z)$$

ossia, la variabile casuale  $Z$  segue una distribuzione  $N(0, 1)$ .

**Esempio 1.2.4.** Sia  $X \sim \mathcal{U}(0, 1)$  e sia  $Y = h(X) = -\ln(X)$ . Allora

$$h^{-1}(y) = e^{-y} \quad \text{e} \quad \left| \frac{d}{dy} h^{-1}(y) \right| = |-e^{-y}| = e^{-y}$$

sicché

$$f_Y(y) = 1 \cdot e^{-y} \mathbb{1}_{[0, \infty)}(y) = e^{-y} \mathbb{1}_{\mathcal{S}_Y}(y)$$

con  $\mathcal{S}_Y = [0, \infty)$ . Inoltre,

$$F_Y(y) = \int_{-\infty}^y e^{-v} \mathbb{1}_{(0, \infty)}(v) = (1 - e^{-y}) \mathbb{1}_{\mathcal{S}_Y}(y).$$

con  $\mathcal{S}_Y = [0, \infty)$ . Da notare, infine, che  $Y = -\ln(X) \sim Exp(\theta = 1)$ .

**Esempio 1.2.5.** Sia  $X \sim \mathcal{G}(\alpha = n, \beta)$  con  $n \in \mathbb{N}$  e  $\beta > 0$ , ossia

$$f_x(x) = \frac{1}{\Gamma(n)\beta^n} x^{n-1} e^{\frac{x}{\beta}} \mathbb{1}_{\mathbb{R}^+}(x)$$

e si vuole trovare la funzione di densità di  $Y = h(X) = \frac{1}{X}$ . In questo caso  $\mathcal{S}_X = \mathcal{S}_Y = \mathbb{R}^+$  e  $h^{-1}(y) = y^{-1}$  sicché  $\frac{d}{dy} h^{-1}(y) = -y^{-2}$ . Allora

$$\begin{aligned} f_Y(y) &= f_X(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right| \mathbb{1}_{\mathbb{R}^+}(y) \\ &= \frac{1}{(n-1)! \beta^n} y^{-(n-1)} e^{-\frac{1}{\beta} y^{-1}} y^{-2} \mathbb{1}_{\mathbb{R}^+}(y) \\ &= \frac{1}{(n-1)! \beta^n} y^{-(n+1)} e^{-\frac{1}{\beta} y^{-1}} \mathbb{1}_{\mathbb{R}^+}(y) \end{aligned}$$

che è un caso speciale della densità di una distribuzione (assolutamente) continua nota come distribuzione Gamma inversa.

In molte occasioni la funzione  $Y = h(X)$  di cui interessa ricavare la distribuzione non è monotona sull'intero intervallo  $\mathcal{S}_Y$ ; magari, però, lo è a tratti ovvero su sotto-intervalli disgiunti dell'originario intervallo  $\mathcal{S}_Y$ .

**Corollario 1.2.1** (Monotonia a tratti). Sia  $X$  una v.c. continua avente funzione di densità  $f_X(x)$  e sia  $Y = h(X)$ . Supponiamo esistere una partizione  $A_0, A_1, \dots, A_k$  del supporto  $S_X$  di  $X$  tale che  $P(X \in A_0) = 0$  e sia  $f_X(x)$  continua su ogni  $A_i$ . Supponiamo, inoltre, esistere le funzioni  $h_1(x), h_2(x), \dots, h_k(x)$  rispettivamente definite su  $A_1, A_2, \dots, A_k$  che soddisfano alle seguenti condizioni:

- $h(x) = h_i(x)$  per  $x \in A_i$ ,  $i = 1, 2, \dots, k$
- $h_i(x)$  è monotona su  $A_i$
- l'insieme  $\mathcal{S}_Y = \{y : y = g_i(x) \text{ per qualche } x \in A_i\}$  è lo stesso per ogni  $i$ , con  $i = 1, 2, \dots, k$
- $h_i^{-1}(y)$  ha derivata continua su  $\mathcal{S}_Y$  per ogni  $i = 1, 2, \dots, k$ .

Allora

$$f_Y(y) = \sum_{i=1}^k f_X(h_i^{-1}(y)) \left| \frac{d}{dy} h_i^{-1}(y) \right| \mathbb{1}_{\mathcal{S}_Y}(y) \quad (1.7)$$

*Dimostrazione.* Analoga a quella del Teorema 1.2.2. □

**Esempio 1.2.6.** (Distribuzione chi-quadrato) Sia  $Z \sim N(0, 1)$  e  $Y = h(Z) = Z^2$ , sicché  $h$  è monotona sui tratti  $A_0 = 0$ ,  $A_1 = (-\infty, 0)$ ,  $A_2 = (0, +\infty)$ . Consideriamo  $h_1(z) = z^2$  per  $z < 0$  mentre  $h_2(z) = z^2$  per  $z > 0$  e dunque  $h_1^{-1}(y) = -\sqrt{y}$  (NB:  $h_1^{-1}(y) \in A_1 \ \forall y \geq 0$ ) mentre  $h_2^{-1}(y) = \sqrt{y}$  (NB:  $h_2^{-1}(y) \in A_2 \ \forall y \geq 0$ ). Ora

$$\frac{d}{dy} h_1^{-1}(y) = -\frac{1}{2\sqrt{y}} \quad \text{e} \quad \frac{d}{dy} h_2^{-1}(y) = \frac{1}{2\sqrt{y}}$$

sono entrambe continue su  $\mathbb{R}^+$  e

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(-\sqrt{y})^2}{2}} \left| \frac{1}{2\sqrt{y}} \right| \mathbb{1}_{\mathbb{R}^+}(y) + \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{y})^2}{2}} \left| \frac{1}{2\sqrt{y}} \right| \mathbb{1}_{\mathbb{R}^+}(y) \\ &= \frac{1}{\sqrt{2\pi}y} e^{-(y/2)} \mathbb{1}_{\mathbb{R}^+}(y) \end{aligned}$$

Dalla forma della precedente funzione di densità si conclude subito che  $Y$  segue una distribuzione chi quadrato con  $\nu = 1$  gradi di libertà, caso particolare di una distribuzione Gamma di parametri  $\alpha = 1/2$  e  $\beta = 2$ .

In generale, consideriamo  $Z_1, Z_2, \dots, Z_n$  v.c. indipendenti e identicamente distribuite come  $N(0, 1)$  e sia

$$W_n = \sum_{i=1}^n Z_i^2$$

Allora per il Teorema 1.3.6, ovvero per la proprietà di riproducibilità, si trova immediatamente che

$$W_n = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

ovvero  $W_n$  ha una *distribuzione chi-quadrato* con  $\nu = n$  gradi di libertà che è un caso particolare di distribuzione Gamma di parametri  $\alpha = n/2$  e  $\beta = 2$ .

A scopo riassuntivo, la tavola 1.1 presenta alcune delle distribuzioni più comunemente utilizzate in statistica matematica e le reciproche relazioni di generazione siano esse *esatte* (linea continua) o *asintotiche* (linea tratteggiata). Avremo modo di tornare su quest'ultimo aspetto, in seguito e con maggior dettaglio.

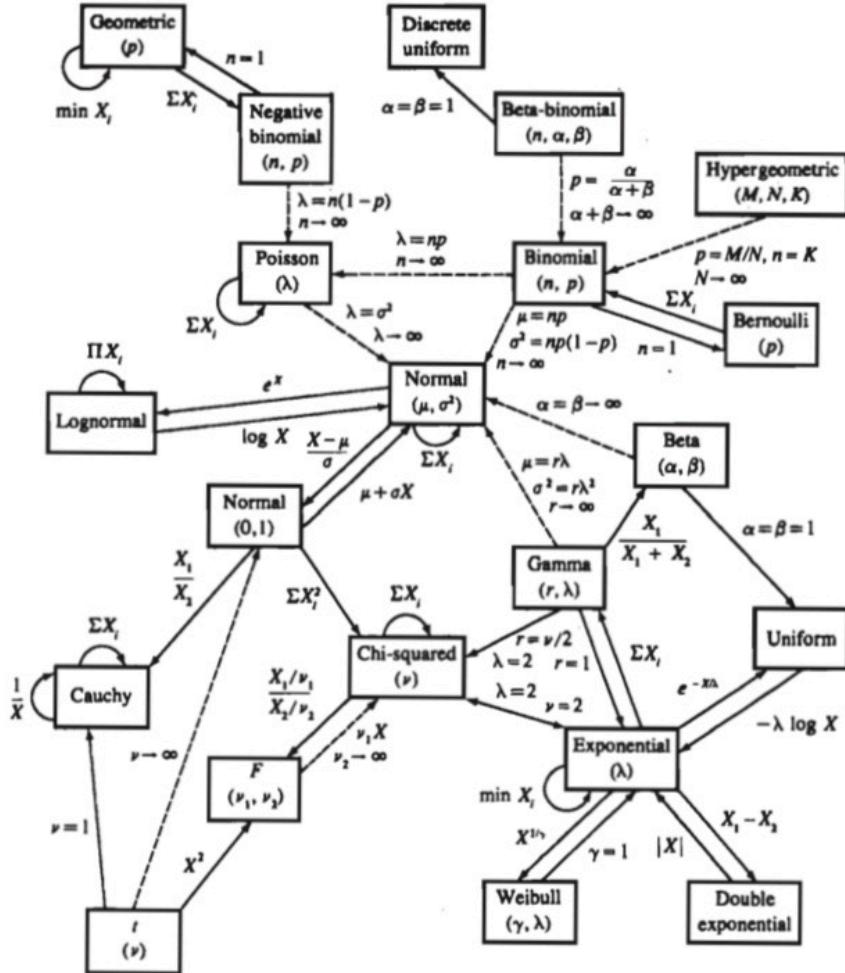


Figura 1.1: Tavola delle distribuzioni comunemente usate in statistica matematica

### 1.3 Valori attesi

Cominciamo col ricordare la definizione di *valore atteso*, o *valore di aspettazione* che dir si voglia, di una funzione  $g(X)$  della v.c.  $X$ .

**Definizione 1.3.1** (Valore atteso). Sia  $X$  una v.c. la cui distribuzione è indicizzata da un parametro (scalare o vettore)  $\theta$  e sia  $g(X)$  una sua funzione. Diremo *valore atteso* di  $g(X)$ , rispettivamente per variabili discrete e (assolutamente) continue, la

quantità

$$\mathbb{E}_\theta[g(X)] = \sum_x g(x) \cdot P_\theta(X = x) \mathbb{1}_A(x) \quad \text{o} \quad \mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) \cdot f(x, \theta) \mathbb{1}_A(x) dx \quad (1.8)$$

a condizione che somme e integrali esistano. Se ciò non è, diremo che il corrispondente valore atteso non esiste.

Molte delle proprietà di cui gode il valore atteso, e che sono spesso di aiuto nei calcoli, seguono da proprietà dell'integrale o della somma.

**Teorema 1.3.1.** Sia  $X$  una v.c. e siano  $a, b, c$  costanti. Allora per qualunque funzione  $g_1(x)$  e  $g_2(x)$  delle quali esistono i corrispondenti valorio attesi

- a) se  $c$  è una costante allora  $\mathbb{E}(c) = c$
- b)  $\mathbb{E}[a g_1(X) + b g_2(X) + c] = a \mathbb{E}[g_1(X)] + b \mathbb{E}[g_2(X)] + c$
- c) se  $g_1(x) \geq 0$  per ogni  $x$  allora  $\mathbb{E}[g_1(X)] \geq 0$
- d) se  $g_1(x) \geq g_2(x)$  per ogni  $x$  allora  $\mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)]$
- e) se  $a \leq g_1(x) \leq b$  per ogni  $x$  allora  $a \leq \mathbb{E}[g_1(X)] \leq b$ .

Quando si devono valutare valori attesi di *funzioni non lineari* della v.c.  $X$  - che supponiamo essere (assolutamente) continua - possiamo procedere in due diversi modi:

1.  $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x; \theta) \mathbb{1}_{S_X}(x) dx$
2. trovata la funzione di densità  $f_Y(y; \theta)$  di  $Y = g(X)$  si calcola

$$\mathbb{E}[g(X)] = \mathbb{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y; \theta) \mathbb{1}_{S_Y}(y) dy \quad (1.9)$$

**Esempio 1.3.1.** Sia  $X \sim \mathcal{U}(0, 1)$  e sia  $Y = g(X) = -\ln(X)$ . Via teorema di trasformazione è facile trovare che  $Y = g(X) = -\ln(X)$  segue una distribuzione esponenziale di parametro  $\beta = 1$  sicché si ha immediatamente che  $\mathbb{E}[-\ln(X)] = 1$ .

**Nota.** Si può procedere in maniera del tutto analoga nel caso di  $X$  v.c. discreta, sostituendo le funzioni di densità con le funzioni di massa e gli integrali con opportune somme.

Veniamo ora a un'importante classe di valori attesi nota come *classe dei momenti di ordine s. centrati in a*

**Definizione 1.3.2** (Momento di ordine  $s$  centrato in  $a$ ). Data una funzione del tipo

$$g(X) = (X - a)^s \quad s \in \mathbb{N} \quad (1.10)$$

il suo valore atteso  $\mathbb{E}[(X - a)^s]$ , laddove esista, è definito *momento di ordine s centrato in a* e si indica con  $\mu_s(a)$ .

Risulta significativo soffermarsi su due particolari casi limite della precedente:

- i) se  $a = 0$ , si parla di *momento centrato nell'origine di ordine s* o di *momento non centrato*, e si indica con le seguenti notazioni equivalenti  $\mathbb{E}(X^s) = \mu_s(0) = \mu'_s$
- ii) se  $a = \mathbb{E}(X)$  si parla di *momento centrato nella media di ordine s* e lo si indica con le notazioni equivalenti  $\mu_s(\mathbb{E}(X)) = \mu_s(\mu'_1) = \mathbb{E}(X - \mu'_1)^s = \mu_s$

Si noti che non tutte le distribuzioni ammettono momenti; un esempio di distribuzione che non ammette momenti è la distribuzione di Cauchy.

Il calcolo dei momenti in statistica matematica può risultare particolarmente significativo in quanto alcuni di essi forniscono delle informazioni in merito alla distribuzione ad essi associata. In particolare,  $\mu'_1$  coincide evidentemente con il valor medio, e  $\mu_2 = \mathbb{E}(X - \mu'_1)^2$  coincide con la definizione di varianza. Si dice anche che  $\mu'_1$  fornisce informazioni sul *baricentro* della distribuzione,  $\mu_2$  sulla sua inerzia. Dal momento di ordine 3 si ricavano informazioni sulla simmetria della distribuzione; infine, da  $\mu_4$  è possibile ottenere alcune informazioni sulle code della distribuzione, ovvero sulla sua *curtosi*: una distribuzione che non presenta una pendenza particolarmente elevata in prossimità delle code è detta platicurtica, una distribuzione che prima delle code abbia una pendenza significativa è detta leptocurtica. Nei casi intermedi, si parla di distribuzione normocurtica.

**Definizione 1.3.3** (Funzione generatrice dei momenti). Sia  $X$  una variabile casuale (discreta o assolutamente continua). Se esiste  $t_0 > 0$  tale per cui  $\mathbb{E}(e^{tX}) < +\infty \forall t \in (-t_0, t_0)$ , chiameremo la funzione

$$M_X(t) := \mathbb{E}(e^{tX}) \quad (1.11)$$

*funzione generatrice dei momenti*, o semplicemente *fgm*, di  $X$ .

**Esempio 1.3.2.** E' noto che

$$\sum_{x=1}^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6} \quad (1.12)$$

Allora

$$\begin{aligned} P(X = x) &= \begin{cases} \frac{6}{\pi^2} \frac{1}{x^2}, & x = 1, 2, 3, \dots \\ 0 & \text{altrimenti} \end{cases} \\ &= \frac{6}{\pi^2 x^2} \mathbb{1}_{\mathbb{N}}(x) \end{aligned} \quad (1.13)$$

è una funzione di massa di una v.c. discreta  $X$ . La corrispondente *fgm*, se esiste, è per definizione data da

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x=1}^{\infty} e^{tx} P(X = x) = \sum_{x=1}^{\infty} \frac{6 e^{tx}}{\pi^2 x^2} \quad (1.14)$$

Ma ci rendiamo immediatamente conto che la precedente serie *diverge* se  $t > 0$ ; infatti basta ricorrere al test del rapporto, ponendo  $a_x = \frac{6 e^{tx}}{\pi^2 x^2}$ , per cui

$$\lim_{x \rightarrow \infty} \frac{a_{x+1}}{a_x} = e^t \left( \frac{x}{x+1} \right)^2 = e^t > 1 \quad (1.15)$$

e di conseguenza, non esiste un numero positivo  $t_0$  tale che  $M_X(t)$  esista per  $-t_0 < t < t_0$ . Possiamo allora concludere che la distribuzione di probabilità che ha come funzione di massa (1.13) non ammette funzione generatrice dei momenti.

**Esempio 1.3.3** (Funzione generatrice dei momenti di una v.c. Normale). Sia  $Z \sim N(0, 1)$  e allora

$$M_Z(t) = \mathbb{E}(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (1.16)$$

Ora

$$e^{tz} e^{-z^2/2} = e^{-\frac{1}{2}z^2 + tz} = e^{-\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2} \quad (1.17)$$

sicché

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2} dz \\ &= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz \end{aligned} \quad (1.18)$$

Ma l'integrandà di cui sopra corrisponde alla densità di una v.c.  $N(\mu = t, \sigma^2 = 1)$  per cui l'integrale sulla retta reale vale 1; di conseguenza,

$$M_Z(t) = e^{t^2/2}, \quad t > 0 \quad (1.19)$$

è la *fgm* di una v.c. Normale standard. Poniamo ora  $X = \sigma Z + \mu$  con  $\sigma > 0$  e  $\mu \in \mathbb{R}$ . Allora

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) = \mathbb{E}(e^{t(\sigma Z + \mu)}) = \mathbb{E}(e^{t\sigma Z}) \mathbb{E}(e^{\mu t}) \\ &= e^{(t^2\sigma^2)/2} e^{t\mu} \\ &= e^{t\mu + \frac{1}{2}t^2\sigma^2} \end{aligned} \quad (1.20)$$

è la *fgm* di una v.c. Normale di media  $\mu$  e varianza  $\sigma^2$ .

Val la pena riflettere sul seguente risultato

**Teorema 1.3.2.** La *fgm* di una v.c. X (discreta o continua che sia) esiste solo se esistono i momenti di qualsiasi ordine

*Dimostrazione.*

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) = \mathbb{E}\left(\sum_{s=0}^{\infty} \frac{t^s}{s!} X^s\right) \\ &= \sum_{s=0}^{\infty} \frac{t^s}{s!} \mathbb{E}(X^s) \\ &= \sum_{s=0}^{\infty} \frac{t^s}{s!} \mu'_s \end{aligned} \quad (1.21)$$

e dunque l'esistenza dei momenti di qualsiasi ordine  $s$  (ergo la loro finitezza) comporta l'esistenza della associata *fgm* essendo la quantità  $\frac{t^s}{s!}$  finita per ogni  $s \in \mathbb{N}$ .  $\square$

Nel paragrafo successivo si avrà ragione del nome di *generatrice dei momenti* dato alla funzione  $M_X(t)$ .

**Nota.**

Data una funzione del tipo  $g(X) = t^X$ , la funzione  $P(t) = E(t^X)$  è nota come *funzione generatrice della probabilità*.

Riportiamo di seguito due importanti risultati relativi a identità fondamentali a proposito di valori attesi che giocheranno un ruolo strategico nell'individuazioni di procedure di costruzione di strumenti inferenziali i ottimali.

**Teorema 1.3.3** (Valore atteso iterato). Siano  $X$  e  $Y$  due v.c. (discrete o continue). Allora

$$\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X | Y)]$$

*Dimostrazione.* Cominciamo con osservare che

- a)  $\mathbb{E}(X)$  è una costante, funzione del parametro che indicizza la distribuzione di  $X$
- b)  $\mathbb{E}(X | Y = y)$  è una funzione deterministica (=non casuale) di  $y$
- c)  $\mathbb{E}(X | Y)$  è una funzione della v.c.  $Y$

Ora, nel caso in cui le v.c.  $X$  e  $Y$  siano continue e omettendo, per semplicità di notazione, la dipendenza delle densità coinvolte dal parametro che indicizza la famiglia di distribuzioni, possiamo scrivere

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X | Y)] &= \int \mathbb{E}(X | Y) f_Y(y) dy \\ &= \int \left[ \int x f_{X|Y}(x, y) dx \right] f_Y(y) dy \\ &= \int \int x f_{X|Y}(x, y) f_Y(y) dx dy \\ &= \int \int x f_{X,Y}(x, y) dy dx \\ &= \int x \left[ \int f_{X,Y}(x, y) dy \right] dx \\ &= \int x f_X(x) dx \\ &= \mathbb{E}(X). \end{aligned}$$

□

**NB:** laddove  $X$  e  $Y$  fossero v.c. discrete, gli integrali andranno sostituiti con opportune somme o serie.

**Teorema 1.3.4** (Scomposizione della varianza). Siano  $X$  e  $Y$  due v.c. (discrete o continue). Allora

$$\mathbb{V}ar(X) = \mathbb{V}ar[\mathbb{E}(X | Y)] + \mathbb{E}[\mathbb{V}ar(X | Y)]$$

*Dimostrazione.* La varianza di  $E(X | Y)$  è per definizione data da

$$\begin{aligned}\mathbb{V}ar[\mathbb{E}(X | Y)] &= \mathbb{E}\left[\mathbb{E}(X | Y) - \mathbb{E}[\mathbb{E}(X | Y)]\right]^2 \\ &= \mathbb{E}\left[[\mathbb{E}(X | Y)]^2\right] - \left[\mathbb{E}[\mathbb{E}(X | Y)]\right]^2 \\ &= \mathbb{E}\left[[\mathbb{E}(X | Y)]^2\right] - [\mathbb{E}(X)]^2\end{aligned}\tag{1.22}$$

poichè, per quanto visto in Teorema 1,  $\mathbb{E}[\mathbb{E}(X | Y)] = \mathbb{E}(X)$ .

Ora, sempre per definizione, la varianza di  $X$  condizionatamente a  $Y$  è data da

$$\mathbb{V}ar(X | Y) = \mathbb{E}(X^2 | Y) - [\mathbb{E}(X | Y)]^2$$

sicchè

$$[\mathbb{E}(X | Y)]^2 = \mathbb{E}(X^2 | Y) - \mathbb{V}ar(X | Y)$$

e di conseguenza, prendendo il valore atteso su ambo i membri della precedente identità, si ha

$$\mathbb{E}[[\mathbb{E}(X | Y)]^2] = \mathbb{E}[\mathbb{E}(X^2 | Y)] - \mathbb{E}[\mathbb{V}ar(X | Y)]$$

Ma allora, sostituendo la precedente in (1.22) si ha

$$\begin{aligned}\mathbb{V}ar[\mathbb{E}(X | Y)] &= \left\{ \mathbb{E}[\mathbb{E}(X^2 | Y)] - \mathbb{E}[\mathbb{V}ar(X | Y)] \right\} - [\mathbb{E}(X)]^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}[\mathbb{V}ar(X | Y)] - [\mathbb{E}(X)]^2 \\ &= \{\mathbb{E}(X^2) - [\mathbb{E}(X)]^2\} - \mathbb{E}[\mathbb{V}ar(X | Y)] \\ &= \mathbb{V}ar(X) - \mathbb{E}[\mathbb{V}ar(X | Y)]\end{aligned}$$

e da quest'ultima identità segue immediatamente la tesi

$$\mathbb{V}ar(X) = \mathbb{V}ar[\mathbb{E}(X | Y)] + \mathbb{E}[\mathbb{V}ar(X | Y)].$$

□

### 1.3.1 Funzione generatrice dei momenti e momenti di una distribuzione

I momenti rappresentano delle costanti caratteristiche riferite alla distribuzione associata alla variabile casuale (o al vettore casuale)  $X$ ; caratteristiche nel senso che descrivono, caratterizzando, la distribuzione stessa fornendo informazioni su elementi importanti quali centro, dispersione e forma (e su molti altri aspetti, anche se di minor interesse in statistica matematica).

Dal momento che una distribuzione  $F_X(x; \theta)$  che ammette fgm  $M_X(t)$  ne è completamente determinata, non deve sorprendere troppo il fatto che si possono ottenere importanti elementi caratterizzanti la distribuzione  $F_X(x; \theta)$ , quali sono i momenti (centrati e non) direttamente da  $M_X(t)$ .

**Definizione 1.3.4.** Se una variabile casuale  $fgm$  derivabile infinite volte in un intorno di  $t = 0$  e se tutti i suoi momenti sono finiti, allora definiamo il momento di ordine  $s$  non centrato:

$$\mu'_s := \mathbb{E}(X^s) = \frac{d^s}{dt^s} M_X(t) |_{t=0} \quad (1.23)$$

La relazione (1.23) è facilmente verificabile nel caso generale:

$$\begin{aligned} \frac{d^s}{dt^s} M_X(t) |_{t=0} &= \frac{d^s}{dt^s} \mathbb{E}(e^{tX}) |_{t=0} = \frac{d^s}{dt^s} \mathbb{E} \left( \sum_{s=0}^{\infty} \frac{t^s}{s!} X^s \right) |_{t=0} \\ &= \frac{d^s}{dt^s} \sum_{s=0}^{\infty} \frac{t^s}{s!} \mathbb{E}(X^s) |_{t=0} \\ &= \frac{d^s}{dt^s} \sum_{s=0}^{\infty} \frac{t^s}{s!} \mu'_s |_{t=0} = \mu'_s \end{aligned}$$

Esiste una semplice relazione che lega momenti non centrati e momenti centrati come si può evincere dal seguente teorema.

**Teorema 1.** Sia  $X$  una v.c. che ammette momenti non centrati  $\mu'_m$  almeno fino all'ordine  $r \geq s$ . Allora,

$$\mu_s = \mathbb{E}[(X - \mu'_1)^s] = \sum_{m=0}^s (-1)^m \binom{s}{m} \mu'_{s-m} (\mu'_1)^m \quad (1.24)$$

*Dimostrazione.* La dimostrazione segue immediatamente dallo sviluppo della potenza  $s$ -ma del binomio e dalle proprietà dell'operatore valore atteso.  $\square$

Di seguito alcuni esempi.

**Esempio 1.3.4** (Distribuzione di Bernoulli). Consideriamo una variabile casuale  $X \sim b(1, p)$  e calcoliamone la funzione generatrice dei momenti:

$$M_X(t) = E_p(e^{tX}) = \sum_{x=0}^1 e^{tx} p^x (1-p)^{1-x} = 1 - p + pe^t \quad (1.25)$$

Nota  $M_X(t)$ , possiamo inoltre calcolare i momenti non centrati di primo e secondo ordine:

$$\mu'_1 = \frac{d}{dt} M_X(t) \Big|_{t=0} = pe^t \Big|_{t=0} = p \quad \mu'_2 = \frac{d^2}{dt^2} M_X(t) \Big|_{t=0} = pe^t \Big|_{t=0} = p \quad (1.26)$$

Concludiamo dunque che il valor medio e la varianza relativi a  $X$  sono rispettivamente:  $E(X) = p$  e  $\text{Var}(X) = \mu'_2 - (\mu'_1)^2 = p(1-p)$ .

**Esempio 1.3.5** (Distribuzione Gamma). Consideriamo una variabile casuale  $X \sim \mathcal{G}(\alpha, \beta)$  e calcoliamone la funzione generatrice dei momenti:

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^{+\infty} e^{tx} x^{\alpha-1} e^{-\frac{x}{\beta}} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^{+\infty} x^{\alpha-1} e^{(t-\frac{1}{\beta})x} dx \quad \text{sost.: } v = \left(\frac{1}{\beta} - t\right)x \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\frac{\beta}{1-\beta t}\right)^\alpha \int_0^{+\infty} e^{-v} v^{\alpha-1} dv \\ &= \frac{1}{(1-\beta t)^\alpha}, \quad \text{con } t < \frac{1}{\beta} \end{aligned} \quad (1.27)$$

dove al penultimo passaggio notiamo che l'integrale restituisce la funzione Gamma valutata in  $\alpha$ ,  $\Gamma(\alpha)$  che elide con la medesima quantità al denominatore della costante di normalizzazione  $\frac{1}{\Gamma(\alpha)\beta^\alpha}$ .

Calcoliamo ora i *momenti non centrati* di primo e secondo ordine:

$$\begin{aligned} \mu'_1 &= \frac{d}{dt} M_X(t) \Big|_{t=0} = \frac{d}{dt} \frac{1}{(1-\beta t)^\alpha} \Big|_{t=0} = \frac{\alpha\beta}{(1-\beta t)^{\alpha+1}} \Big|_{t=0} = \alpha\beta \\ \mu'_2 &= \frac{d^2}{dt^2} M_X(t) \Big|_{t=0} = \frac{\alpha(\alpha+1)\beta^2}{(1-\beta t)^{\alpha+2}} \Big|_{t=0} = \alpha^2\beta^2 + \alpha\beta^2 \end{aligned} \quad (1.28)$$

Concludiamo dunque che il valor medio e la varianza relativi a  $X$  sono rispettivamente  $E(X) = \alpha\beta$  e  $\text{Var}(X) = \alpha\beta^2$ .

Torniamo alle funzioni generatrici dei momento con due interessanti e, soprattutto assai utili, teoremi:

- i) il primo riguarda la caratterizzazione della distribuzione a opera della funzione generatrice dei momenti. In altre parole, se la *fgm* esiste (il che equivale ad assumere l'esistenza dei momenti di qualsiasi ordine) allora la successione dei momenti identifica univocamente la distribuzione, vale a dire, il *problema dei momenti è determinato*: alla successione dei momenti rimane associata *una e una sola* distribuzione sicché dalla conoscenza dei momenti (ergo della *fgm*) si può risalire alla distribuzione a essi associata;
- ii) il secondo un'importante applicazione della *fgm* per ricavare la distribuzione di particolari trasformazioni di variabili casuali (siano esse *indipendenti* o *indipendenti e identicamente distribuite*) quali sono le somme di variabili casuali: esse rappresentano una classe di trasformazioni assai ricorrenti in statistica matematica, essendo un naturale strumento di sintesi dell'infomazione contenuta in una  $n$ -pa di v.c.

**Teorema 1.3.5.** Siano  $X$  e  $Y$  due variabili casuali e siano  $M_X(t)$  e  $M_Y(t)$  le rispettive funzioni generatrici dei momenti, che esistono in un intorno  $(-t_0, t_0)$  di 0. Allora  $F_X(z) = F_Y(z)$  se e solo se  $M_X(t) = M_Y(t) \forall t \in (-t_0, t_0)$  con  $t_0 > 0$ .

*Dimostrazione.* Per esercizio. □

**Osservazione 1.** Il teorema appena visto ci dice sostanzialmente che, se esiste, la fgm caratterizza la distribuzione della corrispondente v.c.

Dunque, vale il seguente

**Corollario 1.3.1.** Se la funzione generatrice dei momenti di una distribuzione esiste allora essa è unica e quindi caratterizza univocamente la distribuzione.

Il teorema che segue stabilisce il comportamento della fgm su somme di variabili casuali indipendenti.

**Teorema 1.3.6.** Siano  $X_1, \dots, X_n$  variabili casuali indipendenti e siano

$$M_{X_1}(t), \dots, M_{X_n}(t)$$

le rispettive funzioni generatrici dei momenti. Allora la funzione generatrice di  $W_n = h(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i$  è la seguente:

$$M_{W_n}(t) = \mathbb{E}(e^{tW_n}) = \mathbb{E}\left(\exp\left\{t \sum_{i=1}^n X_i\right\}\right) = \prod_{i=1}^n \mathbb{E}(e^{tX_i}) = \prod_{i=1}^n M_{X_i}(t) \quad (1.29)$$

*Dimostrazione.* Per esercizio. □

In altre parole, la funzione generatrice dei momenti di una somma di variabili aleatorie indipendenti è pari al prodotto delle funzioni generatrici dei momenti delle variabili aleatorie stesse. Se oltre che indipendenti le v.c. fossero anche identicamente distribuite, la fgm della loro somma coinciderebbe con la potenza  $n$ -ma della fgm di una qualsiasi di esse, ossia

$$M_Y(t) = [M_{X_1}(t)]^n.$$

Il Teorema 1.3.6, unitamente al Teorema 1.3.5 e al Corollario 1.3.1, dà ragione dell'importanza della fgm non solo quale strumento per generare i momenti di una distribuzione ma anche nell'ambito di particolari trasformazioni di variabili casuali, quantomeno indipendenti,  $(X_1, X_2, \dots, X_n)$  quali sono le somme e del calcolo della relativa distribuzione. L'esempio che segue ne è una prova.

**Esempio 1.3.6.** Siano  $(X_1, \dots, X_n)$  risultati della replicazione di un esperimento casuale dicotomico  $(X_i \sim b(1, p))$ . Vogliamo trovare la distribuzione di  $W_n := \sum_{i=1}^n X_i$ . Calcoliamo quindi la sua fgm: per il Teorema 1.3.6

$$\begin{aligned} M_{W_n}(t) &= \mathbb{E}(e^{tW_n}) = \mathbb{E}\left(e^{t \sum_{i=1}^n X_i}\right) \\ &= \prod_{i=1}^n \mathbb{E}(e^{tX_i}) \\ &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n (pe^t + (1-p)) = (pe^t + (1-p))^n \end{aligned}$$

ovvero  $W_n$  è distribuita come  $b(n, p)$  per il Corollario 1.3.1.

**Esempio 1.3.7** (Distribuzione binomiale). Sia  $X_1, \dots, X_n$  un set di  $n$  variabili casuali indipendenti e identicamente distribuite con distribuzione di Bernoulli  $b(1, p)$ . Allora la variabile casuale  $W_n = \sum_{i=1}^n X_i$  ha funzione generatrice dei momenti

$$M_{W_n}(t) = (1 - p + pe^t)^n \quad (1.30)$$

Si noti, in virtù della relazione biunivoca tra (forma della) *fgm* e distribuzione, che  $W_n \sim b(n, p)$  poiché  $M_{W_n}(t)$  è la funzione generatrice dei momenti della distribuzione binomiale. Calcolando tale funzione applicando direttamente la sua definizione avremmo ottenuto esattamente lo stesso risultato; infatti,

$$M_{W_n}(t) = \mathbb{E}_p(e^{tW_n}) = \sum_{w=0}^n e^{tW_n} \binom{n}{w} p^w (1-p)^{n-w} = (1 - p + pe^t)^n \quad (1.31)$$

**Esempio 1.3.8** (Distribuzione di Poisson). Sia  $X_1, \dots, X_n$  un insieme di  $n$  variabili casuali indipendenti distribuite con distribuzioni di Poisson  $X_i \sim P(\lambda_i)$ , con  $i = 1, \dots, n$ . Determiniamo la funzione generatrice dei momenti associata alla generica variabile casuale di Poisson  $X_i$ :

$$M_{X_i}(t) = E_{\lambda_i}(e^{tX_i}) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda_i} \lambda_i^x}{x!} = e^{-\lambda_i} \sum_{x=0}^{\infty} \frac{(\lambda_i e^t)^x}{x!} = \exp\{\lambda_i(e^t - 1)\} \quad (1.32)$$

Allora la variabile casuale  $W_n = \sum_{i=1}^n X_i$  ha funzione generatrice dei momenti

$$M_{W_n}(t) = \exp\{\lambda_1(e^t - 1)\} \dots \exp\{\lambda_n(e^t - 1)\} = \exp\{(\lambda_1 + \dots + \lambda_n)(e^t - 1)\} \quad (1.33)$$

Pertanto anche  $Y$  segue una distribuzione di Poisson di parametro  $\lambda$  pari alla somma dei parametri delle variabili che la costituiscono gli addendi della somma, sicché

$$W_n \sim \mathcal{P}\left(\sum_{i=1}^n \lambda_i\right). \quad (1.34)$$

**Esempio 1.3.9** (Distribuzione chi-quadrato). Mediante il teorema 1.3.6, le considerazioni effettuate precedentemente in merito alla distribuzione chi-quadrato possono essere estese a un insieme di  $n$  variabili casuali indipendenti e identicamente distribuite  $X_1, \dots, X_n$ , con  $X_i \sim N(0, 1)$ ,  $\forall i = 1, \dots, n$ . Definita la funzione

$$W_n = \sum_{i=1}^n X_i^2 \quad (1.35)$$

abbiamo già mostrato che ogni termine  $X_i^2 \sim \mathcal{G}(1/2, 2)$ . Possiamo applicare il teorema 1.3.6 per determinare la funzione generatrice dei momenti di  $W_n$ , nota la funzione generatrice dei momenti di una generica variabile con distribuzione  $\mathcal{G}(\alpha = 1/2, \beta = 2)$ . Otteniamo dunque

$$M_{W_n}(t) = \prod_{i=1}^n M_{X_i^2}(t) = \prod_{i=1}^n \left(\frac{1}{1-2t}\right)^{1/2} = \left(\frac{1}{1-2t}\right)^{n/2} \quad (1.36)$$

Concludiamo dunque che  $W_n \sim \mathcal{G}(n/2, 2)$ . Tale distribuzione è nota come *distribuzione chi-quadrato con  $n$  gradi di libertà*.

Gli esempi che abbiamo appena visto costituiscono un buon esempio di quella che è nota come *proprietà di riproducibilità*.

**Definizione 1.3.5** (Proprietà di riproducibilità). Siano  $(X_1, X_2, \dots, X_n)$   $n$  v.c. indipendenti e  $X_i \sim F_{X_i}(x; \theta_i)$ ,  $\theta_i \in \Theta$ . Diremo la famiglia di distribuzioni corrispondente a  $F_{X_i}$  godere della *proprietà di riproducibilità* se

$$X = \sum_{i=1}^n X_i \sim F_X(x; \theta), \text{ con } \theta = \sum_{i=1}^n \theta_i. \quad (1.37)$$

Ora, considerato un insieme di v.c. indipendenti  $(X_1, X_2, \dots, X_n)$  proveniente da una famiglia di distribuzioni quali, per esempio, Normale, Gamma (e casi particolari quali esponenziale e chi-quadrato), Poisson, Geometrica, binomiale per  $p$  fissato e comune alle  $n$  v.c. sono *chiuse sotto somma* ovvero la distribuzione associata alla somma  $X = \sum_{i=1}^n X_i$  delle v.c. indipendenti in questione *riproduce* la distribuzione da cui le v.c. provengono opportunamente riparametrizzata. La dimostrazione di questo fatto poggia sulla unicità della *fgm* e sulla sua capacità di identificare la distribuzione a essa associata.

## 1.4 Famiglie esponenziali

**Definizione 1.4.1** (Famiglia esponenziale). Una famiglia di distribuzioni discrete o (assolutamente) continue è detta *famiglia esponenziale* a  $k$  parametri  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$  se la corrispondente funzione di massa o di densità può essere scritta nella forma seguente:

$$f_X(x; \boldsymbol{\theta}) = C(x) D(\boldsymbol{\theta}) \exp \left\{ \sum_{m=1}^k A_m(\boldsymbol{\theta}) B_m(x) \right\} \mathbb{1}_{S_X}(x) \quad (1.38)$$

dove  $A_m(\boldsymbol{\theta})$  e  $D(\boldsymbol{\theta})$  sono funzioni esclusivamente di  $\boldsymbol{\theta}$  mentre  $B_m(x)$  e  $C(x)$  sono funzioni della sola  $x$ .

L'appartenenza di una distribuzione a famiglia esponenziale è talvolta espressa tramite un'interessante *riparametrizzazione* della sua funzione di massa o di densità:

**Definizione 1.4.2** (Parametrizzazione naturale). Posto  $\eta_m = A_m(\boldsymbol{\theta})$  e riscritta la densità (8.2.2) come

$$f(x; \boldsymbol{\eta}) = C(x) \tilde{D}(\boldsymbol{\eta}) \exp \left\{ \sum_{m=1}^k \eta_m B_m(x) \right\} \mathbb{1}_{S_X}(x)$$

si dice che la famiglia esponenziale è ora espressa nella sua *parametrizzazione naturale* con  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)$  *parametro naturale*.

Vale la pena osservare che:

- talvolta l'appartenenza a famiglia esponenziale a  $k$  parametri è definita riscrivendo la (8.2.2) come

$$f_X(x; \boldsymbol{\theta}) = \exp \left\{ \sum_{m=1}^k A_m(\boldsymbol{\theta}) B_m(x) + C^*(x) + D^*(\boldsymbol{\theta}) \right\} \mathbb{1}_{S_X}(x) \quad (1.39)$$

con  $C^*(x) = \ln[C(x)]$  e  $D^*(\boldsymbol{\theta}) = \ln[D(\boldsymbol{\theta})]$

- se una distribuzione appartiene a una famiglia esponenziale il suo supporto  $\mathcal{S}_X$  non può dipendere dal parametro  $\theta$ : se così non fosse, non avremmo modo di scrivere le funzioni  $C$  e  $D$  secondo quanto richiesto dalla definizione
- le famiglie di esponenziali hanno interessanti proprietà matematiche (*proprietà di regolarità*). Dal punto di vista statistico, ciò si traduce in un'interessante conseguenza: tutta l'informazione contenuta nei dati a disposizione  $(X_1, \dots, X_n)$  relativa alla distribuzione della v.c.  $X$  può essere riassunta attraverso  $k$  quantità

$$\left( \sum_{i=1}^n B_1(X_i), \sum_{i=1}^n B_2(X_i), \dots, \sum_{i=1}^n B_k(X_i) \right) \quad (1.40)$$

funzioni di  $(X_1, \dots, X_n)$  che potranno essere impiegate per costruire procedure inferenziali (stima, test per la verifica di ipotesi) riguardanti il parametro  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ . In altre parole, l'appartenenza ad una famiglia esponenziale permette una *riduzione nella dimensione* dei dati  $(X_1, X_2, \dots, X_n)$  tramite le  $k$  quantità in (1.40) anche dette *statistiche naturali*

**Esempio 1.4.1** (Distribuzione bernoulliana). Consideriamo un vettore di  $n$  variabili casuali indipendenti e identicamente distribuite  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  con  $X_i \sim b(1, \theta)$  per ogni  $i = 1, \dots, n$ . La probabilità congiunta associata a una specifica osservazione  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  è la seguente:

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}) &= \prod_{i=1}^n P(X_i = x_i) \mathbb{1}_{\{0,1\}}(x_i) \\ &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i) \\ &= \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(x_i) \\ &= \underbrace{\prod_{i=1}^n \mathbb{1}_{\{0,1\}}(x_i)}_{C(x)} \underbrace{(1 - \theta)^n}_{D(\theta)} \exp \left\{ \underbrace{(\ln \theta - \ln(1 - \theta))}_{A(\theta)} \underbrace{\sum_{i=1}^n x_i}_{\sum_{i=1}^n B(x_i)} \right\} \end{aligned}$$

La funzione  $A(\theta)$  corrispondente alla parametrizzazione necessaria a rendere esplicita l'appartenenza della famiglia di distribuzioni a famiglia esponenziale a  $k = 1$  parametri restituisce il *parametro naturale* che, in questo caso, corrisponde al logaritmo dell'odds ratio,  $A(\theta) = \ln \frac{\theta}{1-\theta}$  intendendo per *odds ratio* il rapporto tra la probabilità di un evento (successo, nel caso di Bernoulli che stiamo trattando) e la probabilità del suo complementare (insuccesso).

La quantità  $\sum_{i=1}^n B(X_i) = \sum_{i=1}^n X_i$ , invece, corrisponde al numero di successi in  $n$  prove bernoulliane ed è *naturale* pensare a questa quantità quale sintesi dell'informazione relativa al parametro  $p$  (probabilità di successo) contenuta nei dati  $(x_1, x_2, \dots, x_n)$ ; in tal senso essa costituisce la *statistica naturale* per il problema in questione. Ed è altrettanto *naturale* osservare che la *frequenza relativa di successo*

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

che della *statistica naturale* è una funzione, possa costituire il *naturale* punto di partenza su cui costruire le diverse procedure inferenziali (stima, test, intervalli di confidenza) relative al parametro  $\theta$ .

Quanto detto in questo esempio particolare ha comunque valenza generale e si può estendere a molti altri contesti distributivi. Riflettete in tal senso negli esempi che seguono individuando *parametri naturali* e *statistiche naturali*.

**Esempio 1.4.2** (Distribuzione binomiale). Consideriamo una variabile casuale  $X \sim b(n, \theta)$ . Con qualche semplice passaggio si può scrivere la funzione di massa della distribuzione binomiale come

$$\begin{aligned} P_\theta(X = x) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \mathbb{1}_{\{0, \dots, n\}}(x) \\ &= \underbrace{\binom{n}{x} \mathbb{1}_B(x)}_{C(x)} \underbrace{(1 - \theta)^n}_{D(\theta)} \exp \left\{ \underbrace{\ln \left( \frac{\theta}{1 - \theta} \right)}_{A(\theta)} \underbrace{x}_{B(x)} \right\} \end{aligned}$$

da cui si evince l'appartenenza della distribuzione binomiale alla classe delle famiglie esponenziali a  $k = 1$ -parametri. Cosa potete dire in questo caso in merito a *parametri* e *statistiche naturali*?

**Esempio 1.4.3** (Distribuzione Normale). Consideriamo un campione casuale di ampiezza  $n$  proveniente da una distribuzione Normale  $N(\mu, \sigma^2)$ . Allora, la densità congiunta associata a una specifica osservazione campionaria  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  è la seguente:

$$\begin{aligned} f(\mathbf{x}; \mu, \sigma^2) &= \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \mathbb{1}_{\mathbb{R}}(x_i) = \\ &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \mathbb{1}_{\mathbb{R}}(x_i) = \\ &= \underbrace{\prod_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_i)}_{C(x_1, x_2, \dots, x_n)} \underbrace{\frac{\exp\{-n\mu^2/2\sigma^2\}}{(\sqrt{2\pi\sigma^2})^n}}_{D(\mu, \sigma^2)} \exp \left\{ -\underbrace{\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2}_{A_1(\mu, \sigma^2)} + \underbrace{\frac{\mu}{\sigma^2} \sum_{i=1}^n x_i}_{A_2(\mu, \sigma^2)} \right\} \end{aligned}$$

Come si può facilmente notare, oltre all'appartenenza della distribuzione Normale a famiglia esponenziale a  $k = 2$ -parametri, la precedente scrittura permette di osservare che la *statistica naturale*

$$\left( \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i \right)$$

consente di riassumere tutta l'informazione relativa al vettore dei parametri  $\theta = (\mu, \sigma^2)$  contenuta nei dati a disposizione.

**Esempio 1.4.4.** Sia  $X$  una v.c. avente densità

$$f_X(x, \theta) = \frac{e^{1-\frac{x}{\theta}}}{\theta} \mathbb{1}_{(\theta, \infty)}(x), \theta > 0. \quad (1.41)$$

Come si può immediatamente notare, la distribuzione di  $X$  non appartiene a famiglia esponenziale. Infatti, il fatto che il supporto  $\mathcal{S}_X = (\theta, \infty)$  di  $X$  dipenda dal parametro  $\theta$  non permette a  $f_X$  di soddisfare quanto richiesto in termini di forma funzionale al fine dell'appartenenza a famiglia esponenziale!

Il teorema che segue contiene un interessante risultato relativo al calcolo del valore atteso delle funzioni  $B_m(X)$ ,  $m = 1, 2, \dots, k$ .

**Teorema 1.4.1.** Se la famiglia esponenziale è espressa nella sua parametrizzazione naturale, vale il seguente risultato:

$$E(B_m(X)) = -\frac{d}{d\eta_m} \ln(\tilde{D}(\boldsymbol{\eta})) \quad (1.42)$$

*Dimostrazione.* Ricordando che in una famiglia esponenziale le operazioni di derivazione e di integrazione si possono scambiare nell'ordine, osserviamo subito che vale la seguente uguaglianza, facilmente verificabile per calcolo diretto:

$$\mathbb{E} \left[ \frac{d}{d\eta_m} \ln f_X(x; \boldsymbol{\eta}) \right] = \frac{d}{d\eta_m} \ln[\tilde{D}(\boldsymbol{\eta})] + \mathbb{E}[B_m(X)] \quad (1.43)$$

Ora, applicando direttamente la definizione di valore atteso si ha

$$\begin{aligned} \mathbb{E} \left[ \frac{d}{d\eta_m} \ln f_X(x; \boldsymbol{\eta}) \right] &= \int_{-\infty}^{+\infty} \frac{d}{d\eta_m} [\ln f_X(x; \boldsymbol{\eta})] f_X(x; \boldsymbol{\eta}) dx \\ &= \int_{-\infty}^{+\infty} \frac{d}{d\eta_m} f(x; \boldsymbol{\eta}) dx \\ &= \frac{d}{d\eta_m} \int_{-\infty}^{+\infty} f(x; \boldsymbol{\eta}) dx = 0. \end{aligned} \quad (1.44)$$

Da (1.43) e (1.44) discende immediatamente l'uguaglianza (1.42).  $\square$

## 1.5 Distribuzioni multivariate

Di seguito forniremo qualche cenno su due importanti distribuzioni multivariate: la distribuzione multinomiale e la distribuzione Normale multivariata. Dedicheremo molta più attenzione e dettaglio alla seconda che non alla prima di cui ci limiteremo alla sola definizione.

### 1.5.1 Distribuzione Multinomiale

Questa famiglia di distribuzioni generalizza quanto visto a proposito di distribuzione binomiale a situazioni in cui ogni prova ha  $k$  (piuttosto che due) possibili risultati. Consideriamo la ripetizione di un esperimento casuale  $n$  volte sotto identiche condizioni (le ripetizioni sono indipendenti). Ad ogni ripetizione, l'esperimento termina in uno dei  $k$  possibili modi, mutuamente esclusivi ed esaustivi, diciamo  $C_1, C_2, \dots, C_k$ . Sia  $p_i$  la probabilità che il risultato dell'esperimento sia un elemento

do  $C_i$  e assumiamo che  $p_i$  rimanga costante durante le  $n$  replicazioni dell'esperimento,  $i = 1, 2, \dots, k$ . Definiamo la v.c.  $X_i$  essere il *numero di risultati* che sono elementi di  $C_i$ ,  $i = 1, 2, \dots, k - 1$ . Inoltre, siano  $x_1, x_2, \dots, x_{k-1}$  interi non negativi tali che  $\sum_{i=1}^{k-1} x_i \leq n$ . Allora la probabilità che  $x_1$  risultati dell'esperimento ripetuto cadano in  $C_1, \dots, x_{k-1}$  risultati cadano in  $C_{k-1}$  e di conseguenza  $n - \sum_{i=1}^{k-1} x_i$  cadano in  $C_k$ , è data da

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = n! \prod_{i=1}^k \frac{p_i^{x_i}}{x_i!} \quad (1.45)$$

dove  $x_k$  è semplicemente un'abbreviazione per  $n - \sum_{i=1}^{k-1} x_i$ .

La (1.45) restituisce la funzione di massa probabilistica associata alla distribuzione multinomiale del vettore casuale  $(X_1, X_2, \dots, X_{k-1})$ .

### 1.5.2 Distribuzione Normale multivariata

Consideriamo un vettore di variabili casuali identicamente distribuite  $\mathbf{Z} = (Z_1, \dots, Z_n)$  dove  $Z_i \sim N(0, 1) \forall i = 1, \dots, n$ . La funzione di densità di probabilità di tale vettore può essere scritta come segue:

$$\begin{aligned} f_{\mathbf{Z}}(z_1, \dots, z_n) &= \prod_{i=1}^n f_{Z_i}(z_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z_i^2}{2}\right\} = \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2} \sum_{i=1}^n z_i^2\right\} = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{\mathbf{Z}'\mathbf{Z}}{2}\right\} \end{aligned} \quad (1.46)$$

Inoltre, si può dimostrare facilmente che  $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbb{I}_n)$ , dove con  $\mathbb{I}_n$  si indica la matrice identità e con  $\mathbf{0}$  si denota un vettore nullo ad  $n$  elementi. Infatti, il valore atteso di  $\mathbf{Z}$  coincide con un vettore avente come elementi i valori attesi delle variabili che lo compongono, i quali sono tutti identicamente nulli:  $E(\mathbf{Z}) = (E(Z_1), \dots, E(Z_n)) = \mathbf{0}$ .

Per descrivere la dispersione associata a un vettore casuale, si utilizza invece la matrice di covarianza, la cui definizione generale è fornita di seguito.

**Definizione 1.5.1** (Matrice di covarianza). Definiamo *matrice di covarianza* associata al vettore  $\mathbf{Z}$  ad  $n$  elementi la matrice  $n \times n$  costruita come segue:

$$\Sigma_{\mathbf{Z}} = \begin{pmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) & \dots & \text{Cov}(Z_1, Z_n) \\ \text{Cov}(Z_2, Z_1) & \text{Var}(Z_2) & \ddots & \vdots \\ \vdots & & \ddots & \end{pmatrix} \quad (1.47)$$

Appare evidente che, per un vettore i cui elementi sono indipendenti vale necessariamente  $\text{Cov}(Z_i, Z_j) = 0 \forall i \neq j$ . Inoltre, se  $Z_i \sim N(0, 1)$ , sappiamo che  $\text{Var}(Z_i) = 1 \forall i = 1, \dots, n$ . Pertanto, per il vettore  $\mathbf{Z}$  la matrice covarianza coincide con la matrice identità:  $\Sigma_{\mathbf{Z}} = \mathbb{I}_n$ .

Dato  $Z \sim N_n(\mathbf{0}, \mathbb{I}_n)$ , sia

$$\mathbf{X} = \Sigma^{\frac{1}{2}} \mathbf{Z} + \mu \quad (1.48)$$

dove con  $\Sigma^{\frac{1}{2}}$  indichiamo la matrice definita in modo tale che  $\Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} = \Sigma$  e  $\mu \in \mathbb{R}^n$ . Allora la funzione generatrice dei momenti del vettore casuale  $\mathbf{X}$  è la seguente:

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= E[e^{\mathbf{t}' \mathbf{x}}] = E\left(\exp\left\{\mathbf{t}' (\Sigma^{\frac{1}{2}} \mathbf{Z} + \mu)\right\}\right) = e^{\mathbf{t}' \mu} E\left(\exp\left\{(\Sigma^{\frac{1}{2}} \mathbf{t})' \mathbf{Z}\right\}\right) = \\ &= e^{\mathbf{t}' \mu} \exp\left\{\frac{1}{2} (\Sigma^{\frac{1}{2}} \mathbf{t})' (\Sigma^{\frac{1}{2}} \mathbf{t})\right\} = \exp\left\{\mathbf{t}' \mu + \frac{1}{2} \mathbf{t}' \Sigma \mathbf{t}\right\} \end{aligned} \quad (1.49)$$

Di conseguenza, possiamo dire che  $\mathbf{X} \sim N_n(\mu, \Sigma)$ .

**Definizione 1.5.2.** Un vettore casuale  $\mathbf{X}$  ha una *distribuzione normale n-variata* se la sua funzione generatrice dei momenti è data da

$$M_{\mathbf{X}}(t) = \exp\left\{\mathbf{t}' \mu + \frac{1}{2} \mathbf{t}' \Sigma \mathbf{t}\right\} \quad (1.50)$$

dove  $\mu$  è il vettore della media di  $\mathbf{X}$  mentre  $\Sigma$  è la matrice di covarianza di  $\mathbf{X}$ . Di conseguenza, la funzione densità di  $\mathbf{X}$  sarà la seguente:

$$f_{\mathbf{X}}(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2} (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu)\right\} \quad \forall \mathbf{x} \in \mathbb{R}^n \quad (1.51)$$

**Teorema 1.5.1.** Sia  $\mathbf{X} \sim N_n(\mu, \Sigma)$  e sia  $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ , dove  $A$  è una matrice  $m \times n$ . Allora  $\mathbf{Y} \sim N_m(A\mu + \mathbf{b}, A\Sigma A')$ .

*Dimostrazione.* Determiniamo la funzione generatrice dei momenti del vettore  $\mathbf{Y}$ :

$$\begin{aligned} M_{\mathbf{Y}}(t) &= E\left(e^{\mathbf{t}' \mathbf{Y}}\right) = E\left(e^{\mathbf{t}' (A\mathbf{X} + \mathbf{b})}\right) = e^{\mathbf{t}' \mathbf{b}} E\left(e^{\mathbf{t}' A\mathbf{X}}\right) = e^{\mathbf{t}' \mathbf{b}} E\left(e^{(A'\mathbf{t})' \mathbf{X}}\right) = \\ &= e^{\mathbf{t}' \mathbf{b}} \exp\left\{(A'\mathbf{t})' \mu + \frac{1}{2} (A'\mathbf{t})' \Sigma (A'\mathbf{t})\right\} = e^{\mathbf{t}' \mathbf{b}} \exp\left\{\mathbf{t}' A\mu + \frac{1}{2} \mathbf{t}' A\Sigma A'\mathbf{t}\right\} = \\ &= \exp\left\{\mathbf{t}' (A\mu + \mathbf{b}) + \frac{1}{2} \mathbf{t}' A\Sigma A'\mathbf{t}\right\} \end{aligned} \quad (1.52)$$

Confrontando il risultato con la definizione 1.5.2 si nota che  $\mathbf{Y} \sim N_m(A\mu + \mathbf{b}, A\Sigma A')$ .  $\square$

**Corollario 1.5.1.** Sia  $\mathbf{X} \sim N_n(\mu, \Sigma)$ ; sia  $\mathbf{X}_1$  il vettore costituito dai primi  $m$  elementi di  $\mathbf{X}$  e  $\mathbf{X}_2$  il vettore costituito dai restanti  $n - m$  elementi. Sia inoltre  $\Sigma$  la matrice di covarianza di  $\mathbf{X}$ , i cui elementi siano raggruppati come segue:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (1.53)$$

Con  $\Sigma_{11}$  matrice di dimensione  $m \times m$  e  $\Sigma_{22}$  matrice di dimensione  $(n-m) \times (n-m)$ . Sia inoltre  $\mu_1$  il vettore costituito dai primi  $m$  elementi di  $\mu$  e sia  $\mu_2$  il vettore costituito dai restanti  $n - m$  elementi. Allora  $\mathbf{X}_1 \sim N_m(\mu_1, \Sigma_{11})$ .

*Dimostrazione.* Il vettore  $\mathbf{X}_1$  può essere scritto come  $\mathbf{X}_1 = A\mathbf{X} + \mathbf{0}$ , dove  $A = [\mathbb{I}_m, 0]$  matrice di dimensione  $n \times m$ . Per il teorema 1.5.1 si ha che

$$\mathbf{X} \sim N_m(A\mu, A\Sigma A') = N_m(\mu_1, \Sigma_{11}) \quad (1.54)$$

La tesi è dunque dimostrata.  $\square$

**Teorema 1.5.2.** Sia  $\mathbf{X} \sim N_m(\mu, \Sigma)$  partizionato come nel corollario precedente. Allora  $\mathbf{X}_1$  e  $\mathbf{X}_2$  sono indipendenti se e solo se  $\Sigma_{12} = \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$ .

*Dimostrazione.* Calcoliamo la funzione generatrice dei momenti di  $\underline{\mathbf{X}}$ :

$$\begin{aligned} M_{\mathbf{X}_1 \mathbf{X}_2}(\mathbf{t}_1 \mathbf{t}_2) &= \exp \left\{ \mathbf{t}'_1 \mu_1 + \mathbf{t}'_2 \mu_2 + \frac{1}{2} (\mathbf{t}'_1 \Sigma_{11} \mathbf{t}_1 + \mathbf{t}'_2 \Sigma_{22} \mathbf{t}_2 + \mathbf{t}'_2 \Sigma_{21} \mathbf{t}_1 + \mathbf{t}'_1 \Sigma_{12} \mathbf{t}_2) \right\} = \\ &\quad | \text{ impongo che } \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = 0 \\ &= \exp \left\{ \left( \mathbf{t}'_1 \mu_1 + \frac{1}{2} \mathbf{t}'_1 \Sigma_{11} \mathbf{t}_1 \right) + \left( \mathbf{t}'_2 \mu_2 + \frac{1}{2} \mathbf{t}'_2 \Sigma_{22} \mathbf{t}_2 \right) \right\} = \\ &= \exp \left\{ \mathbf{t}'_1 \mu_1 + \frac{1}{2} \mathbf{t}'_1 \Sigma_{11} \mathbf{t}_1 \right\} \exp \left\{ \mathbf{t}'_2 \mu_2 + \frac{1}{2} \mathbf{t}'_2 \Sigma_{22} \mathbf{t}_2 \right\} = M_{\mathbf{X}_1}(\mathbf{t}_1) M_{\mathbf{X}_2}(\mathbf{t}_2) \end{aligned} \quad (1.55)$$

Da tale proprietà della funzione generatrice dei momenti discende che  $\mathbf{X}_1$  e  $\mathbf{X}_2$  sono indipendenti.  $\square$

**Teorema 1.5.3.** Sia  $\mathbf{X} \sim N_m(\mu, \Sigma)$  e sia data l'usuale partizione del vettore in  $\mathbf{X}_1$  e  $\mathbf{X}_2$ . Assumiamo inoltre che  $\Sigma$  sia definita positiva. Allora la distribuzione condizionata di  $\mathbf{X}_1 | \mathbf{X}_2$  è la seguente:

$$N_m \left( \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{X}_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right) \quad (1.56)$$

*Dimostrazione.* Definiamo  $\mathbf{W} = \mathbf{X}_1 - \Sigma_{12} \Sigma_{22}^{-1} \mathbf{X}_2$  e consideriamo la distribuzione di  $\mathbf{W}$  e  $\mathbf{X}_2$  che si ottiene dalla seguente trasformazione:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{W} \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbb{I}_m & -\Sigma_{12} \Sigma_{22}^{-1} \\ 0 & \mathbb{I}_{n-m} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = A\mathbf{X} \quad (1.57)$$

Per il teorema 1.5.1 vale che  $\mathbf{Y} \sim N_n(A\mu, A\Sigma A')$ . Inoltre, appare evidente che il valore atteso di  $\mathbf{Y}$  può essere partizionato come segue:

$$E(\mathbf{Y}) = \begin{pmatrix} \mu_1 - \Sigma_{12} \Sigma_{22}^{-1} \mu_2 \\ \mu_2 \end{pmatrix} \quad \text{da cui} \quad \begin{cases} E(\mathbf{W}) = \mu_1 - \Sigma_{12} \Sigma_{22}^{-1} \mu_2 \\ E(\mathbf{X}_2) = \mu_2 \end{cases} \quad (1.58)$$

Ci aspettiamo inoltre che la matrice covarianza sia costruita come segue:

$$\Sigma_Y = \begin{pmatrix} \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \text{da cui} \quad \begin{cases} \text{Var}(\mathbf{W}) = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \\ \text{Var}(\mathbf{X}_2) = \Sigma_{22} \end{cases} \quad (1.59)$$

Per il teorema 1.5.2, poiché la covarianza tra  $\mathbf{W}$  e  $\mathbf{X}_2$  è nullla, tali vettori sono indipendenti. Pertanto:

$$\mathbf{W}|\mathbf{X}_2 \sim \mathbf{W} \sim N_m(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \quad (1.60)$$

Poiché, per definizione di  $\mathbf{W}$ , si ha che  $\mathbf{X}_1 = \mathbf{W} + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{X}_2$  allora  $\mathbf{X}_1|\mathbf{X}_2$  è la seguente:

$$\mathbf{X}_1|\mathbf{X}_2 = \mathbf{W} + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{X}_2 \sim N_m(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \quad (1.61)$$

Tale distribuzione è nota come modello di regressione lineare<sup>1</sup> di  $\mathbf{X}_1$  su  $\mathbf{X}_2$ .  $\square$

---

<sup>1</sup>Dato  $\mathbf{X}_1 = \beta_0 + \beta_1\mathbf{X}_2$  esso può essere stimato mediante la seguente quantità:  $\hat{\mathbf{X}}_1 = (\mathbf{X}_1 - b_1\mathbf{X}_2) + \frac{\sigma_{12}}{\sigma_2^2}\mathbf{X}_2$

## 2 Popolazioni, campioni e statistiche

La statistica matematica è una disciplina che utilizza gli strumenti del calcolo delle probabilità per fini statistici legati alla modelizzazione e alla previsione di fenomeni stocastici. Però, a differenza di quanto accade in ambito probabilistico, il ragionamento proprio della statistica matematica è un *ragionamento* di tipo *induttivo*: estende all'universo di riferimento (più propriamente, *popolazione target*) le conclusioni inerenti gli aspetti di interesse ottenute sulla base dell'osservazione di un suo sottoinsieme rappresentativo (comunemente detto *campione*). Vale la pena qui spendere qualche riga per dare le definizioni di alcuni concetti fondamentali in parte già richiamati.

**Definizione 2.0.1** (Popolazione statistica). In statistica per *popolazione* (o collettivo statistico o aggregato) si intende l'insieme degli elementi che sono oggetto di studio, ovvero l'insieme delle unità (dette unità statistiche) sulle quali viene effettuata la rilevazione delle modalità con le quali il fenomeno studiato si presenta.

**Definizione 2.0.2** (Campione casuale (semplice)). Un *campione casuale* è una  $n$ -upla di variabili casuali  $X_1, \dots, X_n$  indipendenti ed identicamente distribuite (*i.i.d.*), sottoinsieme della popolazione. Si definisce inoltre *spazio campionario* (o spazio dei campioni) lo spazio  $\mathfrak{X}$  di tutti i possibili valori (o determinazioni campionarie)  $(x_1, x_2, \dots, x_n)$  della  $n$ -pla  $(X_1, X_2, \dots, X_n)$ .

In altre parole, possiamo anche dire che per popolazione si intende un'insieme di unità statistiche (numeri, misure o osservazioni) che si vogliono esaminare e relativamente al quale interessa poter dire qualcosa in merito a uno o più suoi aspetti di interesse; mentre per campione si intende quella parte di unità statistiche estratte casualmente dalla popolazione per realizzare, sfruttando l'informazione in essa contenuta, l'obiettivo conoscitivo che ci si è prefissato. Ed è proprio la *casualità* del meccanismo di estrazione del campione dalla popolazione a garantire che esso riproduca la struttura della popolazione stessa ovvero ne sia rappresentativo in termini di struttura di frequenza; e ciò è di garanzia per le conclusioni inferenziali basate sul meccanismo di ragionamento induttivo che sta alla base del ragionamento statistico.

Indicata con  $N$  la dimensione della popolazione e con  $n$  quella del campione (=ampiezza campionaria), usualmente il campionamento da una popolazione è effettuato *senza riposizione* e, sebbene in questo caso non sia possibile parlare di indipendenza, per  $N$  grande (quando finito) rispetto a  $n$  possiamo comunque parlare di una situazione di *quasi-indipendenza* (in questo caso gli effetti dei due schemi di campionamento *con riposizione* e *senza riposizione* sono di fatto *indistinguibili*). Come già ricordato, un campione casuale contiene una certa quantità di informazione (che, in genere, cresce con  $n$ ) relativamente al *meccanismo generatore dei dati* restituito dal *modello statistico*.

**Definizione 2.0.3** (Modello statistico). Un *modello statistico*  $\mathcal{F}$  è una famiglia di distribuzioni di probabilità per  $(X_1, X_2, \dots, X_n)$  che si presume approssimi sufficientemente bene, o addirittura contenga, il meccanismo probabilistico che ha generato i dati disponibili. Inoltre, un modello statistico  $\mathcal{F}$  è detto *parametrico* se esiste  $\Theta \subseteq \mathbb{R}^k$  e una mappa (biettiva)  $\Theta \rightarrow \mathcal{F}$ ; scriveremo allora  $\mathcal{F}_\theta$ .

Ogni procedura statistica (e, in ultima analisi, l'intera statistica matematica) si pone, in particolare, due obiettivi: la **sintesi** e la **conservazione** dell'informazione rilevante contenuta in una determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale  $(X_1, X_2, \dots, X_n)$  e relativa al fenomeno di interesse. Si tratta ora di individuare uno strumento opportuno per sintetizzare l'informazione campionaria (magari conservando quella, e solo quella, rilevante per il problema inferenziale di interesse).

**Definizione 2.0.4** (Statistica). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale di una distribuzione  $F_\theta(x)$  (che spesso scriveremo anche  $F_X(x; \theta)$ ). Ogni funzione

$$T(X_1, X_2, \dots, X_n) : \mathfrak{X} \rightarrow \mathbb{R}^k, \text{ con } k \in \mathbb{N} \quad (2.1)$$

che non dipenda in modo alcuno da parametri incogniti è detta *statistica*.

Ora una statistica è

- una caratteristica numerica del campione e che realizza una sintesi dell'informazione su  $\theta \in \Theta$  contenuta nei dati; ed è altresì fondamentale che tale sintesi sia (massimamente) informativa su  $\theta$  o su una qualche sua funzione  $g(\theta)$  cui l'inferenza è dedicata; per esempio l'interesse può riguardare il valore plausibile di  $\theta$  o di una qualche sua funzione
- a sua volta una v.c. (univariata o multivariata) e ha una sua propria peculiare distribuzione di probabilità (che chiameremo *distribuzione campionaria*), mutuata dalla distribuzione del campione e che la caratterizza completamente. Ciò ha come immediata conseguenza il fatto che le proprietà dell'inferenza (*accuratezza* e *precisione*, per esempio) basata su di una certa statistica dipendono dalla distribuzione della stessa statistica (da qui, come vedremo, il ruolo essenziale giocato dai teoremi di trasformazione di v.c.)

Sarà peraltro essenziale porsi il problema di come *individuare la forma funzionale* della statistica sulla quale poggiare la soluzione del problema inferenziale di interesse; studieremo diverse tecniche per l'individuazione di oportune statistiche e tra queste il principio del *plug-in*, quello *pivotal* e quello basato sulla *sufficienza*.

Sarà anche importante individuare alcuni *criteri* per il *confronto* e la *scelta* tra statistiche diverse: alcuni di questi, come vedremo, saranno restituiti da concetti quali quello di

- *non distorsione* (=accuratezza)
- *efficienza* (=precisione)
- *consistenza* (=capacità di produrre inferenze sempre più stabili al crescere dell'ampiezza campionaria)
- *robustezza* (=capacità di resistere a distorsioni indotte dalla presenza di outlier nel campione)

per citarne alcuni tra i più comunemente impiegati.

Consideriamo il seguente semplice esempio per identificare e rendere operativi i concetti fin qui introdotti.

**Esempio 2.0.1.** Supponiamo di voler analizzare statisticamente un set di dati associati al numero di cicloni verificatisi nel nord-est dell'Australia nel corso di 13 stagioni, non necessariamente in successione. Disponiamo dunque dei seguenti dati, riportati in tabella:

Stagione	1	2	3	4	5	6	7	8	9	10	11	12	13
N° di cicloni	6	3	5	6	6	3	12	7	5	2	6	7	5

**Tabella 2.1:** Numero di cicloni nel corso di 13 stagioni

Osserviamo subito come in questo caso le 13 stagioni rappresentino le unità statistiche campionarie e costituiscono il *campione casuale* a nostra disposizione mentre la *popolazione target* è costituita dall'insieme delle stagioni che si sono succedute nel tempo (o in un certo lasso di tempo). Su ciascuna unità statistica (ovvero su ogni elemento del campione) è stata "misurata" l'intensità del fenomeno di interesse, misura che è qui rappresentata dal numero di cicloni verificate nella stagione. A partire dai dati campionari a disposizione, è possibile innanzitutto calcolare la *media* e la *varianza*, definite rispettivamente come

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \quad s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2 \quad (2.2)$$

Possiamo subito identificare nelle due quantità  $\sum_{i=1}^n x_i$  e  $\sum_{i=1}^n x_i^2$  i valori delle due statistiche  $\sum_{i=1}^n X_i$  e  $\sum_{i=1}^n X_i^2$  usate per riassumere l'informazione contenuta dati a disposizione.

Sottolineiamo che è necessario, al fine di trarre delle informazioni rilevanti, calcolare (almeno) entrambe le quantità presenti in (2.2). Infatti, attraverso il solo calcolo della media, si cattura l'informazione relativa al centro della distribuzione e non si è però in grado di quantificare quanto sia variabile il numero dei cicloni nelle diverse stagioni e, in ultima analisi, quanto sia "stabile" il fenomeno nel tempo. Questa informazione può essere parzialmente recuperata mediante il calcolo della varianza<sup>1</sup>, che è un indicatore di quanto i dati si discostano dal loro valore medio (si parla di *dispersione della distribuzione*).

Svolgendo i calcoli per il set di dati proposto in questo esempio, si ottiene  $\bar{x}_n = 5,615$  e  $s^2 = 5,621$ . Appare subito evidente che, in questo caso,  $\bar{x}_n \simeq s^2$  e questa informazione può essere utilizzata per la scelta di una distribuzione di probabilità (ovvero di un *modello statistico* e in questo caso, *parametrico*) che approssimi bene i dati, per esempio la distribuzione di Poisson:  $X \sim \mathcal{P}(\lambda)$ ,  $\lambda > 0$ , dove ora indichiamo con  $X$  la variabile casuale associata al numero di cicloni in una stagione.

<sup>1</sup>Si noti che, nella formula per il calcolo della varianza, la presenza di una somma quadratica garantisce il fatto che ogni termine porti un contributo positivo alla varianza, e pertanto garantisce che non si abbia una somma nulla a meno che ogni termine non sia esattamente pari al valor medio dell'intero insieme di osservazioni.

Ricordiamo che la distribuzione di Poisson fornisce un modello (statistico) parametrico, ovvero un modello matematico di forma funzionale nota e che dipende da un parametro  $\lambda > 0$  non noto, ed è definita come segue:

$$P_\lambda(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \mathbb{1}_{\mathbb{N}^*}(x) \quad \text{con } \mathbb{N}^* = \mathbb{N} \cup \{0\}, \quad \lambda > 0. \quad (2.3)$$

In virtù del fatto che il parametro  $\lambda$  rappresenta la media della distribuzione di Poisson possiamo per esempio porre  $\hat{\lambda} = \bar{x}_n$  dove  $\hat{\lambda}$  rappresenta una stima *plug-in* (o basata sull'equivalente campionario) del parametro  $\lambda$ .

Va da sé che la stima del parametro  $\lambda$  tramite  $\hat{\lambda} = \bar{x}_n$  permette di *identificare* tra gli infiniti modelli di Poisson (uno per ogni valore ammissibile del parametro  $\lambda$ ) che ne costituiscono la famiglia, quello che risulta più vicino al meccanismo (stocastico) che ha generato i dati osservati e che quindi si presta bene a rappresentare.

In sintesi, l'individuazione di un qualsivoglia *modello statistico (parametrico)* per un fenomeno oggetto di interesse si concretizza in tre fasi che possiamo così riassumere:

- *specificazione* del modello statistico, ovvero la ricerca di una relazione funzionale tra le variabili di interesse che descrivono il fenomeno sfruttando la teoria disponibile su di esso, le ipotesi assunte e l'esperienza.
- *stima dei parametri* del modello, ovvero la fase in cui si stimano i valori dei parametri incogniti del modello scelto
- *verifica del modello* che consiste nel verificare la bontà di adattamento (o *goodness-of-fit*) del modello ipotizzato e stimato ai dati ai fini di una possibile adozione dello stesso a descrivere/rappresentare il fenomeno a cui i dati si riferiscono e a fornire la base su cui poggiare le tecniche inferenziali sugli aspetti di interesse del fenomeno considerato.

I metodi della statistica parametrica utilizzati per la soluzione di problemi di carattere univariato e multivariato hanno, come limitazione, la necessità di dover ricorrere all'introduzione di ipotesi piuttosto restrittive, spesso formulate ad hoc per rendere praticabili le procedure inferenziali che riassumono quel ragionamento induttivo quale caratterizzante la statistica a cui si faceva poc'anzi riferimento. A questo si deve aggiungere che le assunzioni che rendono valida l'applicazione di tali metodi (normalità, omoschedasticità, indipendenza e identica distribuzione della componente stocastica erratica) quand'anche soddisfatte, portano a situazioni in cui i risultati si possono ottenere solo tramite *procedure di approssimazione* avvalendosi di **teoremi limite** che garantiscano la convergenza in *probabilità* (Legge Debole dei Grandi Numeri o più semplicemente *LDGN*) o in *distribuzione* (Teorema Limite Centrale, o più semplicemente *TLC*) degli oggetti statistici coinvolti nelle procedure inferenziali apprendo il grande capitolo delle *procedure asintotiche*. Anche su questo argomento, nel corso delle lezioni, avremo modo di tornare con calma.

## 2.1 Funzione di distribuzione empirica e principio del plug-in

Alla base delle procedure inferenziali in ambito di stima e test di ipotesi vi sono una serie di tecniche tramite le quali estrarre l'informazione contenuta nella determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale  $(X_1, X_2, \dots, X_n)$  relativamente a

uno o più aspetti di interesse della popolazione (siano essi parametri o funzioni di parametri).

Una di queste tecniche consiste nel *principio del plug-in*; esso consiste essenzialmente nello sfruttare l'equivalente campionario del parametro o altro aspetto di interesse e, da punto di vista formale, poggia su quella che è nota come *distribuzione empirica*.

Sia ora  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una popolazione avente funzione di distribuzione  $F_X(x; \theta)$  con  $\theta \in \Theta \subseteq \mathbb{R}^k$ , che per semplicità di notazione indicheremo con  $F_X(x)$ , omettendo la dipendenza sul parametro  $\theta$ ; e sia  $(x_1, x_2, \dots, x_n)$  una sua determinazione.

**Definizione 2.1.1** (Distribuzione di probabilità empirica). Sia  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  una determinazione del campione casuale  $(X_1, X_2, \dots, X_n)$ . La *distribuzione di probabilità empirica* associata a  $\mathbf{x}$  e si indica con  $\hat{P}_n$  la distribuzione probabilistica discreta che assegna probabilità  $1/n$  ad ogni  $x_i$ :

$$\hat{P}_n(\{x_i\}) = \frac{1}{n}, \quad \forall i = 1, \dots, n \quad (2.4)$$

Si noti che  $\hat{P}_n$  è un'approssimazione della distribuzione da cui proviene il campione; dato un evento  $A$ ,

$$P(A) \simeq \hat{P}_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(x_i) \quad (2.5)$$

dove  $\hat{P}_n(A)$  altro non è che la *frequenza relativa osservata* (o *empirica* o *campionaria*) di  $A$ .

**Definizione 2.1.2** (Distribuzione cumulata empirica). La *distribuzione cumulata empirica* (o semplicemente, *distribuzione empirica*)  $\hat{F}_n$  è la funzione di distribuzione associata alla distribuzione di probabilità empirica  $\hat{P}_n$ :

$$\hat{F}_n(x) = \hat{P}_n(X \leq x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i) \quad (2.6)$$

La distribuzione cumulata empirica, qualsiasi sia la distribuzione probabilistica a cui è associata, gode delle seguenti proprietà:

- $\hat{F}_n(-\infty) = 0$  e  $\hat{F}_n(\infty) = 1$
- è *non decrescente*
- è *continua da destra*

**Teorema 2.1.1.** Per un valore  $x \in \mathbb{R}$  fissato (seppur arbitrario) si ha che

$$n \hat{F}_n(x) \sim b(n, F(x)) \quad (2.7)$$

sicché

1.  $\mathbb{E}(\hat{F}_n(x)) = F(x)$

$$2. \operatorname{Var}\left(\hat{F}_n(x)\right) = \frac{F(x)[1-F(x)]}{n}$$

*Dimostrazione.* Conviene subito ricordare che

$$\hat{F}_n(x) = \hat{P}_n(X \leq x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i) \quad (2.8)$$

e dunque essa è una *somma di v.c. indipendenti* di Bernoulli con probabilità di successo  $p = F(x)$ . Di conseguenza,  $n \hat{F}_n(x)$  è una v.c. binomiale sicchè media e varianza di  $\hat{F}_n(x)$  dono date da 1. e 2., rispettivamente.  $\square$

Inoltre, come vedremo tra poco, è un *buon* stimatore della funzione di distribuzione  $F(x)$  al crescere dell'ampiezza del campione su cui è calcolata.

La figura 2.1 riporta la funzione di distribuzione cumulata di una distribuzione Normale standard e l'andamento della funzione di distribuzione empirica per diverse ampiezze campionarie ( $n = 20, 100, 300, 30000$ ); come si può notare, al crescere di  $n$ ,  $\hat{F}_n$  è sempre più vicina a  $F$ .

### 2.1.1 Uso della funzione di distribuzione empirica

Molte volte non siamo direttamente interessati alla funzione di distribuzione  $F_X(x)$  della v.c.  $X$  quanto piuttosto ad alcune funzioni che la coinvolgono direttamente conosciute con il termine di *funzionali statistici* quali per esempio media, varianza, mediana, quartili e così via che si prestano a rappresentare, sinteticamente e spesso a fini di confronto, aspetti caratteristici della distribuzione di  $X$ . Va da sé che queste quantità risulteranno funzioni del parametro incognito  $\theta \in \Theta$ , scalare o vettore che sia, che indicizza la distribuzione e in quanto tali andranno stimate sulla base dei dati disponibili.

In molti casi possiamo ricavare degli stimatori per le quantità di cui sopra semplicemente sostituendo  $F_X(x) = F_X(x; \theta)$  con  $\hat{F}_n(x)$  ottenendo quella che si dice essere una loro stima *plug-in*.

**Definizione 2.1.3** (Media campionaria). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione  $F_X(x)$  e sia  $\mu = \mu'_1 = \mathbb{E}(X) = \int_{-\infty}^{\infty} x dF(x)$  la media della distribuzione  $F(x)$  di  $X$ , assunto  $\mathbb{E}(X)$  esistere. La stima *stima plug-in* di  $\mu$  è data da

$$\hat{\mu}_n = \int_{-\infty}^{\infty} x d\hat{F}_n(x) = \sum_{i=1}^n X_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n \quad (2.9)$$

nota come *media campionaria*.

La media campionaria  $\bar{X}_n$  è, com'è evidente dalla sua definizione, una variabile casuale e come tale avrà una sua distribuzione che dipenderà dalla distribuzione da cui proviene il campione casuale su cui essa è calcolata.

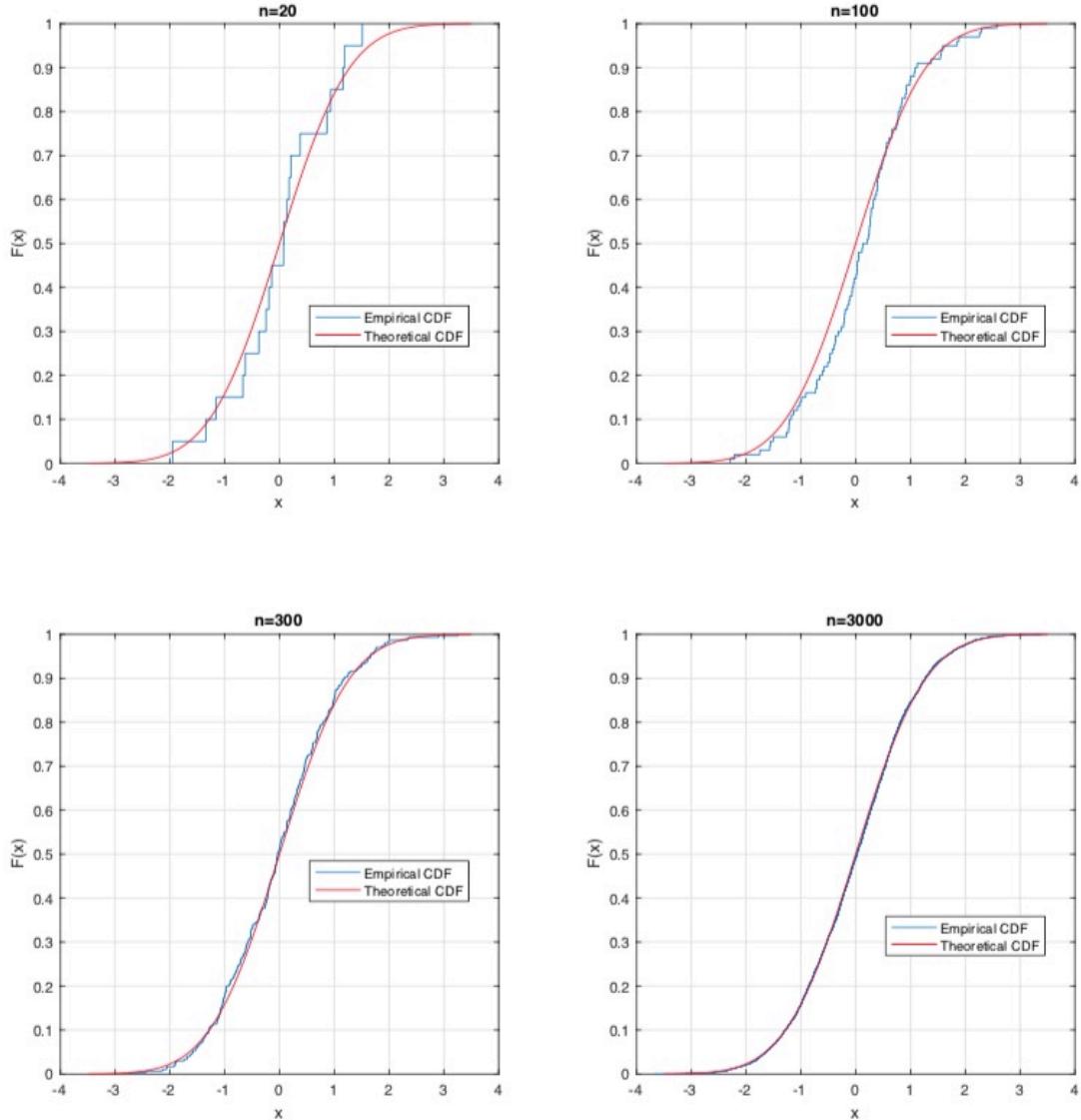


Figura 2.1: Andamento di  $\hat{F}_n(x)$  al crescere dell'ampiezza campionaria  $n$

**Teorema 2.1.2** (Varianza di  $\bar{X}_n$ ). Sia  $X_1, \dots, X_n$  un campione casuale proveniente da una distribuzione  $F_X(x)$ . La varianza della media campionaria  $\bar{X}_n$  è data da

$$\text{Var}_\theta(\bar{X}_n) = \frac{\sigma^2}{n} \quad (2.10)$$

dove  $\sigma^2 = \text{Var}(X) = \mathbb{E}[(X - \mu)^2]$  è la varianza della popolazione da cui proviene il campione.

*Dimostrazione.* La dimostrazione è immediata e lasciata per esercizio.  $\square$

Veniamo ora alla definizione dello *stimatore plug-in* per la varianza della popolazione  $\sigma^2 = \text{Var}(X)$ .

**Definizione 2.1.4.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione  $F_X(x)$  e sia  $\sigma^2 = \mu_2 = \mathbb{E}[(X - \mu)^2]$  la varianza della distribuzione di  $X$ , assunto  $\mathbb{E}[(X - \mu)^2]$  esistere. Lo stimatore *stima plug-in* di  $\sigma^2$  è dato da

$$\begin{aligned}\hat{\sigma}_n^2 &= \int_{-\infty}^{\infty} (x - \bar{x}_n)^2 d\hat{F}_n(x) = \sum_{i=1}^n (X_i - \bar{X}_n)^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2\end{aligned}\tag{2.11}$$

**Definizione 2.1.5 (Varianza campionaria).** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione  $F_X(x)$ . La *varianza campionaria*  $S_n^2$  è data da

$$S_n^2 = \frac{n}{n-1} \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2\tag{2.12}$$

Trovare la varianza della v.c.  $S_n^2$  richiede un po' più di lavoro di quello richiesto per la varianza della v.c.  $\bar{X}_n$ ; nel ricavare tale quantità si ha anche modo di incontrare un'interessante applicazione della funzione generatrice dei momenti. Vale la pena spenderci un po' di tempo.

**Teorema 2.1.3 (Varianza di  $S_n^2$ ).** Sia  $X_1, \dots, X_n$  un campione casuale proveniente da una distribuzione  $F_X(x; \theta)$ , con  $\theta \in \Theta$ , e sia  $S_n^2$  definito come segue:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2\tag{2.13}$$

Allora la varianza di  $S_n^2$  può essere scritta in funzione dei momenti di ordine 2 e 4, come segue:

$$\mathbb{V}ar_\theta(S_n^2) = \frac{1}{n} \left( \mu_4 - \frac{n-3}{n-1} \mu_2^2 \right)\tag{2.14}$$

*Dimostrazione.* Dalla definizione di varianza di una v.c. si ha

$$\begin{aligned}\mathbb{V}ar_\theta(S_n^2) &= \mathbb{E}_\theta [S_n^2 - \mathbb{E}_\theta(S_n^2)]^2 \\ &= \mathbb{E}_\theta [S_n^2]^2 - \mu_2^2 \\ &= \mathbb{E}_\theta [S_n^4] - \sigma^4\end{aligned}$$

avendo già dimostrato che  $\mathbb{E}_\theta [S_n^2] = \sigma^2$  ovvero che  $S_n^2$  è uno stimatore non distorto

(=accurato) per la varianza della popolazione  $\mu_2 = Var_\theta(X) = \sigma^2$ . Inoltre,

$$\begin{aligned}
S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \sum_{i=1}^n [X_i - \bar{X}_n + \mu - \mu]^2 \\
&= \frac{1}{n-1} \sum_{i=1}^n [(X_i - \mu) - (\bar{X}_n - \mu)]^2 \\
&= \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \text{ con } Z_i = (X_i - \mu), \bar{Z}_n = (\bar{X}_n - \mu) \\
&= \frac{1}{n-1} \sum_{i=1}^n Z_i^2 - \frac{n}{n-1} \bar{Z}_n^2 \\
&= \frac{1}{n-1} \sum_{i=1}^n Z_i^2 - \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^n Z_i \right)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^n Z_i^2 - \frac{n}{n(n-1)} \left( \sum_{i=1}^n Z_i \right)^2 \\
&= \frac{1}{n(n-1)} \left[ n \sum_{i=1}^n Z_i^2 - \left( \sum_{i=1}^n Z_i \right)^2 \right]
\end{aligned}$$

Ma allora,

$$[S_n^2]^2 = S_n^4 = \frac{1}{[n(n-1)]^2} \left\{ n^2 \left[ \sum_{i=1}^n Z_i^2 \right]^2 - 2n \sum_{i=1}^n Z_i^2 \left[ \sum_{i=1}^n Z_i \right]^2 + \left[ \sum_{i=1}^n Z_i \right]^4 \right\}$$

e dunque

$$\mathbb{E}_\theta [S_n^4] = \frac{1}{[n(n-1)]^2} \left\{ n^2 \mathbb{E}_\theta \left[ \sum_{i=1}^n Z_i^2 \right]^2 - 2n \mathbb{E}_\theta \left[ \sum_{i=1}^n Z_i^2 \left( \sum_{i=1}^n Z_i \right)^2 \right] + \mathbb{E}_\theta \left[ \sum_{i=1}^n Z_i \right]^4 \right\} \quad (2.15)$$

Ora essendo  $X_1, X_2, \dots, X_n$  v.c. indipendenti poiché elementi del campione casuale anche  $Z_1, Z_2, \dots, Z_n$ , con  $Z_i = (X_i - \mu)$ ,  $i = 1, 2, \dots, n$ , saranno v.c. indipendenti di media nulla (vale a dire,  $\mathbb{E}_\theta(Z_i) = 0, \forall i$ ) sicché

$$\mathbb{E}_\theta(Z_i \cdot Z_j) = 0, \mathbb{E}_\theta(Z_i^3 \cdot Z_j) = 0, \mathbb{E}_\theta(Z_i^2 \cdot Z_j \cdot Z_k) = 0, i \neq j \neq k$$

$$\mathbb{E}_\theta(Z_i^2 \cdot Z_j^2) = \mathbb{E}_\theta(Z_i^2) \cdot \mathbb{E}_\theta(Z_j^2) = \mu_2 \cdot \mu_2 = \mu_2^2, \mathbb{E}_\theta(Z_i^4) = \mu_4$$

dove  $\mu_2$  e  $\mu_4$  sono i momenti centrati, rispettivamente di ordine 2 e 4, della distribuzione  $F_X(x; \theta)$  da cui proviene il campione casuale. Ora,

$$\mathbb{E}_\theta \left[ \sum_{i=1}^n Z_i^2 \right]^2 = n\mu_4 + n(n-1)\mu_2^2$$

$$\begin{aligned}\mathbb{E}_\theta \left[ \sum_{i=1}^n Z_i^2 \left( \sum_{i=1}^n Z_i \right)^2 \right] &= n\mu_4 + n(n-1)\mu_2^2 \\ E_\theta \left[ \sum_{i=1}^n Z_i \right]^4 &= n\mu_4 + 3n(n-1)\mu_2^2\end{aligned}$$

da cui sostituendo in (2.15), con qualche passaggio e semplificazione, si ottiene

$$\mathbb{E}_\theta [S_n^4] = \mathbb{E}_\theta [S_n^2]^2 = \frac{(n-1)\mu_4 + (n^2 - 2n + 3)\mu_2^2}{n(n-1)}$$

Infine si ha

$$\begin{aligned}Var_\theta(S_n^2) &= \mathbb{E}[S_n^2]^2 - [\mathbb{E}_\theta(S_n^2)]^2 \\ &= \frac{(n-1)\mu_4 + (n^2 - 2n + 3)\mu_2^2}{n(n-1)} - (\mu_2)^2 \\ &= \frac{1}{n} \left( \mu_4 - \frac{n-3}{n-1} \mu_2^2 \right)\end{aligned}$$

ed è quello che si voleva dimostrare.

Ora, nel caso in cui il campione provenga da una distribuzione Normale,  $N(\mu, \sigma^2)$ , si ha

$$\mu_2 = \mathbb{E}_{\mu, \sigma^2}(X - \mu)^2 = \sigma^2$$

$$\mu_4 = \mathbb{E}_{\mu, \sigma^2}(X - \mu)^4 = 3\sigma^4$$

sicché

$$\begin{aligned}Var_{\mu, \sigma^2}(S_n^2) &= \frac{1}{n} \left[ 3\sigma^4 - \frac{n-3}{n-1} (\sigma^2)^2 \right] \\ &= \frac{1}{n} \left[ \frac{3(n-1)\sigma^4 - (n-3)\sigma^4}{n-1} \right] \\ &= \frac{1}{n} \cdot \frac{2n\sigma^4}{(n-1)} \\ &= \frac{2\sigma^4}{n-1}\end{aligned}$$

□

In termini del tutto generali,

$$\hat{\eta}_n = \int_{-\infty}^{\infty} g(x) d\hat{F}_n(x) = \sum_{i=1}^n g(X_i) \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n g(X_i) \quad (2.16)$$

è lo stimatore *plug-in* di  $\eta = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) dF(x)$ .

### Nota

Il principio del *plug-in* poggia su una procedura di *campionamento ripetuto* e, in ultima analisi, sull'*equivalenza* - al crescere dell'ampiezza campionaria  $n$  - della struttura di frequenza del campione casuale e della popolazione da cui il campione proviene.

## 2.1.2 Il metodo dei momenti

Il metodo dei momenti consente di individuare, convolgendo nel calcolo i momenti della popolazione e i loro analogi campionari ottenuti tramite *plug-in*, i valori di quantità di interesse come per esempio i parametri di una distribuzione.

Per applicare questo metodo, è necessario partire dai cosiddetti *momenti campionari*, ovvero quei momenti che dipendono esclusivamente dal campione.

**Definizione 2.1.6** (Momento campionario di ordine  $s$ ). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione  $F_X(x; \theta)$  con  $\theta \in \Theta \subseteq \mathbb{R}^k$  e sia  $s \in \mathbb{N}$ . Definiamo momento campionario (non centrato) di ordine  $s$  (o *stimatore plug-in* di  $\mu'_s = \mathbb{E}(X^s)$ ) la quantità

$$m'_s = \frac{1}{n} \sum_{i=1}^n X_i^s \quad (2.17)$$

Uguagliando i primi  $k$  momenti della popolazione  $\mu'_1, \mu'_2, \dots, \mu'_k$  ai corrispondenti momenti campionari otterremo un sistema di  $k$  equazioni

$$\mu'_s = m'_s \quad (2.18)$$

per  $s = 1, 2, \dots, k$  e tenuto conto del fatto che i momenti  $\mu'_s$  sono funzioni del parametro  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ , (2.18) costituisce un sistema di  $k$  equazioni in  $k$  incognite la cui soluzione fornisce una stima di  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$  detta anche *stima con il metodo dei momenti*.

**Esempio 2.1.1.** Sia  $X$  una variabile casuale bernoulliana  $X \sim b(1, p)$  e sia  $(X_1, X_2, \dots, X_n) \in \mathfrak{X}$  un icampione casuale di ampiezza  $n$ . Il modello statistico associato a questa distribuzione è la distribuzione bernoulliana

$$P(X = x) = p^x (1 - p)^{1-x} \mathbb{1}_{\{0,1\}}(x) \quad (2.19)$$

e abbiamo già avuto modo di vedere che  $\mu'_1 = \mathbb{E}(X) = p$ . Ora essendo la distribuzione di Bernoulli parametrizzata da un solo parametro  $\theta_1 = p$ , il sistema di equazioni (2.18) si riduce a

$$\mu'_1 = m'_1 \quad \text{ovvero} \quad p = \frac{1}{n} \sum_{i=1}^n X_i \quad (2.20)$$

per cui lo stimatore del parametro  $p$  ottenuto con il metodo dei momenti è uguale a  $\tilde{p}_n = \bar{X}_n$ . quest'ultimo è, com'era da aspettarsi, a sua volta una v.c. caratterizzata da una propria distribuzione le cui caratteristiche saranno fondamentali per giudicare della bontà dell'approssimazione (=stima) di  $p$  fornita da  $\tilde{p}_n = \bar{X}_n$ .

**Esempio 2.1.2** (Distribuzione Normale). Supponiamo di voler approssimare il set di dati  $(X_1, X_2, \dots, X_n)$  a nostra disposizione mediante una distribuzione Normale  $N(\mu, \sigma^2)$ . Per farlo, è necessario trovare i valori dei parametri  $\mu$  e  $\sigma^2$  ovvero valor medio e varianza. Abbiamo già mostrato che, per definizione, il valor medio coincide con il momento di ordine uno centrato nell'origine ( $\mu'_1 = \mathbb{E}(X) = \mu$ ), mentre il momento di ordine 2 centrato nel valor medio è pari alla varianza:  $\mathbb{E}(X - \mu'_1)^2 = E(X - \mu)^2 = \sigma^2$ . Se ne deduce che  $\mu'_2 = E(X^2) = \sigma^2 + \mu^2$ . Infatti:

$$\sigma^2 = E(X - \mu)^2 = E(X^2) + \mu^2 - 2\mu E(X) = E(X^2) - \mu^2 \quad (2.21)$$

A questo punto, è sufficiente imporre che i momenti campionari,  $m'_1$  e  $m'_2$ , siano uguali ai due corrispondenti momenti della distribuzione, rispettivamente  $\mu'_1$  e  $\mu'_2$ :

$$\begin{cases} \mu = \frac{1}{n} \sum_{i=1}^n X_i \\ \sigma^2 + \mu^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases} \quad (2.22)$$

Risolvendo il precedente sistema rispetto  $\mu$  e  $\sigma^2$  otteniamo lo stimatore del metodo dei momenti di  $\theta = (\mu, \sigma^2)$  dato da

$$\tilde{\theta}_n = \left( \frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right) \quad (2.23)$$

**Esempio 2.1.3** (Distribuzione Gamma). Supponiamo ora di voler approssimare il set di dati  $(X_1, X_2, \dots, X_n)$  a nostra disposizione mediante una distribuzione Gamma di parametri  $\alpha$  e  $\beta$ , entrambi positivi, applicando il metodo dei momenti. Abbiamo già mostrato che per tale distribuzione i momenti di ordine 1 e 2 centrati nell'origine sono rispettivamente  $\mu'_1 = E(X) = \alpha\beta$  e  $\mu'_2 = E(X^2) = \alpha\beta^2 + \alpha^2\beta^2$ . Come visto nell'esempio precedente, è sufficiente imporre che i momenti campionari,  $m'_1$  e  $m'_2$ , siano uguali ai due corrispondenti momenti della distribuzione, rispettivamente  $\mu'_1$  e  $\mu'_2$ , ottenendo il sistema di (due) equazioni in  $\alpha$  e  $\beta$

$$\begin{cases} \alpha\beta = \frac{1}{n} \sum_{i=1}^n X_i \\ \alpha\beta^2 + \alpha^2\beta^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases} \quad (2.24)$$

Risolvendo rispetto  $\alpha$  e  $\beta$  di ottiene lo stimatore del metodo dei momenti per  $\theta = (\alpha, \beta)$  dato da

$$\left( \frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2 / n}, \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n\bar{X}_n} \right) \quad (2.25)$$

Vale la pena osservare che, con alcuni semplici passaggi, si può scrivere la funzione di densità della distribuzione Gamma nei termini richiesti per l'appartenenza a famiglia esponenziale a  $k = 2$ -parametri; infatti

$$f(x; \alpha, \beta) = \mathbb{1}_{\mathbb{R}^+}(x) \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp\left\{(\alpha - 1) \ln(x) - \frac{1}{\beta}x\right\} \quad (2.26)$$

sicché in presenza di un campione casuale di ampiezza  $n$  da una distribuzione Gamma di parametri  $(\alpha, \beta)$  la *statistica naturale* sarà data da

$$\left( \sum_{i=1}^n \ln(X_i), \sum_{i=1}^n X_i \right) \quad (2.27)$$

da cui si possono ricavare i seguenti *momenti caratterizzanti* della distribuzione Gamma:

$$\left( \frac{1}{n} \sum_{i=1}^n \ln(X_i), \frac{1}{n} \sum_{i=1}^n X_i \right) \quad (2.28)$$

che, come si può facilmente osservare, non coincidono con i primi due momenti interi  $m'_1$  e  $m'_2$  su cui poggia la stima dei parametri basata sul metodo dei momenti: la mancata coincidenza tra momenti "caratterizzanti" e momenti interi avrà delle importanti ripercussioni in termini di *bontà* della stima dei parametri della distribuzione Gamma tramite metodo dei momenti e, in ultima analisi, sulla attendibilità del conseguente processo inferenziale a essi relativo.

## 2.2 Statistiche ordinate

Cominciamo col dare la definizione di ciò che intendiamo con *statistiche ordinate*.

**Definizione 2.2.1.** Sia  $(X_1, \dots, X_n)$  un campione casuale con distribuzione  $F_X(x, \theta)$ , densità  $f_X(x, \theta)$  e supporto  $\mathcal{S}_X := (a, b) \subseteq \mathbb{R}$  ove  $X \in \{X_1, \dots, X_n\}$  e  $-\infty \leq a < b \leq +\infty$ . Definiamo ricorsivamente le seguenti variabili casuali:

- $X_{(1)} := \min(\{X_1, \dots, X_n\})$
- $X_{(m)} := \min(\{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(m-1)}\}), \forall 1 < m \leq n.$

Chiameremo allora  $X_{(m)}$  la  $m$ -ma *statistica ordinata* del campione.

**Osservazione:** La statistica ordinata consiste semplicemente nel vettore per il quale le variabili casuali vengono appunto ordinate in base al valore che assumono in un determinato punto del loro dominio comune, usualmente in *ordine crescente*. In particolare  $X_{(m)}$  sarà la  $m$ -esima variabile più piccola. Naturalmente, se il campione ha lunghezza  $n$ , allora  $X_{(n)} = \max(\{X_1, \dots, X_n\})$ . Osserviamo che la funzione

$$(X_1, \dots, X_n) \mapsto (X_{(1)}, \dots, X_{(n)})$$

è essa stessa una statistica e contiene la *stessa* quantità di informazione sul parametro  $\theta$  o su qualsiasi altro aspetto della popolazione da cui proviene il campione del campione casuale  $(X_1, X_2, \dots, X_n)$  stesso.

**Teorema 2.2.1** (Distribuzione della  $m$ -ma statistica ordinata). Sia  $X$  una v.c. (assolutamente) continua avente funzione di distribuzione  $F_X(x; \theta)$  cui corrisponde la funzione di densità  $f_X(x; \theta)$  e sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da  $F_X(x; \theta)$ . La funzione di densità della  $m$ -ma statistica ordinata campionaria  $X_{(m)}$ ,  $m = 1, 2, \dots, n$ , è data da:

$$f_{X_{(m)}}(x; \theta) = \binom{n}{m} m [F_X(x; \theta)]^{m-1} [1 - F_X(x; \theta)]^{n-m} f_X(x; \theta).$$

**Dimostrazione.** Per semplicità di notazione, nel seguito, ometteremo il parametro  $\theta$  dalla notazione delle funzioni di distribuzione e delle funzioni di densità, scrivendo  $F_X(x)$  in luogo di  $F_X(x; \theta)$  e  $f_X(x)$  in luogo di  $f_X(x; \theta)$ .

Sia

$$Y = \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i)$$

il numero degli elementi del campione casuale in cui valore è inferiore o al più uguale a  $x$ ; allora

$$Y \sim b(n, p = F_X(x))$$

Perciò, essendo l'evento  $\{X_{(m)} \leq x\}$  equivalente all'evento  $\{Y \geq m\} = \{\text{il numero degli elementi del campione casuale minori o uguali a } x \text{ è almeno pari a } m\}$ , si avrà

$$F_{X_{(m)}}(x) = P_\theta(Y \geq m) = \sum_{k=m}^n \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}.$$

Ricordiamo ora che  $X$  è una v.c. assolutamente continua e, pertanto, lo sarà pure  $X_{(m)}$ ; perciò

$$f_{X_{(m)}}(x) = \frac{d}{dx} F_{X_{(m)}}(x)$$

ovvero

$$\begin{aligned} f_{X_{(m)}}(x) &= \binom{n}{m} m [F_X(x)]^{m-1} [1 - F_X(x)]^{n-m} f_X(x) + \\ &\quad + \sum_{k=m+1}^n \binom{n}{k} k [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x) + \\ &\quad - \sum_{k=m}^{n-1} \binom{n}{k} (n-k) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x) \end{aligned} \quad (2.29)$$

tenuto conto del fatto che l'ultimo addendo della terza somma in (2.29) è nullo per  $k = n$  e, pertanto, quest'ultima può essere arrestata a  $n - 1$ . Inoltre, le due somme presenti in (2.29) si elidono; infatti, procedendo a un cambiamento di variabili nella prima delle due somme, ponendo  $h = k - 1$ , si ha

$$\begin{aligned} &\sum_{k=m+1}^n \binom{n}{k} k [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x) = \\ &= \sum_{h=m}^{n-1} \binom{n}{h+1} (h+1) [F_X(x)]^h [1 - F_X(x)]^{n-(h+1)} f_X(x) \\ &= \sum_{h=m}^{n-1} \binom{n}{h} (n-h) [F_X(x)]^h [1 - F_X(x)]^{n-(h+1)} f_X(x) \end{aligned}$$

essendo

$$\begin{aligned} \binom{n}{h+1} (h+1) &= \frac{n!}{(h+1)! [n-(h+1)]!} (h+1) \\ &= \frac{n!}{h! (n-h)!} (n-h) \\ &= \binom{n}{h} (n-h). \end{aligned}$$

Ma allora (reintroducendo il parametro  $\theta$  nella notazione della funzione di densità)

$$f_{X_{(m)}}(x; \theta) = \binom{n}{m} m [F_X(x; \theta)]^{m-1} [1 - F_X(x; \theta)]^{n-m} f_X(x; \theta)$$

risulta essere la funzione di densità della  $m$ -ma statistica ordinata  $X_{(m)}$  per  $m = 1, 2, \dots, n$ .  $\square$

A partire dalle statistiche ordinate è possibile costruire altre statistiche che forniscono informazioni su aspetti peculiari della popolazione da cui proviene il campione su cui sono state calcolate.

### 2.2.1 I cinque numeri magici e l'analisi esplorativa dei dati

Nelle primissime pagine di questo capitolo abbiamo sottolineato il fatto che statistica consiste in una serie di operazioni legate alla *sintesi* dei dati e alla *conservazione dell'informazione rilevante* presente in essi.

I *cinque numeri magici* che tra poco introdurremo si occupano proprio di questo e costituiscono il fulcro di quella che è nota come *analisi esplorativa dei dati*; essi, come avremo modo di vedere, altro non sono che particolari statistiche ordinate.

Sia  $(X_{(1)}, \dots, X_{(n)})$  l'insieme delle statistiche ordinate di un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da una popolazione avente funzione di distribuzione  $F_X(x; \theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}^k$ ,  $k \in \mathbb{N}$ . Allora possiamo definire le seguenti nuove statistiche (variabili casuali):

- *Range* campionario (o ) di  $X$

$$R(X_1, X_2, \dots, X_n) = X_{(n)} - X_{(1)}$$

che fornisce una prima, seppur grezza, *misura di dispersione o variabilità*;

- *Mid Range* campionario

$$MR(X_1, X_2, \dots, X_n) = \frac{X_{(1)} + X_{(n)}}{2}$$

che fornisce informazione sul centro della distribuzione ovvero una prima *misura di centralità*;

- *Mediana* campionaria

$$Me(X_1, X_2, \dots, X_n) = \begin{cases} X_{(\frac{n+1}{2})}, & \forall n \text{ dispari} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & \forall n \text{ pari} \end{cases} \quad (2.30)$$

Come si può facilmente osservare dalla seconda in (2.30), in presenza di un campione casuale con un numero  $n$  pari di elementi, e di conseguenza due posizioni mediane, la mediana campionaria risulterà dalla semi-somma dei due elementi che occupano dette posizioni.

- *Quantile* campionario di *ordine p*

Sia  $\xi_p \in F_X^{-1}(p)$  per  $p \in (0, 1)$ ; chiameremo  $\xi_p$  *quantile della popolazione* di ordine  $p$ . Quest'ultimo può essere stimato via *plug-in* nella maniera che segue. Possiamo definire l'intero

$$k_p := \lfloor p(n+1) \rfloor \quad (2.31)$$

che risulta essere compreso tra 1 e  $n$ , al variare di  $p$ , e restituisce un'approssimazione all'intero più vicino al reale  $p(n+1)$ , vale a dire alla posizione occupata dal  $p$ -mo quantile nell'insieme dei dati campionari, ordinati in senso crescente.

La statistica

$$Q_p^*(X_1, X_2, \dots, X_n) = X_{(k_p)} \quad (2.32)$$

è detta *quantile campionario* di ordine  $p$  ed è la naturale stima *plug-in* del percentile di ordine  $p$  della popolazione,  $\xi_p$ .

In molte applicazioni siamo interessati a quantili di un particolare ordine e, più precisamente, di ordine  $p = 0.25$ ,  $p = 0.50$  e  $p = 0.75$  cui daremo rispettivamente i nomi di *primo*, *secondo* e *terzo quartile* il cui significato è presto detto: il primo quartile  $Q_1 = Q_{0.25}^*(X_1, X_2, \dots, X_n)$  è quel valore che lascia alla sua sinistra il 25% delle osservazioni campionarie, il secondo quartile  $Q_2 = Q_{0.50}^*(X_1, X_2, \dots, X_n) = Me(X_1, X_2, \dots, X_n)$  è quel valore che lascia alla sua sinistra il 50% delle osservazioni campionarie e infine il terzo quartile  $Q_3 = Q_{0.75}^*(X_1, X_2, \dots, X_n)$  è quel valore che lascia alla sua sinistra il 75% delle osservazioni campionarie e risulteranno fondamentali in quella che va sotto il nome di *analisi esplorativa dei dati* in quanto costituiscono una sorta di *pietre miliari* che suddividono l'insieme *ordinato* dei valori della distribuzione o del campione in quattro parti di uguale numerosità.

Siamo ora in grado di costruire l'insieme dei *cinque numeri magici* dell'*analisi esplorativa dei dati*: *massimo*, *minimo*, *primo*, *secondo* e *terzo quartile*; questi trovano una sintesi grafica in un interessante strumento: il *box-plot*. La costruzione del *box-plot* poggia sui citati numeri magici e sulle due seguenti quantità rispettivamente note come *lower fence* e *upper fence*

$$\begin{aligned} LF &= \max \left( X_{(1)}, Q_1 - \frac{3}{2} \cdot IQR \right) \\ UF &= \min \left( X_{(n)}, Q_3 + \frac{3}{2} \cdot IQR \right) \end{aligned} \quad (2.33)$$

dove  $IQR = (Q_3 - Q_1)$  è la *distanza interquartilica* (o *Inter-Quartile Range*).

*Lower fence* e *upper fence* giocano un ruolo fondamentale nell'individuare osservazioni discordanti con quello che è il corpo centrale dei dati e che potenzialmente possono costituire *outlier*, valori potenzialmente pericolosi per le misure di sintesi in quanto possono alterarne profondamente il valore. Più precisamente, i valori che stanno a sinistra di  $LF$  e a destra di  $UF$  sono da considerare *outlier* e vanno debitamente trattati nel calcolo delle misure di sintesi.

**Esempio 2.2.1.** Consideriamo il seguente set di  $n = 15$  osservazioni relative alla v.c.  $X$  che abbiamo preventivamente *ordinato in senso crescente*.

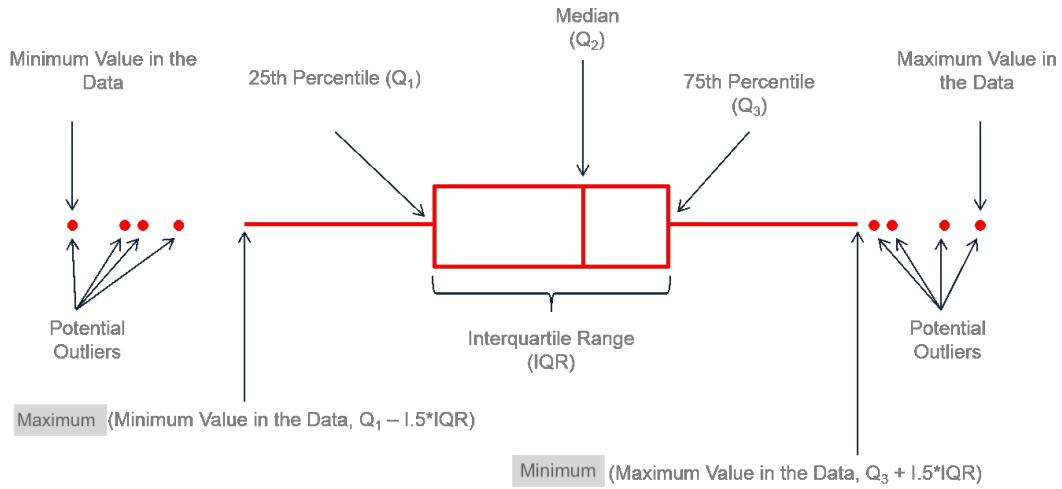


Figura 2.2: I cinque numeri magici e Box-plot

Oss.ne	Valore di $X$
1	56
2	70
3	89
<b>4</b>	<b>94</b>
5	96
6	101
7	102
<b>8</b>	<b>102</b>
9	102
10	105
11	106
<b>12</b>	<b>108</b>
13	110
14	113
15	116

A partire da questi dati calcoliamo subito media e varianza:  $\bar{x}_n = 98.0$  e  $s_n^2 = 258.0$ .

Ora procediamo nella costruzione del box-plot; a questo scopo ricaviamo subito valore *minimo*  $x_{(1)} = 56$  e valore *massimo*  $x_{(15)} = 116$  e osserviamo che, per  $n = 15$  (*dispari*), le posizioni occupate dal primo, secondo e terzo quartile (valori che equiripartiscono l'insieme dei dati in quattro parti di uguale numerosità) sono rispettivamente date da

$$Pos(Q_1) : \frac{1}{4} (15 + 1) = 4, \quad Pos(Q_2) : \frac{2}{4} (15 + 1) = 8, \quad Pos(Q_3) : \frac{3}{4} (15 + 1) = 12$$

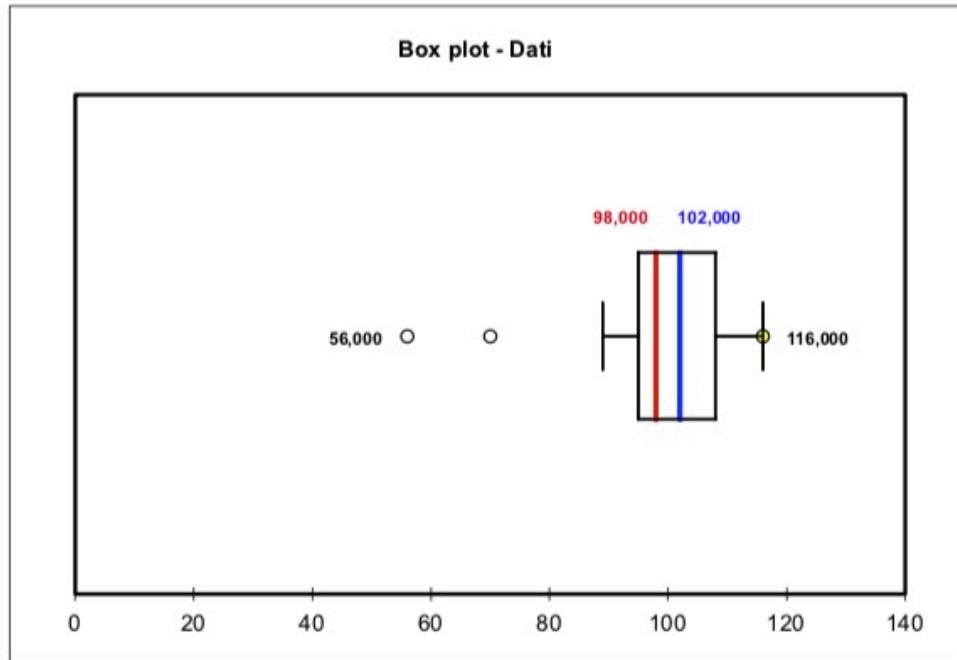
e di conseguenza,

$$Q_1 = x_{(4)} = 94, \quad Q_2 = x_{(8)} = 102, \quad Q_3 = x_{(12)} = 108$$

Possiamo ora trovare i valori di Lower e Upper fence rispettivamente dati da

$$\begin{aligned} LF &= \max \left( 56, 94 - \frac{3}{2} \cdot (108 - 94) \right) = \max(54, 73) = 73 \\ UF &= \min \left( 116, 108 + \frac{3}{2} \cdot (108 - 94) \right) = \min(116, 129) = 116 \end{aligned} \quad (2.34)$$

Ora abbiamo tutti gli elementi per costruire il *box-plot* per i dati in questione.



**Figura 2.3:** Box-plot per le  $n = 15$  osservazioni

Mediana e media campionarie sono rispettivamente indicate con il segmento in blu e con il segmento in rosso; dall'esame del box-plot emerge la presenza di due outlier (valori 56 e 70) ed è subito evidente il loro effetto sulla media e sulla varianza dei dati in questione se ne ricalcoliamo i valori *escludendo* questi due outlier

$$\bar{x}_n = 103.8 \quad \text{e} \quad s_n^2 = 57.26 \quad (2.35)$$

e come possiamo subito notare, la media si è ora avvicinata al valore mediano che risulta più *robusto* (ovvero non influenzabile dalla presenza di outlier perché poggia il suo valore sulla posizione e non sui valori) di quanto non sia la media stessa mentre la dispersione dei dati è decisamente diminuita (si confrontino i valori delle varianze *prima* e *dopo* l'eliminazione degli outlier).

**Osservazione:** più la media si *allontana* dalla mediana, più vi è *asimmetria* nella distribuzione. Infatti una distribuzione è *simmetrica* se la sua funzione di densità (analogo discorso può essere fatto per la funzione di massa) soddisfa a

$$\exists x_0 \in \mathbb{R} : f(x_0 + x) = f(x_0 - x), \forall x \in \text{Dom}(f). \quad (2.36)$$

Ora, ponendo che la funzione di ripartizione sia iniettiva e la funzione di densità sia simmetrica, si vede immediatamente che la mediana della popolazione, ovvero il quantile di popolazione di ordine  $p = \frac{1}{2}$ , coincide con il media della distribuzione della v.c., la quale a sua volta deve coincidere con  $x_0$ .

In definitiva è' importante notare una cosa: se fatte *dialogare tra loro*, le misure di sintesi (qui per esempio, *media* e *mediana*) possono fornire informazioni su aspetti della distribuzione (per esempio, la *simmetria*) che, prese singolarmente, non riescono a cogliere in quanto non sono state pensate per coglierli. Dunque, l'estrazione dell'*informazione rilevante*, da cui dipende la *bontà* dell'analisi svolta sui dati, passa quindi anche attraverso la capacità di far *interagire* tra loro le singole misure di sintesi.

### 3 Non distorsione, consistenza e distribuzioni limite

In probabilità e in statistica è spesso necessario considerare la distribuzione di una v.c.  $Y$  che è, a sua volta, una funzione di una o più v.c., ovvero  $Y = g(X_1, X_2, \dots, X_n)$ ; sfortunatamente, trovare la *distribuzione esatta* di  $Y$  può essere talvolta molto difficile, complicato o assai laborioso, anche quando la forma funzionale della distribuzione di  $(X_1, X_2, \dots, X_n)$  sia nota. O addirittura impossibile quando abbiamo solo una parziale informazione circa la distribuzione congiunta di  $(X_1, X_2, \dots, X_n)$  ed è perciò impossibile determinare la distribuzione di  $Y$  tramite gli usuali teoremi di trasformazione.

Tuttavia, quando l'ampiezza campionaria  $n$  è sufficientemente grande, è possibile ottenere un'approssimazione della distribuzione esatta di  $Y$  anche in presenza di informazione parziale sulla distribuzione di  $(X_1, X_2, \dots, X_n)$ . Tali approssimazioni possono essere decisamente accurate e poggiano su opportuni processi di convergenza.

In particolare considereremo l'idea - piuttosto fantasiosa se volete - di permettere all'ampiezza campionaria  $n$  di tendere a infinito e studieremo il comportamento delle statistiche al succedere di ciò. Sebbene la nozione di *ampiezza campionaria infinita* sia un puro artificio teorico, essa spesso può fornirci qualche utile approssimazione per il caso di ampiezza campionaria finita perché spesso accade che le espressioni analitiche delle distribuzioni, al limite, diventino più semplici.

Vale la pena osservare che in ambito statistico  $Y_n = g(X_1, X_2, \dots, X_n)$  è spesso una statistica coinvolta in processi inferenziali (stima puntuale o intervallare, test di ipotesi) riguardanti una quantità incognita  $\theta = (\theta_1, \dots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$  o sue funzioni  $g(\theta)$ , vale a dire statistiche quali possono essere estimatori o statistiche test per  $\theta$  o  $g(\theta)$ , le cui proprietà (ottimali) frequentemente dipendono dall'ampiezza campionaria  $n$ . Di conseguenza,  $\{Y_n\}_{n \in \mathbb{N}}$  è una successione di statistiche e, usualmente, siamo in grado di studiare le proprietà di  $Y_n$  solo asintoticamente ossia quando  $n \rightarrow \infty$ .

Quello in cui, in ultima analisi, confidiamo è che lo studio delle proprietà di  $\{Y_n\}_{n \in \mathbb{N}}$ , e in particolare del limite di  $\{Y_n\}_{n \in \mathbb{N}}$ , finisce per essere illuminante in merito alle proprietà statistiche di  $Y_n$  confidando nel fatto che il *comportamento* di  $Y_n$  su un grande campione (quale quello che stiamo effettivamente osservando) e il *comportamento* dello stesso  $Y_n$  su di un campione di ampiezza infinita (quello che analizziamo prendendo il limite di  $Y_n$ ) sia lo stesso.

Lo studio dei diversi tipi di convergenza costituisce il nucleo della *teoria asintotica* che costituisce uno dei cardini della statistica matematica.

Vi sono molti differenti tipi di processi di convergenza per successioni  $\{Y_n\}_{n \in \mathbb{N}}$  di v.c. che sostanzialmente dipendono da come si misura la distanza tra gli elementi  $Y_n$  della successione  $\{Y_n\}_{n \in \mathbb{N}}$  e l'elemento limite  $Y$ .

Ci occuperemo essenzialmente di tre tipi di convergenza: la *convergenza in probabilità*, la *convergenza in distribuzione* e la *convergenza in media quadratica*. Diremo, infine, qualcosa anche in merito alla *convergenza delle funzioni generatrici dei momenti*.

## 3.1 Convergenza in probabilità

Cominciamo con il richiamare due importanti disuguaglianze in ambito probabilistico: la *disuguagliaza di Markov* e la *disuguagliaza di Chebychev*.

**Teorema 3.1.1** (Disuguagliaza di Markov). Sia  $X$  una variabile casuale non negativa. Allora per un qualunque valore  $\varepsilon > 0$  vale che

$$P(X \geq \varepsilon) \leq \frac{\mathbb{E}(X)}{\varepsilon} \quad (3.1)$$

*Dimostrazione.* Supponiamo  $X \sim F_X(x; \theta)$  con densità di probabilità  $f_X(x; \theta)$ . Calcoliamo il valore atteso di tale variabile casuale:

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{+\infty} x f_X(x; \theta) \mathbb{1}_{S_X}(x) dx \\ &= \int_{-\infty}^{\varepsilon} x f_X(x; \theta) \mathbb{1}_{S_X}(x) dx + \int_{\varepsilon}^{+\infty} x f_X(x; \theta) \mathbb{1}_{S_X}(x) dx \\ &\quad \left| \begin{array}{l} \text{poiché entrambi questi integrali sono certamente non negativi,} \\ \text{essendo la variabile casuale definita non negativa, possiamo dire} \\ \text{che} \end{array} \right. \\ &\geq \int_{\varepsilon}^{+\infty} x f_X(x; \theta) \mathbb{1}_{S_X}(x) dx \\ &\geq \varepsilon \int_{\varepsilon}^{+\infty} f(x; \theta) dx = \varepsilon P(X \geq \varepsilon) \end{aligned}$$

Dalla disuguagliaza ottenuta discende naturalmente la tesi del teorema. □

**Teorema 3.1.2** (Disuguagliaza di Chebychev). Sia  $X$  una variabile casuale di cui si conoscono la media  $\mu$  e la varianza  $\sigma^2$ , entrambe finite. Allora, dato un qualunque valore  $\varepsilon > 0$  vale la disuguagliaza

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\mathbb{V}ar(X)}{\varepsilon^2} \quad (3.2)$$

*Dimostrazione.* Si consideri l'evento per cui vale  $|X - \mu| \geq \varepsilon$ : i valori di  $X, \mu$  e  $\varepsilon$  per cui vale la disuguagliaza sono gli stessi per cui vale la disuguagliaza  $(X - \mu)^2 \geq \varepsilon^2$ . La probabilità che si verifichi una delle precedenti disuguaglianze è la stessa; inoltre, la variabile casuale  $(X - \mu)^2$  è non negativa. Per la disuguagliaza di Markov si ha che

$$P(|X - \mu| \geq \varepsilon) = P((X - \mu)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{\varepsilon^2} = \frac{\mathbb{V}ar(X)}{\varepsilon^2} \quad (3.3)$$

La tesi è dunque dimostrata. □

Il concetto di convergenza in probabilità si basa sulla seguente intuizione: due variabili casuali sono *vicine l'una all'altra* se vi è un'alta probabilità che la loro *differenza* sia molto *piccola*.

**Definizione 3.1.1** (Convergenza in probabilità). Sia  $\{Y_n\}_{n \in \mathbb{N}}$  una successione di variabili casuali definite su uno spazio campionario  $\Omega$  e sia  $Y$  una variabile casuale definita anch'essa sul medesimo spazio campionario. Diciamo che  $Y_n$  converge in probabilità a  $Y$  se e solo se

$$\lim_{n \rightarrow +\infty} P(|Y_n - Y| \geq \varepsilon) = 0, \quad \forall \varepsilon > 0 \quad (3.4)$$

e scriveremo  $Y_n \xrightarrow{P} Y$ .

Risulta significativo svolgere le seguenti osservazioni.

- le v.c.  $Y_1, Y_2, \dots, Y_n, \dots$  della definizione 3.1.1 (e delle altre definizioni che daremo in seguito trattando di processi di convergenza) **non** sono necessariamente i.i.d. come accade in un campione casuale.

La distribuzione di  $Y_n = g(X_1, X_2, \dots, X_n)$  *cambia al cambiare* dell'indice  $n$  e ciò che faremo sarà studiare le diverse maniere in cui la distribuzione di  $Y_n$  si *avvicina* (vale a dire, *converge*) alla distribuzione del limite  $Y$  della successione  $\{X_n\}_{n \in \mathbb{N}}$  al crescere di  $n$

- $\{P(|Y_n - Y| \geq \varepsilon)\}_{n \in \mathbb{N}}$  è una successione di numeri reali. Pertanto, il limite implicato nella definizione di convergenza in probabilità è l'usuale limite di una successione di numeri reali.
- giusto per avere un'interpretazione fisica del processo di convergenza in probabilità, possiamo dire che se  $Y_n \xrightarrow{P} Y$  allora la *massa associata alla differenza*  $|Y_n - Y|$  converge a zero
- in statistica spesso il limite  $Y$  è una *costante* cioè una v.c. *degenera* (per esempio, un *parametro* o una funzione di parametro) con tutta la sua massa concentrata in qualche punto  $a \in \mathbb{R}$ . In questo caso scriviamo  $Y_n \xrightarrow{P} a$ .

**Teorema 3.1.3** (LDGN). Sia  $(X_1, X_2, \dots, X_n, \dots)$  una successione di variabili indipendenti e identicamente distribuite, e supponiamo che le  $X_i$  ammettano momento secondo. Definita  $Y_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , si ha

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \geq \varepsilon) = 0, \quad \text{per ogni } \varepsilon > 0 \quad (3.5)$$

In altre parole,  $\bar{X}_n \xrightarrow{P} \mu$ .

*Dimostrazione.* Immediata via Chebychev. Sia  $\mu = E[X_n]$  e  $\sigma^2 = \text{Var}([X_n])$ . Per la linearità del valor medio,  $\bar{X}_n = \mu$ . È inoltre immediato verificare che  $\text{Var}(\bar{X}_n) = \sigma^2/n$ ; infatti

$$\mathbb{E}[(\bar{X}_n - \mu)^2] = \text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2 n}{n^2} = \frac{\sigma^2}{n} \quad (3.6)$$

Applichiamo ora la disuguaglianza di Chebychev e prendiamone il limite per  $n \rightarrow \infty$  su entrambi i versanti della disuguaglianza

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0 \quad (3.7)$$

ottenendo infine la tesi del teorema.  $\square$

Ovviamente l'utilità della *legge debole dei grandi numeri* risulta evidente quando siamo interessati a fare inferenza sulla media  $\mu$  di una popolazione (o sulla probabilità  $\theta$  di successo di una popolazione bernoulliana) basandoci sul suo stimatore *plug-in*  $\bar{X}_n$ .

La proprietà riassunta dalla LDGN e relativa al fatto che la successione delle statistiche (stimatori di  $\mu$ )  $\bar{X}_n$  si avvicina sempre più alla costante  $\mu$ , media della popolazione (o a  $\theta$ , probabilità di successo) al crescere di  $n$ , è nota come *consistenza*.

**Definizione 3.1.2** (Stimatore consistente). Sia  $X$  una v.c. avente funzione di ripartizione  $F_X(x; \theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}^k$  e sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da quest'ultima distribuzione. Sia  $T_n = T_n(X_1, X_2, \dots, X_n)$  uno stimatore di  $\theta$ . Diremo  $T_n$  stimatore (*debolmente o semplicemente*) *consistente* di  $\theta$  se

$$T_n \xrightarrow{P} \theta. \quad (3.8)$$

**Definizione 3.1.3** (Stimatore non distorto). Sia  $X$  una v.c. avente funzione di ripartizione  $F_X(x; \theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}^k$  e sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da quest'ultima distribuzione. Sia  $T_n = T_n(X_1, X_2, \dots, X_n)$  uno stimatore di  $\theta$ . Diremo  $T_n$  stimatore *non distorto* di  $\theta$  se

$$\mathbb{E}_\theta(T_n) = \theta, \forall \theta \in \Theta. \quad (3.9)$$

### Osservazione

- i) Guardando alle due definizioni 3.1.2 e 3.1.3 si nota immediatamente che la *consistenza* è una proprietà della successione degli stimatori quindi una proprietà *asintotica* mentre la *non distorsione* è quel che si dice una proprietà *finita* ovvero, laddove presente, vale per ogni valore di  $n$ .
- ii) laddove non valga la (3.9) diremo lo stimatore *distorto* e chiameremo la quantità

$$\mathbb{B}_\theta(T_n) = \mathbb{E}_\theta(T_n) - \theta \quad (3.10)$$

*distorsione* di  $T_n$

- iii) considerata la successione  $\{T_n\}_{n \in \mathbb{N}}$  degli stimatori di  $\theta$ , qualora

$$\lim_{n \rightarrow \infty} \mathbb{B}_\theta(T_n) = 0 \quad (3.11)$$

diremo lo stimatore  $T_n$  *asintoticamente non distorto* per  $\theta$ .

**Esempio 3.1.1** (Media campionaria). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una qualche distribuzione e supponiamo di essere interessati a fare inferenza in merito alla media  $\mu = \mathbb{E}(X)$  della distribuzione stessa. E' immediato pensare allo stimatore *media campionaria*  $\bar{X}_n$  dal momento che è uno stimatore di  $\mu$  *accurato*

(sinonimo di *non distorto*) perché il suo valore atteso coincide con  $\mu$  e pure *preciso* in quanto la sua varianza tende a 0 al crescere di  $n$ .

Se consideriamo al contrario la quantità  $W = (X_1 + X_2)/2$ , è facile verificare che tale quantità non è un buon stimatore di  $\mu$ . Infatti, si tratta di uno stimatore accurato in quanto  $E(W) = \mu$ ,  $\forall \mu \in \mathbb{R}$  ma non preciso poiché  $\text{Var}(W) = \sigma^2/2$ ,  $\sigma^2 \in \mathbb{R}^+$ . Di conseguenza, l'aumentare dell'ampiezza campionaria  $n$  non influisce sulla precisione della stima di  $\mu$ . E questa è una gran brutta cosa: al crescere dell'ampiezza campionaria, cresce il contenuto informativo del campione casuale relativo all'aspetto (parametro o funzione del parametro) su cui si intende fare inferenza. Ma è evidente che uno stimatore quale  $W$  non riesce ad "approfittare" di ciò e questa è la ragione per cui non siamo interessati a stimatori che non siano consistenti.

**Esempio 3.1.2** (Frequenza relativa campionaria). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale con distribuzione  $b(1, p)$ . Vogliamo fare inferenza su  $p$ . Applicando il principio del plug-in, possiamo stimare  $p$  mediante la quantità

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (3.12)$$

Lo stimatore  $\hat{p}_n$  è uno stimatore non distorto di  $p$  poiché  $E_n(\hat{p}_n) = p$ ,  $\forall p \in [0, 1]$  e inoltre  $\text{Var}(\hat{p}_n) = p(1-p)/n$ . Applicando la diseguaglianza di Chebychev, otteniamo

$$P(|\hat{p}_n - p| \geq \varepsilon) \leq \frac{\text{Var}(\hat{p}_n)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0 \quad (3.13)$$

e di conseguenza, possiamo dire che  $\hat{p}_n$  converge in probabilità a  $p$  o che è uno stimatore consistente del parametro  $p$ , probabilità di successo nell' singola prova bernoulliana.

**Esempio 3.1.3** (Varianza campionaria). Sia  $\{X_i\}_{i \in \mathbb{N}}$  una successione di variabili casuali indipendenti identicamente distribuite con  $E(X_i) = \mu$  e  $\text{Var}(X_i) = \sigma^2$ , entrambe finite. Sia

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (3.14)$$

la varianza campionaria calcolata sulle prime  $n$  variabili casuali della successione. Ora  $S_n^2$  è uno stimatore non distorto di  $\sigma^2$ ; infatti,

$$\begin{aligned} E(S_n^2) &= \mathbb{E} \left[ \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right] \\ &= \frac{1}{n-1} \mathbb{E} \left[ \sum_{i=1}^n X_i^2 - n \bar{X}_n^2 \right] \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n \mathbb{E}(X_i^2) - n \mathbb{E}(\bar{X}_n^2) \right] \\ &= \frac{1}{n-1} \left[ n(\sigma^2 + \mu^2) - n \left( \frac{\sigma^2}{n} + \mu^2 \right) \right] \\ &= \frac{1}{n-1} [n\sigma^2 - \sigma^2] = \sigma^2, \quad \forall \sigma^2 > 0 \end{aligned} \quad (3.15)$$

poiché  $\mathbb{E}(X_i^2) = \text{Var}(X_i) + \mu^2 = \sigma^2 + \mu^2$  mentre  $\mathbb{E}(\bar{X}_n^2) = \text{Var}(\bar{X}_n) + \mu^2 = \sigma^2/n + \mu^2$ . Al contrario, lo stimatore *plug-in*

$$\hat{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (3.16)$$

è uno stimatore distorto per  $\sigma^2$ ; infatti, se calcoliamo il suo valore atteso in maniera analoga a quanto fatto per  $S_n^2$ , otteniamo

$$E(\hat{S}_n^2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2 \quad (3.17)$$

da cui la distorsione

$$B_{\sigma^2}(\hat{S}_n^2) = -\frac{1}{n} \sigma^2 \quad (3.18)$$

distorsione che però si risolve al crescere di  $n$  dal momento che la quantità (3.18) tende a 0 al crescere dell'ampiezza campionaria  $n$ . Possiamo dunque dire che lo stimatore *plug-in*  $\hat{S}_n^2$  è comunque *asintoticamente non distorto*.

È altresì noto che la varianza dello stimatore varianza campionaria  $S_n^2$  può essere scritta, quale che sia la distribuzione da cui proviene il campione casuale su cui è calcolata, in funzione dei momenti centrati nella media di quest'ultima secondo la relazione:

$$\text{Var}(S_n^2) = \frac{1}{n} \left[ \mu_4 - \frac{n-3}{n-1} \mu_2^2 \right] \quad (3.19)$$

della quale si può trovare la dimostrazione nell'appendice alla fine di questo capitolo. Possiamo dunque applicare la diseguaglianza di Chebychev per verificare se  $S_n^2$  converge in probabilità a  $\sigma^2$ :

$$P[|S_n^2 - \sigma^2| \geq \varepsilon] \leq \frac{E[(S_n^2 - \sigma^2)^2]}{\varepsilon^2} = \frac{\text{Var}(S_n^2)}{\varepsilon^2} \xrightarrow[n \rightarrow \infty]{} 0 \quad (3.20)$$

Concludiamo che  $S_n^2$  converge in probabilità a  $\sigma^2$  e pertanto anche la varianza campionaria  $S_n^2$  è uno stimatore (debolmente) consistente della varianza  $\sigma^2$  della popolazione. Si può facilmente verificare che anche lo stimatore *plug-in*  $\hat{S}_n^2$  è uno stimatore consistente per la varianza  $\sigma^2$  della popolazione; infatti,

$$\begin{aligned} \text{Var}(\hat{S}_n^2) &= \text{Var}\left(\frac{n-1}{n} S_n^2\right) \\ &= \left[\frac{n-1}{n}\right]^2 \text{Var}(S_n^2) \\ &= \frac{(n-1)^2}{n^3} \left[ \mu_4 - \frac{n-3}{n-1} \mu_2^2 \right] \end{aligned} \quad (3.21)$$

quantità che tende a zero al crescere di  $n$  ergo anche  $\hat{S}_n^2$  è uno stimatore consistente per la varianza  $\sigma^2$  della popolazione.

### 3.1.1 Alcuni utili risultati relativi alla convergenza in probabilità

Vi sono una serie di risultati riguardanti la convergenza in probabilità che sono estremamente utili nelle applicazioni. Vediamone qualcuno (senza darne la dimostrazione).

**Teorema 3.1.4.** Supponiamo che  $X_n \xrightarrow{P} X$  e che  $Y_n \xrightarrow{P} Y$ . Allora

$$X_n + Y_n \xrightarrow{P} X + Y$$

**Teorema 3.1.5.** Supponiamo che  $X_n \xrightarrow{P} X$  e sia  $a$  una costante. Allora

$$aX_n \xrightarrow{P} aX$$

In buona sostanza, i teoremi 3.1.4 e 3.1.5 affermano che la *convergenza in probabilità* è chiusa sotto *linearità*.

**Teorema 3.1.6** (Continuous mapping). Sia  $X_n \xrightarrow{P} a$  e sia  $g$  una funzione reale continua in  $a$ . Allora

$$g(X_n) \xrightarrow{P} g(a)$$

**Corollario 3.1.1.** Sia  $X_n \xrightarrow{P} a$  con  $a$  costante. Allora valgono i seguenti limiti in probabilità:

- $X_n^2 \xrightarrow{P} a^2$
  - $\frac{1}{X_n} \xrightarrow{P} \frac{1}{a}$ , purché  $a \neq 0$
  - $\sqrt{X_n} \xrightarrow{P} \sqrt{a}$ , purché  $a \geq 0$
- (3.22)

**Teorema 3.1.7.** Supponiamo che  $X_n \xrightarrow{P} X$  e  $Y_n \xrightarrow{P} Y$ . Allora

$$X_n \cdot Y_n \xrightarrow{P} X \cdot Y$$

**Esempio 3.1.4.** Consideriamo  $Y_n \sim b(n, p)$  con  $Y_n = \sum Y_i$  e  $Y_i \sim b(1, p)$ . Per i teoremi enunciati precedentemente, e per quanto noto sulla distribuzione binomiale, valgono i seguenti limiti in probabilità:

$$\frac{Y_n}{n} \xrightarrow{P} p \quad 1 - \frac{Y_n}{n} \xrightarrow{P} 1 - p \quad \frac{Y_n}{n} \left(1 - \frac{Y_n}{n}\right) \xrightarrow{P} p(1 - p) \quad (3.23)$$

Consideriamo ora un altro esempio, interessante in quanto introduce dei nuovi oggetti utili nell'inferenza quali sono le *statistiche ordinate*; queste altro non sono che il frutto della riorganizzazione degli elementi del campione in ordine crescente indicando quindi con  $X_{(1)} = \min(X_1, X_2, \dots, X_n)$ , con  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$  e con  $X_{(m)}$ ,  $m = 2, 3, \dots, (n-1)$  i rimanenti  $(n-1)$  elementi di  $(X_1, X_2, \dots, X_n)$  disposti in ordine crescente a partire dal minimo.

**Esempio 3.1.5** ( $n - ma$  statistica ordinata). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una popolazione avente distribuzione Uniforme su  $[0, \theta]$  con  $\theta \in \Theta$  parametro non noto.

Il problema inferenziale che ora affronteremo sarà quello della stima di  $\theta$  a partire dall'informazione disponibile (ovvero dal campione casuale). Una soluzione intuitiva consiste nello scegliere l'equivalente campionario di  $\theta$ , vale a dire lo stimatore *plug-in* fornito dal massimo campionario (ovvero dalla  $n - ma$  statistica ordinata)

$$X_{(n)} = \max(X_1, X_2, \dots, X_n) \quad (3.24)$$

Nel capitolo precedente abbiamo ricavato la funzione di densità della  $m$ -ma statistica ordinata. In conseguenza di quel risultato, posto  $m = n$ , la funzione di densità della  $n - ma$  statistica ordinata di un campione casuale proveniente da una distribuzione Uniforme su  $[0, \theta]$  è data da

$$f_{X_{(n)}}(x; \theta) = \frac{n}{\theta^n} x^{n-1} \mathbb{1}_{[0, \theta]}(x), \quad \theta > 0 \quad (3.25)$$

Ma allora

$$\mathbb{E}_\theta(X_{(n)}) = \int_0^\theta x \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{n+1}\theta \neq \theta \quad (3.26)$$

Ne discende che il massimo campionario è uno stimatore distorto di  $\theta$  sebbene sia comunque asintoticamente non distorto. Il risultato ottenuto può tuttavia aiutarci a individuare uno stimatore non distorto di  $\theta$ . Possiamo però supporre che  $Y_{(n)} = \frac{(n+1)}{n} X_{(n)}$  sia uno stimatore non distorto; infatti,

$$\mathbb{E}(Y_{(n)}) = \frac{n+1}{n} \mathbb{E}(X_{(n)}) = \theta, \quad \forall \theta \in \Theta \quad (3.27)$$

Verifichiamo immediatamente che lo stimatore  $Y_{(n)}$  sia anche consistente per  $\theta$ , calcolandone prima la varianza e successivamente applicando la diseguaglianza di Chebychev; ora,

$$\begin{aligned} \mathbb{V}ar(X_{(n)}) &= \mathbb{E}(X_{(n)}^2) - \mathbb{E}(X_{(n)})^2 \\ &= \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n} dx - \left(\frac{n}{n+1}\theta\right)^2 \\ &= \frac{n}{\theta^n} \int_0^\theta x^{n+1} dx - \left(\frac{n}{n+1}\theta\right)^2 \\ &= \frac{\theta^2}{n(n+2)} - \left(\frac{n}{n+1}\theta\right)^2 \\ &= \frac{n\theta^2}{(n+1)^2(n+2)} \end{aligned} \quad (3.28)$$

sicché

$$\mathbb{V}ar(Y_{(n)}) = \mathbb{V}ar\left(\frac{(n+1)}{n} X_{(n)}\right) = \frac{(n+1)^2}{n^2} \mathbb{V}ar(X_{(n)}) = \frac{\theta^2}{n(n+2)} \quad (3.29)$$

e per Chebychev,

$$P(|Y_{(n)} - \mathbb{E}_\theta(Y_{(n)})| \geq \varepsilon) \leq \frac{\mathbb{V}ar(Y_{(n)})}{\varepsilon^2} = \frac{\theta^2}{n(n+2)\varepsilon^2} \xrightarrow[n \rightarrow \infty]{} 0 \quad (3.30)$$

Abbiamo di conseguenza mostrato che  $Y_{(n)}$  è uno stimatore consistente di  $\theta$ , in quanto converge in probabilità a  $\mathbb{E}(Y_{(n)}) = \theta$ .

Si dimostra anche che  $X_{(n)} \xrightarrow{P} \theta$  ossia che anche  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$  è uno stimatore consistente (sebbene distorto) per  $\theta$ . Infatti,

$$\begin{aligned} P(|X_{(n)} - \theta| > \varepsilon) &= P[(\theta - X_{(n)}) > \varepsilon], \quad \forall \varepsilon > 0 \\ &= P[X_{(n)} \leq \theta - \varepsilon] \\ &= F_{X_{(n)}}(\theta - \varepsilon) \\ &= \left(\frac{\theta - \varepsilon}{\theta}\right)^n \\ &= \left(1 - \frac{\varepsilon}{\theta}\right)^n \end{aligned} \tag{3.31}$$

e perciò

$$\lim_{n \rightarrow \infty} P(|X_{(n)} - \theta| > \varepsilon) = \lim_{n \rightarrow \infty} \left(1 - \frac{\varepsilon}{\theta}\right)^n = 0 \tag{3.32}$$

poiché  $(1 - \frac{\varepsilon}{\theta}) < 1$  da cui segue immediatamente che

$$X_{(n)} \xrightarrow{P} \theta. \tag{3.33}$$

Notiamo però subito che possiamo anche ottenere altri stimatori consistenti e non distorti per  $\theta$ . In particolare, sappiamo che se  $X \sim U(0; \theta)$  allora  $E_\theta(X) = \theta/2$ , da cui  $\theta = 2E_\theta(X)$ . Per il principio del *plug-in* possiamo quindi approssimare  $E_\theta(X)$  con  $\bar{X}_n$  e stimare quindi  $\theta$  mediante lo stimatore  $T_n = 2\bar{X}_n$  che, tra l'altro, risulta essere lo stimatore di  $\theta$  ottenuto con il *metodo dei momenti*, soluzione rispetto a  $\theta$  dell'equazione

$$\frac{\theta}{2} = \mu'_1 = m'_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n. \tag{3.34}$$

Si verifica facilmente che  $E_\theta(2\bar{X}_n) = \theta$  e ricorrendo alla LDGN che  $2\bar{X}_n \xrightarrow{P} \theta$ ; pertanto anche  $T_n$  è uno stimatore non distorto e consistente per il parametro  $\theta$ .

Si pone dunque un interessante (e cruciale) *quesito*: quale stimatore per il parametro  $\theta$  scegliere tra i tre competitori  $X_{(n)}$ ,  $Y_{(n)}$  e  $T_n$  (ma potremmo costruirne molti di più)?

A parità di non distorsione e consistenza (questa duplice condizione esclude lo stimatore  $X_{(n)}$ ), risulta opportuno scegliere lo stimatore *più preciso*, vale a dire quello con varianza più piccola perché più affidabile in termini inferenziali. Confrontiamo allora la varianza dei due stimatori non distorti di  $\theta$ ,  $Y_{(n)}$  e  $T_n$ :

$$\mathbb{V}ar_\theta(Y_{(n)}) = \frac{\theta^2}{n(n+2)} \quad \text{e} \quad \mathbb{V}ar_\theta(2\bar{X}_n) = \frac{\theta^2}{3n} \tag{3.35}$$

Appare evidente che, per ogni valore di  $n > 1$ ,

$$\mathbb{V}ar_\theta(Y_{(n)}) \leq \mathbb{V}ar_\theta(2\bar{X}_n)$$

e pertanto lo stimatore più preciso (ergo, affidabile) sarà  $Y_{(n)}$ .

Possiamo anche dire che  $Y_{(n)}$  è *relativamente più preciso* di  $T_n$ . Ma è il *miglior stimatore in assoluto* del parametro  $\theta$  e quindi migliore rispetto a un qualsiasi altro stimatore (non distorto) di  $\theta$ ?

Rispondere a questa domanda, e farlo in termini generali che quindi prescindono dall'esempio particolare che stiamo trattando, ci porterà via un po' di tempo e comporterà introdurre il concetto di *limite inferiore della varianza* di un qualsiasi stimatore non distorto del parametro (o della funzione del parametro) di interesse.

## 3.2 Convergenza in distribuzione

Il concetto di convergenza in distribuzione si basa sulla seguente intuizione: due variabili casuali sono *vicine* l'una all'altra se le loro *distribuzioni lo sono*.

**Definizione 3.2.1** (Convergenza in distribuzione). Sia  $\{X_n\}_{n \in \mathbb{N}}$  una successione di variabili casuali e sia  $F_{X_n}(x) = F_n(x)$  la funzione di ripartizione dell'elemento  $X_n$  della successione. Sia inoltre  $X$  una variabile casuale avente funzione di ripartizione  $F_X(x)$  e sia  $C(F_X)$  l'insieme dei punti in cui  $F_X(x)$  è continua. Diremo che  $X_n$  converge in distribuzione a  $X$  se e solo se

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x), \quad \forall x \in C(F_X) \quad (3.36)$$

o, in altri termini, se e solo se la successione delle funzioni di ripartizione  $\{F_{X_n}(x)\}_{n \in \mathbb{N}}$  converge a  $\{F_X(x)\}_{n \in \mathbb{N}}$  per ogni scelta di  $x$  in  $C(F_X)$  (tranne al più un insieme "speciale" di valori di  $x$  dove  $F_{X_n}(x)$  non è continua in  $x$ ). Sinteticamente scriveremo

$$X_n \xrightarrow{D} X$$

e chiameremo la v.c.  $X$  *limite in distribuzione* della successione  $\{X_n\}_{n \in \mathbb{N}}$ .

Vale la pena osservare che:

- a) nonostante diciamo che una successione di v.c. converge in distribuzione, sono in realtà le funzioni di ripartizione corrispondenti agli elementi della successione  $X_n$  che convergono alla distribuzione limite  $F_X(x)$  di  $X$  e non le v.c. di per sé. In questo senso, nella convergenza in distribuzione non è più necessario che  $X_n$  e  $X$  siano definite sullo stesso spazio campionario, diversamente da quanto accadeva per la convergenza in probabilità.
- b)  $F_X(x)$  è anche detta *distribuzione asintotica* di  $X_n$ .

**Esempio 3.2.1.** Sia  $\{X_n\}_{n \in \mathbb{N}}$  una successione di variabili casuali aventi funzione di massa

$$P(X_n = k) = \begin{cases} \frac{n 2^k + 3^{k-1}}{(n+1)4^k} & \text{se } k \in \mathbb{N} \\ 0 & \text{altrimenti} \end{cases} \quad (3.37)$$

Vogliamo, innanzitutto, dimostrare che la funzione appena definita è una funzione di massa probabilistica. Per farlo, dimostriamo che la sua sommatoria sui naturali

ha valore unitario.

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{n2^k + 3^{k-1}}{(n+1)4^k} &= \frac{n}{n+1} \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k + \frac{1}{3(n+1)} \sum_{k=1}^{\infty} \left(\frac{1}{4}\right)^k = \\ &= \frac{n}{n+1} \left( \frac{1}{1-1/2} - 1 \right) + \frac{1}{3(n+1)} \left( \frac{1}{1-1/4} - 1 \right) = 1 \end{aligned} \quad (3.38)$$

Possiamo, inoltre, dimostrare che  $X_n$  converge in distribuzione a una distribuzione geometrica di parametro  $p$  e calcolare il valore di tale parametro. È evidente che la funzione di ripartizione associata alla distribuzione di  $X_n$  è la seguente

$$F_{X_n}(x) = P(X_n \leq x) = \sum_{k=1}^x \frac{2^k n + 3^{k-1}}{(n+1)4^k} \quad (3.39)$$

Calcoliamo dunque il limite per  $n \rightarrow \infty$  di tale sommatoria, che abbiamo già scomposto in modo conveniente in (3.38)

$$\lim_{n \rightarrow \infty} \left[ \frac{n}{n+1} \sum_{k=1}^x \left(\frac{1}{2}\right)^k + \frac{1}{3(n+1)} \sum_{k=1}^x \left(\frac{1}{4}\right)^k \right] = \sum_{k=1}^x \left(\frac{1}{2}\right)^k = \frac{1}{2} \sum_{k=1}^x \left(1 - \frac{1}{2}\right)^{k-1} \quad (3.40)$$

e ricordando che se  $X \sim Geo(p)$  allora la sua funzione di ripartizione è data da

$$F_X(x) = p \sum_{k=1}^x (1-p)^{k-1}, \quad x \in \mathbb{N} \quad (3.41)$$

abbiamo dunque dimostrato che  $X_n \xrightarrow{D} X \sim G(1/2)$  sicché  $p = \frac{1}{2}$

**Esempio 3.2.2** (Distribuzione del massimo campionario per campionamento da Uniforme). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da  $U(0, \theta)$  e sia  $X_{(n)} = \max\{X_1, \dots, X_n\}$ , il massimo campionario. Vogliamo trovare la distribuzione asintotica di

$$Z_n = n(\theta - X_{(n)}).$$

Calcoliamo, innanzitutto, la funzione densità e la funzione di ripartizione associate a  $X_{(n)}$  usando i risultati che conosciamo:

$$\begin{aligned} f_{X_{(n)}}(x; \theta) &= \frac{n}{\theta^n} x^{n-1} 1_{(0, \theta]}(x) \quad \text{con } \theta > 0 \\ F_{X_{(n)}}(x; \theta) &= \int_0^x \frac{n}{\theta^n} (v^{n-1}) dv = \left(\frac{x}{\theta}\right)^n \end{aligned} \quad (3.42)$$

La distribuzione associata a  $Z_n$  è dunque la seguente:

$$\begin{aligned} F_{Z_n}(z) &= P(Z_n \leq z) \\ &= P(n(\theta - X_{(n)}) \leq z) \\ &= P\left(X_{(n)} \geq \theta - \frac{z}{n}\right) = 1 - P\left(X_{(n)} < \theta - \frac{z}{n}\right) \\ &= 1 - F_{X_{(n)}}\left(\theta - \frac{z}{n}\right) \\ &= 1 - \left(\frac{\theta - z/n}{\theta}\right)^n \\ &= 1 - \left(1 + \left(-\frac{z}{n\theta}\right)^n\right) \xrightarrow{n \rightarrow \infty} 1 - e^{-z/\theta} \end{aligned} \quad (3.43)$$

Pertanto  $Z_n$  tende asintoticamente a una distribuzione esponenziale di parametro  $\theta$  ovvero,  $Z_n \xrightarrow{d} Z \sim \text{Exp}(\theta)$ ,  $\theta > 0$ .

### 3.2.1 Alcuni utili risultati relativi alla convergenza in distribuzione

Daremo di seguito alcuni risultati (senza fornirne sempre la dimostrazione) utili nelle applicazioni dove è richiesta la convergenza in distribuzione; tra questi il teorema più importante del calcolo delle probabilità e della statistica matematica noto come *Teorema Limite Centrale* dove l'aggettivo centrale sta a sottolinearne proprio l'importanza. Daremo, inoltre, una serie di risultati che legano la convergenza in distribuzione alla convergenza in probabilità.

**Teorema 3.2.1** (Continuous mapping). Si supponga  $X_n$  convergere in distribuzione a  $X$  e sia  $g$  una funzione continua di  $X_n$  sul supporto di  $X$ . Allora vale

$$g(X_n) \xrightarrow{D} g(X).$$

In altre parole, funzioni continue preservano la convergenza in distribuzione.

**Teorema 3.2.2.** Sia  $X_n \xrightarrow{D} X$  e sia  $Y_n \xrightarrow{P} 0$ . Allora

$$X_n + Y_n \xrightarrow{D} X.$$

Il teorema che segue è spesso usato in associazione con il Teorema Limite Centrale quale strumento assai utile per realizzare approssimazioni di oggetti di notevole interesse in statistica (avremo modo di tornare sull'argomento più avanti).

**Teorema 3.2.3** (di Slutsky). Siano  $X_n$ ,  $X$ ,  $A_n$  e  $B_n$  variabili casuali e siano  $a$  e  $b$  due costanti. Se  $X_n \xrightarrow{D} X$ ,  $A_n \xrightarrow{P} a$ ,  $B_n \xrightarrow{P} b$ , allora

$$A_n + B_n X_n \xrightarrow{D} a + bX$$

Il teorema che segue stabilisce una *relazione di implicazione* tra convergenza in probabilità e convergenza in distribuzione, sottolineando come la seconda sia più debole della prima (il che dà anche ragione del nome *convergenza debole* con cui spesso ci si riferisce alla convergenza in distribuzione).

**Teorema 3.2.4.** Se  $X_n \xrightarrow{P} X$  allora  $X_n \xrightarrow{D} X$ .

La relazione di implicazione va in un solo senso: la convergenza in probabilità implica, in generale, la convergenza in distribuzione. Vi è però un caso particolare in cui vale anche l'altro verso dell'implicazione e di ciò si occupa il corollario che segue.

**Corollario 3.2.1.** Se  $X_n \xrightarrow{D} a$ , con  $a$  costante, allora  $X_n \xrightarrow{P} a$ .

Spesso si è interessati al comportamento della distribuzione della media campionaria  $\bar{X}_n$  su grandi campioni e, in particolare, a studiare la sua *distribuzione limite o asintotica* (e, più in generale, potremmo essere interessati al comportamento di somme di potenze delle variabili casuali che costituiscono il campione casuale  $(X_1, X_2, \dots, X_n)$ ). Il seguente teorema offre una sorprendente soluzione.

**Teorema 3.2.5** (Teorema Limite Centrale). Sia  $X_1, X_2, \dots$  una successione di v.c. indipendenti e identicamente distribuite la cui funzione generatrice dei momenti esiste in un intorno dello zero (ossia, esiste  $M_{X_i}(t)$  per  $|t| < t_0, t_0 > 0$ ). Sia  $\mathbb{E}(X_i) = \mu$  e  $\mathbb{V}ar(X_i) = \sigma^2 > 0$ . (Sia  $\mu$  che  $\sigma^2$  sono finite dal momento che esiste la fgm). Sia  $\bar{X}_n$  la media campionaria e innidichiamo con  $G_n(y)$  la funzione di ripartizione della v.c.

$$Y_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

Allora per un qualsiasi  $x$ ,  $-\infty < x < \infty$ ,

$$\lim_{n \rightarrow \infty} G_n(y) = \int_{-\infty}^y \frac{1}{\sqrt{2}} e^{-\frac{v^2}{2}} dv \quad (3.44)$$

ovvero,  $Y_n$  ha distribuzione limite (o *asintotica*) Normale standard; in altre parole, possiamo scrivere

$$Y_n \xrightarrow{D} N(0, 1)$$

o, in alternativa,

$$Y_n \underset{a}{\sim} N(0, 1)$$

*Dimostrazione.* La filosofia della dimostrazione poggia sull'unicità della fgm, laddove essa esista e sul fatto che, se esiste, identifica univocamente la distribuzione.

Abbiamo già mostrato nella dimostrazione della LDGN che  $E[\bar{X}_n] = \mu$  e  $\mathbb{V}ar(\bar{X}_n) = \sigma^2/n$ . Sapendo ciò, è facile mostrare che  $E[Y_n] = 0$  e  $\mathbb{V}ar(Y_n) = 1$ . Se le variabili  $X_k$  hanno distribuzione Normale, la tesi è automaticamente dimostrata. In caso contrario, definiamo la successione di variabili casuali

$$Z_k = \frac{X_k - \mu}{\sigma} \quad (3.45)$$

con  $k = 1, 2, \dots$ . Poiché le variabili  $X_k$  sono indipendenti e identicamente distribuite per ipotesi, allora anche le variabili  $Z_k$  lo sono; inoltre,  $E[Z_k] = 0$  e  $\mathbb{V}ar(Z_k) = 1$ . In termini di queste nuove variabili, la variabile casuale  $Y_n$  può essere definita come segue:

$$Y_n = \frac{Z_1 + Z_2 + \dots + Z_n}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^n Z_k \quad (3.46)$$

Definiamo ora  $\xi_k = Z_k/\sqrt{n}$  per comodità di notazione e calcoliamo ora la funzione generatrice dei momenti di  $Y_n$  e successivamente effettuiamo uno sviluppo in serie di McLaurin:

$$M_{Y_n}(t) = \prod_{k=1}^n M_{\xi_k}(t) = [M_{\xi_1}(t)]^n \simeq \left[ M_{\xi_1}(0) + t M'_{\xi_1}(0) + t^2 \frac{M''_{\xi_1}(0)}{2} + o(t^2) \right]^n \quad (3.47)$$

Applicando la definizione di funzione generatrice dei momenti, è facile mostrare che  $M_{\xi_1}(0) = 1$ ,  $M'_{\xi_1}(0) = E[Z_1]/\sqrt{n}$  e  $M''_{\xi_1}(0) = \mathbb{V}ar(\xi_1) + E[Z_1]^2 = 1/n$ . Sostituendo

queste relazioni all'interno dello sviluppo in serie di McLaurin ed effettuando il limite per  $n \rightarrow \infty$  si ottiene

$$\lim_{n \rightarrow \infty} M_{Y_n}(t) = \lim_{n \rightarrow \infty} \left[ 1 + \frac{t^2}{2n} + o(t^2) \right]^n = e^{\frac{t^2}{2}} \quad (3.48)$$

Questa coincide con la funzione generatrice dei momenti della funzione di distribuzione di una variabile Normale standard. Pertanto, per la proprietà di unicità della funzione generatrice dei momenti, la tesi è dimostrata.  $\square$

Si noti che dal TLC discendono naturalmente anche i due seguenti risultati:

$$\begin{aligned} W_n &= \sum_{i=1}^n X_i \xrightarrow{D} N(n\mu, n\sigma^2) \\ \bar{X}_n &= \frac{W_n}{n} \xrightarrow{D} N\left(\mu, \frac{\sigma^2}{n}\right) \end{aligned} \quad (3.49)$$

**Esempio 3.2.3.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Uniforme su  $[0, \theta]$  e sia  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$ . Vogliamo ora di trovare la *distribuzione asintotica* di  $Z_n = n(\theta - X_{(n)})$ .

Come abbiamo già avuto modo di vedere,

$$\begin{aligned} f_{X_{(n)}}(x; \theta) &= \frac{n}{\theta^n} x^{n-1} \mathbb{1}_{[0, \theta]}(x), \quad \theta > 0 \\ F_{X_{(n)}}(x; \theta) &= \left(\frac{x}{\theta}\right)^n \end{aligned} \quad (3.50)$$

Allora,

$$\begin{aligned} F_{Z_n}(z; \theta) &= P_\theta(Z_n \leq z) = P_\theta(n(\theta - X_{(n)}) \leq z) \\ &= P_\theta\left(X_{(n)} \geq \theta - \frac{z}{n}\right) \\ &= 1 - P_\theta\left(X_{(n)} \leq \theta - \frac{z}{n}\right) \\ &= 1 - F_{X_{(n)}}\left(\theta - \frac{z}{n}\right) \\ &= 1 - \left(\frac{\theta - z/n}{\theta}\right)^n \\ &= 1 - \left(1 - \frac{z/n}{\theta}\right)^n \\ &= 1 - \left(1 + \frac{-(z/n)}{\theta}\right)^n \xrightarrow{n \rightarrow \infty} 1 - e^{-\frac{z}{\theta}} \end{aligned} \quad (3.51)$$

essendo  $\lim_{n \rightarrow \infty} \left(1 + \frac{b}{n}\right)^n = e^b$  (limite notevole).

Ma  $1 - e^{-\frac{z}{\theta}}$  è la funzione di distribuzione di una v.c. esponenziale di media  $\theta$  sicché

$$Z_n = n(\theta - X_{(n)}) \underset{a}{\sim} Z \sim \text{Exp}(\theta). \quad (3.52)$$

**Esempio 3.2.4.** Sia  $Y \sim b(n = 100, p = 1/2)$ , ovvero sia  $Y = \sum_{i=1}^{100} X_i$  con  $X_i \sim b(1, p = 1/2)$  indipendenti e identicamente distribuite. Vogliamo dimostrare che una variabile casuale  $X \sim N(\mu = 500, \sigma^2 = 25)$  può essere usata per approssimare la distribuzione di  $Y$ . Infatti, per il teorema del limite centrale, si ha che

$$Y \xrightarrow{D} N(np, np(1-p)) = N(50, 25) \quad (3.53)$$

Vogliamo ora stabilire che errore commettiamo se effettuiamo tale approssimazione. Per farlo, calcoliamo  $P(40 \leq Y \leq 65)$  per entrambe le distribuzioni (esatta e asintotica): nel primo caso (distribuzione esatta), si ha

$$\begin{aligned} P(40 \leq Y \leq 65) &= \sum_{y=0}^{65} \binom{100}{y} \left(\frac{1}{2}\right)^y \left(1 - \frac{1}{2}\right)^{100-y} - \sum_{y=0}^{39} \binom{100}{y} \left(\frac{1}{2}\right)^y \left(1 - \frac{1}{2}\right)^{100-y} \\ &= \frac{1}{2^{100}} 1.24421 \cdot 10^{30} \simeq 0.9815 \end{aligned} \quad (3.54)$$

mentre, usando l'approssimazione suggerita dal TLC, applicando la *correzione per la continuità*, si ha

$$\begin{aligned} P(40 \leq Y \leq 65) &= P(40 - 0.5 \leq Y \leq 65 + 0.5) \\ &= \Phi_Z\left(\frac{65.5 - 50}{5}\right) - \Phi_Z\left(\frac{39.5 - 50}{5}\right) \\ &= \Phi_Z(3.1) - [1 - \Phi_Z(2.1)] \\ &= 0.9990 - [1 - 0.9821] = 0.9811 \end{aligned} \quad (3.55)$$

L'errore di approssimazione commesso usando l'approssimazione suggerita dal TLC è dunque pari a 0.0004, assolutamente sopportabile tenuto conto della grande semplificazione nel calcolo da effettuare.

**Osservazione 2 (Correzione per la continuità).** Laddove si intende approssimare la distribuzione di una v.c. discreta con quanto suggerito dal TLC è opportuno, in presenza di piccoli campioni (conventionalmente, per  $n < 30$ ), applicare quella che è nota come *correzione per la continuità*: essa consiste semplicemente nel ridefinire l'intervallo di cui interessa la probabilità, diciamo  $[a, b]$ , come  $[a - \frac{1}{2}, b + \frac{1}{2}]$ . Ed è esattamente quanto abbiamo fatto in (3.55) anche se in questo caso, si poteva evitare essendo  $n = 100 > 30$  in quanto il guadagno in termini di bontà di approssimazione che segue a tale correzione è decisamente più limitato di quanto non sarebbe per valori di  $n$  decisamente più piccoli (per esempio, se  $n$  fosse stato pari a  $12 < 30$ ).

### 3.2.2 Delta method

Come abbiamo già avuto occasione di rimarcare molte altre volte, un problema che comunemente ricorre in statistica matematica è quello in cui, conoscendo la distribuzione di una v.c. (univariata o multivariata che sia), si è interessati a trovare la distribuzione di una sua trasformata

$$X \sim F_X(x; \theta) \longmapsto Y = g(X) \sim ? \quad (3.56)$$

Questa situazione si ripropone anche in ambito *asintotico* e i teoremi del *continuous mapping* e di *Slutsky* ne sono illustrazione.

Un altro utile (e fondamentale) risultato è il *metodo delta*: esso permette di determinare le *distribuzioni* campionarie *asintotiche* di stimatori per *funzioni*  $g(\cdot)$  del parametro  $\theta$  di un modello statistico  $(\mathfrak{X}, f_{\mathbf{X}}(\mathbf{x}; \theta), \theta \in \Theta)$ . Il *metodo delta* può essere visto come un Teorema Limite Centrale molto generale e si basa sul seguente teorema:

**Teorema 3.2.6** (Delta Method (univariato)). Data una successione di v.c.  $\{X_n\}_{n \in \mathbb{N}}$  tale che

$$\sqrt{n}(X_n - \theta) \xrightarrow{D} N(0, \sigma_0^2) \quad (\text{dove } \sigma_0^2 \text{ è la varianza asintotica di } X_n) \quad (3.57)$$

allora per ogni funzione (continua)  $g(\cdot)$ , con derivata prima in  $\theta$  finita e diversa da 0, si ha:

$$\sqrt{n}\{g(X_n) - g(\theta)\} \xrightarrow{D} N\left(0, \sigma_0^2 \cdot [g'(\theta)]^2\right).$$

*Dimostrazione.* La dimostrazione si basa su uno sviluppo di Taylor per  $g(X_n)$  attorno al punto  $X_n = \theta$ , arrestato al primo termine:

$$g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + R(\theta, X_n), \quad (3.58)$$

dove  $R(\theta, X_n) = O([X_n - \theta]^2)$  è il resto della serie, ovvero una funzione di  $\theta$  e di  $X_n$  che tende a zero in probabilità con la stessa velocità di  $(X_n - \theta)^2$  (o, in modo equivalente, che tende a zero più velocemente di  $(X_n - \theta)$ ).

Prendendo il valore atteso a sinistra e a destra della (3.58) e trascurando  $R(\theta, X_n)$ , si ottiene la seguente approssimazione:

$$\mathbb{E}[g(X_n)] \simeq g(\theta) + g'(\theta)\mathbb{E}[(X_n - \theta)].$$

Dato che, per l'assunzione in (3.57),  $\mathbb{E}(X_n - \theta) = 0$ , si ottiene che  $\mathbb{E}[g(X_n)] \simeq g(\theta)$ . Da questo risultato e da quanto in (3.58) si ricava la seguente approssimazione per la varianza di  $g(X_n)$ :

$$\begin{aligned} \mathbb{V}ar[g(X_n)] &\stackrel{d}{=} \mathbb{E}[g(X_n) - \mathbb{E}(g(X_n))]^2 \simeq \mathbb{E}[g(X_n) - g(\theta)]^2 \\ &\simeq \mathbb{E}[g'(\theta)(X_n - \theta)]^2 \\ &= [g'(\theta)]^2 \mathbb{E}[X_n - \theta]^2 \\ &= [g'(\theta)]^2 \mathbb{V}ar(X_n). \end{aligned}$$

Trascurando  $R(\theta, X_n)$  nella (3.58) e moltiplicando la medesima a sinistra e a destra per  $\sqrt{n}$ , si ha:

$$\sqrt{n}[g(X_n) - g(\theta)] \simeq g'(\theta)[\sqrt{n}(X_n - \theta)].$$

Per l'assunzione in (3.57), la parte destra converge *in distribuzione* a una variabile casuale  $N(0, \sigma_0^2)$ , moltiplicata per  $g'(\theta)$ ; quindi, per il teorema di Slutsky, si ottiene che:

$$\sqrt{n}\{g(X_n) - g(\theta)\} \xrightarrow{D} N\left(0, \sigma_0^2 \cdot [g'(\theta)]^2\right).$$

□

In altre parole, il *delta method* permette di ricavare, sotto opportune condizioni, la distribuzione *asintotica* di  $g(X_n)$  a partire dalla distribuzione *asintotica* di  $X_n$ .

**Esempio 3.2.5.** Sia  $\{X_n\}_{n \in \mathbb{N}}$  una successione di variabili casuali *asintoticamente Normali* con media (*asintotica*)  $\mu = 0$  e varianza (*asintotica*)  $\sigma_0^2 = 1$  sicché

$$\sqrt{n} \bar{X}_n \underset{a}{\sim} N(0, 1).$$

Vogliamo determinare la distribuzione asintotica di  $\exp\{\bar{X}_n\}$ .

La funzione  $g(\mu) = \exp\{\mu\}$  soddisfa i requisiti per l'applicazione del *delta method* essendo  $g(\mu) = 1$  e  $g'(\mu)|_{\mu=0} = 1$ ; espandiamo dunque  $\exp\{\bar{X}_n\}$  in serie di Taylor intorno a  $\mu = 0$ , arrestando lo sviluppo al primo ordine, ottenendo

$$\exp\{\bar{X}_n\} \simeq \exp(0) + \frac{d}{d\mu} \exp(\mu) |_{\mu=0} (\bar{X}_n - \mu)$$

sicché

$$\mathbb{E}(\exp\{\bar{X}_n\}) = 1 \quad \text{e} \quad \mathbb{V}ar(\exp\{\bar{X}_n\}) = 1^2 \mathbb{V}ar(\bar{X}_n) = \frac{1}{n}$$

Perciò, per il *delta method*

$$\sqrt{n} (\exp\{\bar{X}_n\} - 1) \xrightarrow{D} N(0, 1) \quad (3.59)$$

e dunque,

$$\exp\{\bar{X}_n\} \xrightarrow{D} N\left(1, \frac{1}{n}\right) \quad (3.60)$$

che possiamo anche scrivere come

$$\exp\{\bar{X}_n\} \underset{a}{\sim} N\left(1, \frac{1}{n}\right) \quad (3.61)$$

notazione, quest'ultima, spesso preferita per riferirsi alla *distribuzione asintotica* di una v.c. dove la scrittura  $\underset{a}{\sim}$  si legge "*asintoticamente distribuita come...*".

**Esempio 3.2.6.** Supponiamo  $X_1, X_2, \dots$  essere v.c. indipendenti e identicamente distribuite con media  $\mu$  e varianza  $\sigma^2$  finite e di essere interessati a trovare la distribuzione asintotica di  $\bar{X}_n^2$ .

Ora per il *TLC* si sa che

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2) \quad (3.62)$$

sicché richiamando il *delta method*, posto  $\bar{X}_n^2 = g(\bar{X}_n)$  funzione continua che soddisfa le assunzioni richieste dal teorema in questione,

$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \underset{a}{\sim} N(0, 4\mu^2\sigma^4) \quad (3.63)$$

purché  $\mu \neq 0$ .

Trattiamo separatamente il caso in cui  $\mu = 0$ . Ora,

- $\sqrt{n}(\bar{X}_n - \mu) = \sqrt{n} \bar{X}_n \xrightarrow{D} N(0, \sigma^2)$  per il *TLC*

- $[\sqrt{n}(\bar{X}_n - \mu)]^2 = n \bar{X}_n^2 \xrightarrow{D} \sigma^2 \chi_1^2$

**Esempio 3.2.7.** Consideriamo

$$Y_n = \frac{\chi_n^2 - n}{\sqrt{2n}} = \sqrt{n} \left( \frac{\chi_n^2}{\sqrt{2n}} - \frac{1}{\sqrt{2}} \right) \quad (3.64)$$

con  $\chi_n^2$  distribuzione chi-quadrato con  $n$  gradi di libertà. Ricordiamo che  $\mathbb{E}(\chi_n^2) = n$  e che  $\mathbb{V}ar(\chi_n^2) = 2n$  (discende dal fatto che  $\chi_n^2 \sim \mathcal{G}(\alpha = n/2, \beta = 2)$ ). Affermiamo che  $Y_n \xrightarrow{d} N(0, 1)$ . Infatti:

$$Y_n = \frac{\chi_n^2 - n}{\sqrt{2n}} = \frac{\sum_{i=1}^n X_i^2 - n \cdot 1}{\sqrt{n}\sqrt{2}} \quad (3.65)$$

dove  $X_i \sim N(0, 1)$ , e quindi  $X_i^2 \sim \chi_1^2$ , quindi le  $X_i^2$  hanno media  $\mu = 1$  e varianza  $\sigma^2 = 2$ . Quindi per il Teorema centrale del Limite (vedi sotto) si ha quanto voluto.

Scrivendo ora  $Y_n$  nella forma

$$Y_n = \sqrt{n} \left( \frac{\chi_n^2}{\sqrt{2n}} - \frac{1}{\sqrt{2}} \right)$$

riconosciamo che la prima parte delle ipotesi del *delta method* sono soddisfatte. Consideriamo quindi  $g(t) = \sqrt{t}$ , che è derivabile in  $\vartheta = 1/\sqrt{2}$ ,  $g'(t) = \frac{1}{2\sqrt{t}}|_{\vartheta=1/\sqrt{2}} = 2^{-3/4}$ . Allora

$$\sqrt{n} \left[ g \left( \frac{\chi_n^2}{\sqrt{2n}} \right) - g(\vartheta) \right] = \sqrt{n} \left( \sqrt{\frac{\chi_n^2}{\sqrt{2n}}} - \sqrt{\frac{1}{\sqrt{2}}} \right) \xrightarrow{D} N(0, 1^2 \cdot 2^{-3/2}) \quad (3.66)$$

Spesso il *delta method* è uno strumento ideale per calcolare la varianza asintotica di opportune statistiche (stimatori, nella maggior parte dei casi), quantità che riveste un ruolo cruciale in statistica matematica in quanto legata alla *dispersione* della statistica stessa e quindi, in ultima istanza, alla sua *precisione*.

**Esempio 3.2.8** (Approssimazione di media e varianza). Consideriamo un vettore casuale  $(X_1, X_2, \dots, X_n)$  da una certa distribuzione. Siamo interessati a stimare  $g(\mu)$ . Possiamo supporre di approssimare, via espansione in serie di Taylor arrestata al primo ordine,  $g(\bar{X}_n)$  intorno a  $\mu$ :

$$g(\bar{X}_n) = g(\mu) + g'(\mu)(\bar{X}_n - \mu)$$

In base a questa approssimazione, possiamo dire che

$$E(g(\bar{X}_n)) \simeq g(\mu) \quad \text{e} \quad \mathbb{V}ar(g(\bar{X}_n)) \simeq g'(\mu)^2 \mathbb{V}ar(\bar{X}_n) \quad (3.67)$$

Sia ora  $g(\mu) = 1/\mu$  e supponiamo di stimarlo attraverso lo stimatore *plug-in*  $1/\bar{X}_n$ . Oraabbiamo

$$\begin{aligned} \mathbb{E} \left( \frac{1}{\bar{X}_n} \right) &\simeq \frac{1}{\mu}, \\ \mathbb{V}ar \left( \frac{1}{\bar{X}_n} \right) &\simeq \left( -\frac{1}{\mu^2} \right)^2 \mathbb{V}ar(\bar{X}_n) = \frac{\sigma^2}{n\mu^4} \end{aligned} \quad (3.68)$$

Pertanto, applicando il *delta method*, ricorrendone le condizioni, otteniamo

$$\sqrt{n} \left( \frac{1}{\bar{X}_n} - \frac{1}{\mu} \right) \xrightarrow{a} N \left( 0, \frac{\sigma^2}{n\mu^4} \right) \quad (3.69)$$

E' comunque vero che la quantità  $\frac{\sigma^2}{n\mu^4}$  è incognita e dipende sia da  $\mu$  che da  $\sigma^2$ . Possiamo però *stimarla* ricorrendo a

$$\hat{\mathbb{V}ar}(1/\bar{X}_n) = \frac{S_n^2}{n\bar{X}_n^4}$$

dove  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  è la *varianza campionaria*. Ora sappiamo già che  $\bar{X}_n \xrightarrow{P} \mu$  e  $S_n^2 \xrightarrow{P} \sigma^2$  (ossia entrambi sono stimatori consistenti per i rispettivi parametri) e perciò

$$\left( \frac{1}{\bar{X}_n} \right)^4 S_n^2 \xrightarrow{P} \left( \frac{1}{\mu} \right)^4 \sigma^2 \quad (3.70)$$

e invocando il *teorema di Slutsky*, possiamo concludere che

$$\frac{\sqrt{n} (1/\bar{X}_n - 1/\mu)}{S_n/\bar{X}_n^2} = \frac{\sigma/\mu^2}{S_n/\bar{X}_n^2} \frac{\sqrt{n} (1/\bar{X}_n - 1/\mu)}{\sigma/\mu^2} \xrightarrow{a} N(0, 1) \quad (3.71)$$

**Esempio 3.2.9.** Consideriamo ora un esperimento casuale bernoulliano e supponiamo di voler stimare l'*odds ratio*

$$g(p) = \frac{p}{(1-p)}$$

dove  $p$  è la probabilità di successo.

Sia  $(X_1, X_2, \dots, X_n)$  il risultato della ripetizione sotto identiche condizioni dell'esperimento casuale in questione  $n$  volte (ovvero un campione casuale da  $b(1, p)$ ). Già sappiamo che  $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$  è lo stimatore *plug-in* di  $p$  e che  $g(\hat{p}_n) = \frac{\hat{p}_n}{1-\hat{p}_n}$  è lo stimatore *plug-in* dell'*odds ratio*  $g(p)$  e che

$$g(\hat{p}_n) = \frac{\hat{p}_n}{1-\hat{p}_n} \xrightarrow{P} \frac{p}{1-p} \quad (3.72)$$

Pertanto, tenuto conto della distribuzione asintotica di  $\hat{p}_n$  e del fatto che le condizioni alla base del *delta method* sono soddisfatte, proprio grazie a quest'ultimo risultato abbiamo che

$$g(\hat{p}_n) \xrightarrow{D} g(p).$$

Inoltre,

$$\mathbb{V}ar(g(\hat{p}_n)) \simeq [g'(p)]^2 \mathbb{V}ar(\hat{p}_n) = \frac{p}{n(1-p)^2}.$$

ottenendo così un'approssimazione della varianza dello stimatore dell'*odds ratio*. Ed è del tutto evidente che  $\mathbb{V}ar(g(\hat{p}_n))$  diverrà sempre più piccola al crescere dell'ampiezza campionaria  $n$ .

### 3.2.3 Funzione generatrice dei momenti e convergenza in distribuzione

La *fgm* gioca un ruolo significativo anche nella teoria asintotica. Talvolta lo studio della successione delle *fgm* fornisce una strada percorribile per trovare la distribuzione asintotica di una v.c.  $X_n$ .

Più precisamente, la relazione di *unicivocità* tra *fgm* e distribuzione può essere di fatto sfruttata per stabilire la *convergenza in distribuzione* della successione di v.c.  $\{X_n\}_{n \in \mathbb{N}}$  alla v.c. limite  $X$  deducendo tale convergenza dalla convergenza della successione delle corrispondenti *fgm* di  $X_n$  alla *fgm* della v.c. limite  $X$ .

Il teorema che segue formalizza quanto finora detto.

**Teorema 3.2.7** (Convergenza della successione delle *fgm*). Sia  $\{X_n\}_{n \in \mathbb{N}}$  una successione di v.c. ognuna avente propria *fgm*  $M_{X_n}(t)$ ,  $-t_0 < t < t_0$ ,  $t_0 > 0$  e sia

$$\lim_{n \rightarrow \infty} M_{X_n}(t) = M_X(t), \quad \forall t \in (-t_0, t_0)$$

con  $M_X(t)$  a sua volta *fgm*. Allora esiste un'unica funzione di distribuzione  $F_X(x)$  i cui momenti sono (univocamente) determinati da  $M_X(t)$  e, per tutti di  $x$  in cui  $F_X(x)$  è continua, si avrà

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

ossia

$$X_n \xrightarrow{D} X$$

o, in altre parole,  $X_n$  converge in distribuzione alla variabile casuale  $X$  la cui funzione di distribuzione  $F_X(x)$  risulta (univocamente) determinata da  $M_Y(t)$ .

*Dimostrazione.* La dimostrazione avviene tramite la teoria delle trasformate di Laplace e si può trovare in Feller (1970) - vol. II).  $\square$

**Esempio 3.2.10.** Sia  $X_n \sim b(n, p)$  e  $\mu = np$ , con  $\mu = E(X_n)$  costante. Al fine di ottenere informazioni in merito alla convergenza della distribuzione  $Y_n$  calcoliamo la funzione generatrice dei momenti  $M_{X_n}(t)$ :

$$M_{X_n}(t) = E(e^{tX_n}) = [pe^t + (1-p)]^n = \left[1 + \frac{\mu}{n}(e^t - 1)\right]^n \xrightarrow{n \rightarrow \infty} e^{\mu(e^t - 1)} \quad (3.73)$$

Riconoscendo in  $e^{\mu(e^t - 1)}$  la *fgm* di una v.c. di Poisson, possiamo quindi concludere che la successione  $X_n$  converge alla v.c.  $X \sim \mathcal{P}(\mu)$  ovvero che segue una distribuzione di Poisson di parametro  $\mu$ .

**Esempio 3.2.11.** Sia  $X_n \sim \mathcal{G}(\alpha = n, \beta)$ , con  $\beta > 0$  che non dipende da  $n$ . Vogliamo trovare la distribuzione limite della successione  $Y_n = X_n/n$ . Sappiamo che la funzione densità associata ad  $X_n$  è la seguente:

$$f_{X_n}(x; n, \beta) = \frac{1}{\Gamma(n)\beta^n} x^{n-1} e^{-x/\beta} \mathbb{1}_{\mathbb{R}^+}(x) \quad (3.74)$$

Inoltre, per il teorema di trasformazione, posto  $y = x/n$ , si ha che

$$\begin{aligned} f_{Y_n}(y; n, \beta) &= \frac{1}{\Gamma(n)} \beta^n (ny)^{n-1} e^{ny/\beta} n \mathbb{1}_{\mathbb{R}^+}(x) \\ &= \frac{1}{\Gamma(n)} \left(\frac{\beta}{n}\right)^n y^{n-1} e^{\frac{n}{\beta}y} \mathbb{1}_{\mathbb{R}^+}(x) \end{aligned} \quad (3.75)$$

Possiamo dunque concludere che la successione  $Y_n$  segue una distribuzione Gamma con  $\alpha = n$  e  $\beta^* = \beta/n$  ossian  $Y_n = \mathcal{G}(n, \beta/n)$ .

Calcoliamo allora la funzione generatrice dei momenti associata a questa distribuzione:

$$M_{Y_n}(t) = E(e^{tY_n}) = E\left(e^{\frac{t}{n}X_n}\right) = M_{X_n}\left(\frac{t}{n}\right) = \frac{1}{\left(1 - \frac{\beta t}{n}\right)^n} \xrightarrow{n \rightarrow \infty} e^{\beta t} \quad (3.76)$$

che risulta essere la *fgm* di una *distribuzione degenera* in  $\beta$ . Infatti,

$$\mu'_s = \left. \frac{d^s}{dt^s} e^{\beta t} \right|_{t=0} = \beta^s, \quad s \in \mathbb{N} \quad (3.77)$$

sicché

$$\begin{aligned} \mu'_1 &= \mathbb{E}(Y_n) = \left. \frac{d}{dt} e^{\beta t} \right|_{t=0} = \beta \\ \mu_2 &= \text{Var}(Y_n) = \mu'_2 - (\mu'_1)^2 = \left. \frac{d^2}{dt^2} e^{\beta t} \right|_{t=0} - (\mu'_1)^2 = \beta^2 - \beta^2 = 0 \end{aligned} \quad (3.78)$$

Inoltre,

$$\begin{aligned} \mu_s &= \sum_{m=0}^s (-1)^m \binom{s}{m} (\mu'_m)^m (\mu'_1)^{s-m} \\ &= \sum_{m=0}^s (-1)^m \binom{s}{m} \beta^m \beta^{s-m} \\ &= \beta^s \sum_{m=0}^s (-1)^m \binom{s}{m} = 0, \quad \forall s \geq 3. \end{aligned} \quad (3.79)$$

Concludiamo il paragrafo relativo alla *convergenza in distribuzione* riportando, a scopo riassuntivo, ancora una volta la tavola 3.1 che contiene alcune delle distribuzioni più frequentemente utilizzate in statistica matematica e le reciproche relazioni di generazione siano esse *esatte* (linea continua) o *asintotiche* (linea tratteggiata). Abbiamo ora tutti gli strumenti per leggerla compiutamente e, grazie al *TLC*, avere una giustificazione formale della posizione di centralità occupata dalla famiglia delle distribuzioni Normali che finisce per essere la *distribuzione limite* per molte altre famiglie di distribuzioni.

### 3.3 Convergenza in media quadratica

L'ultimo tipo di convergenza che prenderemo in considerazione va sotto il nome di *convergenza in media quadratica* e tra i tipi di convergenza fin qui considerati è, come vedremo, quello più forte.

Il concetto di *convergenza in media quadratica* poggia sulla seguente intuizione: due v.c. sono vicine l'una all'altra se il quadrato della loro *differenza* è piccolo.

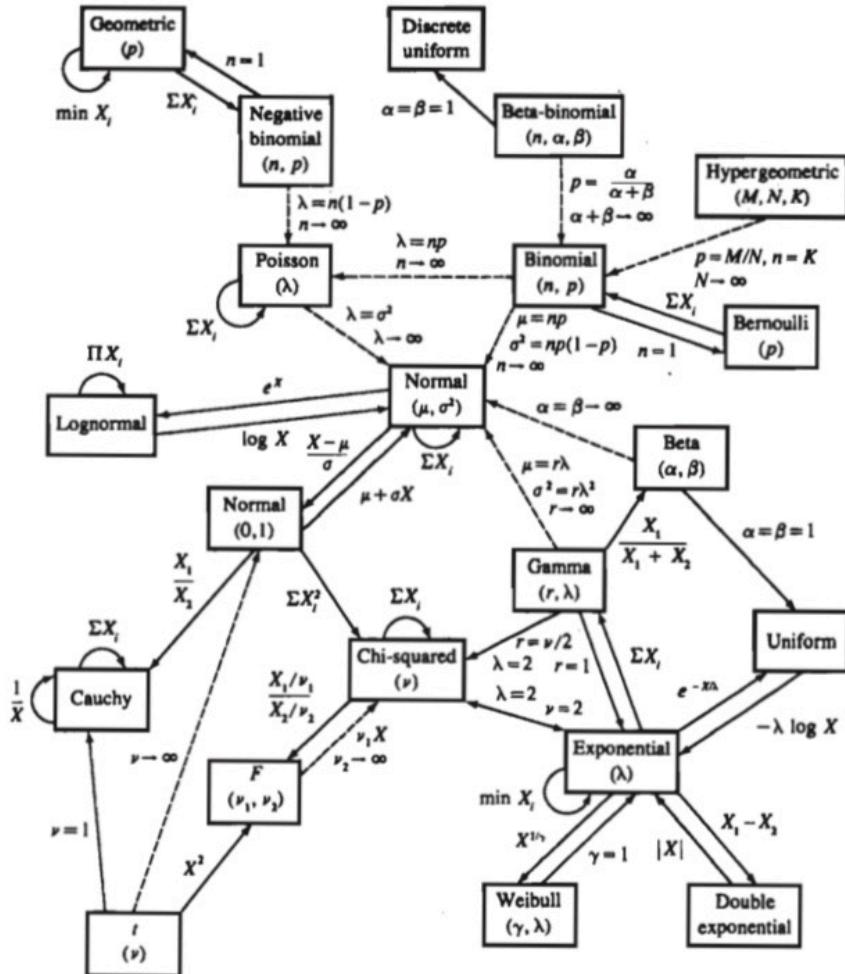


Figura 3.1: Tavola delle distribuzioni comunemente usate in statistica matematica

**Definizione 3.3.1** (Convergenza in media quadratica). Sia  $\{X_n\}_{n \in \mathbb{N}}$  una successione di variabili casuali definite su uno spazio campionario  $\Omega$  e sia  $X$  una variabile casuale anch'essa definita sul medesimo spazio campionario. La successione  $\{X_n\}_{n \in \mathbb{N}}$  si dice *convergente in media quadratica* a  $X$  se  $\{X_n\}_{n \in \mathbb{N}}$  converge a  $X$  secondo la metrica

$$d(X_n, X) = \mathbb{E}_\theta [(X_n - X)^2]$$

ovvero se (e solo se)

$$\lim_{n \rightarrow \infty} \mathbb{E}_\theta [(X_n - X)^2] = 0 \quad (3.80)$$

In maniera concisa, scriveremo  $X_n \xrightarrow{L^2} X$ .

Diamo ora una definizione di fondamentale importanza per quanto diremo nei prossimi capitoli.

**Definizione 3.3.2** (Errore Quadratico Medio). Sia  $\{T_n\}_{n \in \mathbb{N}}$  una successione di stimatori del parametro  $\theta \in \Theta$  che indicizza la distribuzione da cui proviene il campione casuale  $(X_1, X_2, \dots, X_n)$  su cui esso è calcolato. Si definisce *Errore Quadratico Medio* di  $T_n$  (o *MSE*, usando l'acronimo inglese) la funzione di  $\theta$

$$MSE_\theta(T_n) = \mathbb{E}_\theta [(T_n - \theta)^2]. \quad (3.81)$$

**Teorema 3.3.1** (Scomposizione dell'Errore Quadratico Medio). Sia  $\{T_n\}_{n \in \mathbb{N}}$  una successione di stimatori di un parametro  $\theta \in \Theta$ . Allora

$$MSE_\theta(T_n) = \text{Var}_\theta(T_n) + \mathbb{B}_\theta^2(T_n) \quad (3.82)$$

dove  $\mathbb{B}_\theta(T_n) = \mathbb{E}_\theta(T_n) - \theta$  è la *distorsione* di  $T_n$ .

*Dimostrazione.*

$$\begin{aligned} MSE_\theta(T_n) &= \mathbb{E}_\theta[(T_n - \theta)^2] \\ &= \mathbb{E}_\theta[T_n - \theta + \mathbb{E}_\theta(T_n) - \mathbb{E}_\theta(T_n)]^2 \\ &= \mathbb{E}_\theta[((T_n - \mathbb{E}_\theta(T_n)) + (\mathbb{E}_\theta(T_n) - \theta))]^2 \\ &= \mathbb{E}_\theta[(T_n - \mathbb{E}_\theta(T_n))^2] + \mathbb{E}_\theta[(\mathbb{E}_\theta(T_n) - \theta)^2] + 2\mathbb{E}_\theta[(T_n - \mathbb{E}_\theta(T_n))(\mathbb{E}_\theta(T_n) - \theta)] \\ &= \text{Var}(T_n) + [\mathbb{E}_\theta(T_n) - \theta]^2 \\ &= \text{Var}(T_n) + [\mathbb{B}_\theta(T_n)]^2 \end{aligned} \quad (3.83)$$

poiché  $\mathbb{E}_\theta[(T_n - \mathbb{E}_\theta(T_n))(\mathbb{E}_\theta(T_n) - \theta)] = (\mathbb{E}_\theta(T_n) - \theta)\mathbb{E}_\theta[(T_n - \mathbb{E}_\theta(T_n))] = 0$  dal momento che

$$\mathbb{E}_\theta[(T_n - \mathbb{E}_\theta(T_n))] = 0.$$

□

Di conseguenza,

$$T_n \xrightarrow{L^2} \theta \quad (3.84)$$

se (e solo se)

$$\lim_{n \rightarrow \infty} \text{Var}(T_n) + [\mathbb{B}_\theta(T_n)]^2 = 0 \quad (3.85)$$

ovvero se (e solo se) sia la varianza di  $T_n$  che la sua distorsione tendono a zero al crescere di  $n$  (varianza e quadrato della distorsione sono entrambe quantità non negative).

**Definizione 3.3.3** (Consistenza quadratica). Sia  $T_n$  uno stimatore del parametro  $\theta \in \Theta$ . Diremo  $T_n$  essere stimatore *quadraticamente consistente* per  $\theta$  se (e solo se)

$$T_n \xrightarrow{L^2} \theta. \quad (3.86)$$

Possiamo immediatamente verificare che media campionaria, varianza campionaria, frequenza relativa campionaria e massimo campionario sono tutti stimatori quadraticamente consistenti delle corrispondenti quantità che stimano; infatti calcolandone l'Errore Quadratico Medio, per  $n \rightarrow \infty$  si ha:

- $MSE_\mu(\bar{X}_n) = \text{Var}_\mu(\bar{X}_n) + \mathbb{B}_\mu(\bar{X}_n)^2 = \frac{\sigma^2}{n} \xrightarrow{n \rightarrow \infty} 0 \Leftrightarrow \bar{X}_n \xrightarrow{L^2} \mu$
- $MSE_{\sigma^2}(S_n^2) = \text{Var}_{\sigma^2}(S_n^2) + \mathbb{B}_{\sigma^2}(S_n^2)^2 = \frac{1}{n} \left( \mu_4 - \frac{n-3}{n-1} \mu_2^2 \right) \xrightarrow{n \rightarrow \infty} 0 \Leftrightarrow S_n^2 \xrightarrow{L^2} \sigma^2$
- $MSE_p(\hat{p}_n) = \text{Var}_p(\hat{p}_n) + \mathbb{B}_p(\hat{p}_n)^2 = \frac{p(1-p)}{n} \xrightarrow{n \rightarrow \infty} 0 \Leftrightarrow \hat{p}_n \xrightarrow{L^2} p$

- $MSE_\theta(X_{(n)}) = \mathbb{V}ar_\theta(X_{(n)}) + \mathbb{B}_\theta(X_{(n)})^2 = \frac{n\sigma^2}{(n+1)^2(n+2)} + \left(\frac{\theta}{n-1}\right)^2 \xrightarrow[n \rightarrow \infty]{} 0 \Leftrightarrow X_{(n)} \xrightarrow{L^2} \max(\mathcal{S}_X) = \theta < \infty.$

Dati due estimatori  $T_n$  e  $V_n$  del medesimo parametro  $\theta$ , di cui uno *distorto* e uno *non distorto*, per determinare quale tra i due è lo stimatore migliore per  $\theta$  possiamo confrontare i loro *MSE* piuttosto che la loro varianza, in quanto questo tiene simultaneamente conto sia della componente di *dispersione* (la varianza, appunto) che della componente di *distorsione* dello stimatore. Al contrario, se i due estimatori da confrontare sono entrambi non distorti, sarà sufficiente valutarne la varianza a cui si riduce, in questo caso, l'*MSE*.

**Esempio 3.3.1.** Consideriamo una variabile casuale  $X \sim U(0, \theta)$ ; costruire uno stimatore per il parametro  $\theta$ . Per farlo, definiamo  $T_n = 2\bar{X}_n$  e  $V_n = X_{(n)}$ . Se confrontiamo gli *MSE* dei due estimatori in questione (che abbiamo già calcolato nelle pagine precedenti), è evidente che sia da preferire  $V_n$  in quanto il suo *MSE* converge a zero più velocemente rispetto a  $T_n$ .

**Teorema 3.3.2** ( $\xrightarrow{L^2}$  implica  $\xrightarrow{P}$ ). Sia  $\{X_n\}_{n \in \mathbb{N}}$  una successione di variabili casuali che converge in media quadratica alla v.c. degenere  $X = \theta$  ossia

$$\lim_{n \rightarrow \infty} \mathbb{E}_\theta[(X_n - \theta)^2] = 0.$$

Allora  $X_n \xrightarrow{P} \theta$ .

*Dimostrazione.*

$$\begin{aligned} \mathbb{E}_\theta[(X_n - \theta)^2] &= \int_{-\infty}^{\infty} (x_n - \theta)^2 f_{X_n}(x_n; \theta) dx_n \\ &\geq \int_{\theta+\varepsilon}^{\infty} (x_n - \theta)^2 f_{X_n}(x_n; \theta) dx_n, \quad \varepsilon > 0 \\ &\geq c^2 \int_{\theta+\varepsilon}^{\infty} f_{X_n}(x_n; \theta) dx_n = c^2 P(X_n > \theta + \varepsilon) = c^2 P[(X_n - \theta) > \varepsilon] \end{aligned} \tag{3.87}$$

con  $c$  costante positiva. Ma siccome, poiché  $X_n \xrightarrow{L^2} \theta$ ,

$$\lim_{n \rightarrow \infty} \mathbb{E}_\theta[(X_n - \theta)^2] = 0 \tag{3.88}$$

allora anche

$$\lim_{n \rightarrow \infty} c^2 P[(X_n - \theta) > \varepsilon] = 0 \tag{3.89}$$

sicché

$$\lim_{n \rightarrow \infty} P[(X_n - \theta) > \varepsilon] = \lim_{n \rightarrow \infty} P[|X_n - \theta| > \varepsilon] = 0 \tag{3.90}$$

essendo  $(X_n - \theta) > 0$  e, di conseguenza,  $X_n \xrightarrow{P} \theta$ .  $\square$

**Non è però garantito il viceversa** ossia, se una successione di v.c. non converge in media quadratica a  $\theta$  nulla vieta che essa possa convergere a  $\theta$  in probabilità.

**Esempio 3.3.2.** Consideriamo una successione  $\{X_n\}_{n \in \mathbb{N}}$  di variabili causali discrete con

$$P(X_n = x) = \begin{cases} \frac{1}{n} & \text{se } x = n \\ 1 - \frac{1}{n} & \text{se } x = 0 \\ 0 & \text{altrimenti} \end{cases} \quad (3.91)$$

Vogliamo determinare il limite di  $X_n$  in media quadratica. Osserviamo che

$$\lim_{n \rightarrow \infty} P(X_n = 0) = \lim_{n \rightarrow \infty} 1 - \frac{1}{n} = 1 \quad (3.92)$$

Possiamo dunque affermare che  $\{X_n\}$  converge alla costante 0. Verifichiamo ora che tale successione converga anche in media quadratica alla medesima costante:

$$\lim_{n \rightarrow \infty} \mathbb{E}[(X_n - X)^2] = \lim_{n \rightarrow \infty} \mathbb{E}(X_n)^2 = \lim_{n \rightarrow \infty} 0^2 \left(1 - \frac{1}{n}\right) + n^2 \left(\frac{1}{n}\right) = \lim_{n \rightarrow \infty} n = +\infty \quad (3.93)$$

Dunque  $X_n$  non converge in media quadratica a  $X = 0$ . Tuttavia possiamo mostrare che vi converge in probabilità: infatti

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) \lim_{n \rightarrow \infty} P(|X_n| \geq \varepsilon) = \lim_{n \rightarrow \infty} P(X_n = n) = \lim_{n \rightarrow \infty} \frac{1}{n} = 0 \quad (3.94)$$

# 4 Un piccolo assaggio di teoria delle decisioni

## 4.1 Introduzione

Tutte le diverse forme di inferenza (stima puntuale, intervallare e test di ipotesi) spesso implicano il dover prendere decisioni. Questa cosa ci è già oramai familiare; per esempio, stimare un parametro (o un altro aspetto di una distribuzione) implica scegliere tra molti (spesso infiniti) stimatori competitori sulla base di una serie di criteri che reputiamo ragionevoli o ottimali (non distorsione e consistenza, per esempio cui rimangono associati i concetti di accuratezza e precisione).

La teoria delle decisioni è lo studio dei problemi inferenziali in cui tutte le parti del processo di decisione sono formalmente definite, includendo anche i desiderati criteri di ottimalità. Questi ultimi sono usati per confrontare diverse (e alternative) procedure decisionali.

Torniamo per un momento all'esempio 3.1.5 in cui volevamo stimare, a partire da un campione casuale di ampiezza  $n$  proveniente da una distribuzione Uniforme su  $(0, \theta]$ , proprio il parametro  $\theta$ ; in quel caso, tra i molti possibili, abbiamo preso in considerazione due stimatori (o *regole di decisione* nella filosofia della teoria delle decisioni): quello restituito dal *principio del plug-in* basato sulla  $n$ -ma statistica ordinata  $Y_{(n)} = \frac{(n+1)}{n} X_{(n)}$  e quello restituito dal *metodo dei momenti*  $T_n = 2\bar{X}_n$  e proprio in base a un criterio di ottimalità poggiato sulla varianza più piccola (= maggior precisione), abbiamo deciso di stimare  $\theta$  ricorrendo  $Y_{(n)}$ .

La *teoria delle decisioni* fornisce un metodo alternativo per analizzare problemi inferenziali quali quelli legati alla stima o alla verifica di ipotesi circa aspetti di interesse sfruttando l'informazione contenuta nella realizzazione di un campione casuale proveniente da una qualche popolazione; in molti casi finisce col fornire conclusioni simili a quelle che abbiamo finora ottenuto nelle pagine precedenti sebbene talvolta possa fornire sorprendenti nuove suggestioni e soluzioni.

## 4.2 Regole di decisione

In un contesto formale di teoria delle decisioni tutti gli elementi coinvolti nel suo disegno devono essere specificati; più precisamente, ci riferiamo a:

- dati*: questi sono descritti tramite il vettore casuale  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  le cui determinazioni  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  costituiscono lo lo spazio campionario  $\mathfrak{X}$
- modello*: esso è costituito dall'insieme delle distribuzioni di probabilità per  $\mathbf{X}$ , indicizzate dal parametro  $\theta \in \Theta \subseteq \mathbb{R}^k$ ,  $k \in \mathbb{N}$

c) *parametro  $\theta$* : esso rappresenta il vero ma *incognito stato di natura* sul quale vogliamo fare inferenza; come abbiamo già visto  $\theta$  può essere uno *scalare* in  $\mathbb{R}$  o un *vettore* in  $\mathbb{R}^k$ ; l'insieme di tutti i possibili valori di  $\theta$ , indicato con  $\Theta$ , è detto *spazio parametrico*. Di conseguenza il modello a cui facevamo riferimento in b) è restituito dall'insieme ovvero,

$$\mathcal{F}_\theta = \{f_X(\mathbf{x}; \theta), \theta \in \Theta\}$$

dove  $f_X(\mathbf{x}; \theta)$  è una *funzione di massa* o una *funzione di densità* su  $\mathfrak{X}$ .

Una volta che si sono ottenuti i dati e che sono stati opportunamente trasformati in informazioni (statistiche) disponiamo di tutti gli elementi per *prendere una decisione* in merito al problema inferenziale cui siamo interessati e a cui essi si riferiscono.

**Definizione 4.2.1** (Spazio delle decisioni). L'insieme di tutte le possibili decisioni prende il nome di *spazio delle decisioni* o anche *spazio delle azioni* tenuto conto del fatto che *ogni* decisione implica una *azione*. Indicheremo tale spazio con  $\mathcal{A}$ .

La *struttura* dello spazio delle azioni  $\mathcal{A}$  determina il *tipo* di inferenza che si intende fare (*stima puntuale*, *stima intervallare* o *test di ipotesi*). Per esempio, in ambito di *stima puntuale*,  $\mathcal{A} = \Theta$  dal momento che le possibili *azioni* in questo contesto coincidono con *indovinare* (meglio, *intuire alla luce dell'informazione disponibile*) il valore di  $\theta \in \Theta$ .

In un problema di stima, per esempio, l'azione  $a$  altro non sarà che lo stimatore di  $\theta$  che scegliamo appunto per stimare il parametro  $\theta$  che indica la distribuzione da cui provengono i dati; mentre in un problema di verifica delle ipotesi, l'azione  $a$  altro non sarà che quella di rifiutare o non rifiutare l'ipotesi di interesse sempre alla luce dell'informazione contenuta nei dati.

## 4.3 Perdita e rischio

Ogni decisione comporta una scelta tra due o più alternative, ciascuna delle quali avrà delle conseguenze che dipendono dallo *stato di natura* nel quale il processo decisionale si svolge, ovvero dal suo contesto. Una decisione è pertanto costituita da *stati di natura*, *azioni* e *conseguenze*.

Ora se  $\theta \in \Theta$  è il *vero stato di natura* che in generale è parzialmente o totalmente incognito, l'azione  $a \in \mathcal{A}$ , intrapresa alla luce dell'informazione disponibile può essere *corretta*, *sbagliata*, *non del tutto corretta* o *non del tutto sbagliata*: questa situazione relativa alla relazione tra  $a$  e  $\theta$  è *quantificata* da una particolare funzione che prende il nome di *funzione di perdita* (o *loss function*), usualmente indicata con  $loss(a, \theta)$ , la cui finalità è quella di *misurare* la gravità della perdita in cui si incorre nel compiere l'azione  $a$  quando invece lo stato di natura è  $\theta$ .

**Definizione 4.3.1** (Funzione di perdita). Cosiderati uno spazio parametrico  $\Theta$  e lo spazio delle azioni  $\mathcal{A}$ , una funzione di perdita è data da

$$loss(a, \theta) : \Theta \times \mathcal{A} \mapsto [0, +\infty)$$

caratterizzata dalle seguenti proprietà:

- $loss(a, \theta) \geq 0, \quad \forall \theta \in \Theta \text{ e } \forall a \in \mathcal{A}$

- se  $\theta = a$  allora  $loss(a, \theta) = 0$
- $loss(a, \theta)$  è una funzione *non decrescente* della *distanza* tra  $a$  e  $\theta$  (e in tal senso, misura la serità o gravità della perdita patita agendo con l'azione  $a$  quando il vero stato di natura è  $\theta$ .)

**Esempio 4.3.1.** Di seguito due funzioni di perdita spesso usate:

- $loss(a, \theta) = |\theta - a|$  (*Absolute loss*)
- $loss(a, \theta) = (\theta - a)^2$  (*Quadratic loss*)

Alla base di una qualsiasi azione vi è una *regola di decisione*; quest'ultima è una regola che specifica, per ogni  $x \in \mathfrak{X}$ , quale azione  $a \in \mathcal{A}$  sarà presa quando  $\mathbf{X} = x$  sia stato osservato.

Indicata co  $\delta(x)$  la *regola di decisione*, questa è una funzione da  $\mathfrak{X}$  in  $\mathcal{A}$ :

$$\delta(x) = \mathfrak{X} \mapsto \mathcal{A}$$

e converremo indicare con  $\mathcal{D}$  l'insieme di tutte le possibili *regole di decisione*,

In generale, vorremmo che  $\mathcal{D}$  sia il più grande possibile in modo che tutte le regole di decisione siano al suo interno; per esempio,  $\mathcal{D}$  potrebbe essere costituito da tutte le funzioni da  $\mathfrak{X}$  in  $\mathcal{A}$ . Ma ci possiamo subito rendere conto del fatto che, così definito  $\mathcal{D}$  finisce per essere troppo grande per essere operativo; sarà dunque necessario introdurre un *set di proprietà di ottimalità* per stabilire quali siano le regole di decisione  $\delta(x) \in \mathcal{D}$  che è ragionevole considerare in quanto sono *buone* regole di decisione alla luce del *set di proprietà di ottimalità*.

In ambito di teoria delle decisioni, la *qualità* di una *regola di decisione* è quantificata da una funzione associata alla regola di decisione stessa che prende il nome di *funzione di rischio*

**Definizione 4.3.2** (Funzione di rischio). Per una regola di decisione  $\delta(x)$  diremo la funzione di  $\theta$

$$Risk(\theta, \delta) = \mathbb{E}_\theta [loss(\theta, \delta(\mathbf{X}))] \quad (4.1)$$

*funzione di rischio*. Essa esprime la *perdita attesa* associata alla regola di decisione  $\delta(x)$

Nell'ambito di un problema di stima, sia  $T_n(\mathbf{X})$  uno stimatore per il parametro  $\theta \in \Theta$  vale a dire una regola di decisione in merito a  $\theta$ ; se assumiamo una funzione di perdita *quadratica* del tipo

$$loss(\theta, T_n(\mathbf{X})) = [T_n(\mathbf{X}) - \theta]^2$$

avremo una funzione di rischio così definita

$$\begin{aligned} Risk(\theta, T_n(\mathbf{X})) &= \mathbb{E}_\theta [loss(\theta, T_n(\mathbf{X}))] \\ &= \mathbb{E}_\theta [(T_n(\mathbf{X}) - \theta)^2] \\ &= MSE_\theta(T_n) \end{aligned} \quad (4.2)$$

dove la funzione di  $\theta$   $MSE_\theta(T_n)$  prende, come già sappiamo, il nome di *Errore Quadratico Medio* di  $T_n$ . Poiché dal Teorema 3.3.1 si ha

$$MSE_\theta(T_n) = \text{Var}_\theta(T_n) + \mathbb{B}_\theta^2(T_n) \quad (4.3)$$

l'*MSE* è in grado di tenere contemporaneamente conto sia della componente di dispersione che di quella di distorsione dello stimatore  $T_n$  e quindi, ipotizzata una funzione di perdita quadratica, costituisce lo strumento naturale per *confrontare* tra loro stimatori distorti e non distorti per il parametro  $\theta \in \Theta$  al fine di stabilire quale sia lo stimatore *relativamente* migliore. Resta immediatamente inteso che, nel caso di stimatore non distorto, l'errore quadratico medio di riduce alla varianza dello stimatore.

# 5 Metodi elementari di inferenza statistica

Finora ci siamo limitati a considerare *stimatori puntuali*, ancorché vettoriali, basati sul *principio del plug-in* o sul *metodo dei momenti* e l'inferenza che abbiamo sviluppato in merito al parametro  $\theta \in \Theta$  consisteva nell'indovinare (=stimare), sulla base dell'informazione campionaria sinetizzata in  $T_n(X_1, X_2, \dots, X_n)$ , il valore del parametro  $\theta$  o di funzioni  $g(\theta)$ .

Ma che cosa siamo in grado di dire in merito alla *affidabilità* delle *stime*, intendendo con *stima* il *valore che uno stimatore assume su una effettiva determinazione*  $(x_1, x_2, \dots, x_n)$  del campione casuale  $(X_1, X_2, \dots, X_n)$  quali sono media campionaria  $\bar{x}_n$ , varianza campionaria  $s_n^2$ , frequenza relativa campionaria  $\hat{p}_n$ , massimo campionario  $x_{(n)}$  ecc.?

Infatti, la semplice stima puntuale, per esempio  $\bar{x}_n$  di  $\mu$ , non contiene alcuna informazione in merito al suo potenziale *scostamento* da  $\mu$  né un qualche strumento che permetta di *controllare* tale scostamento. Dovremo quindi affrontare il problema della costruzione di procedure di stima in cui sia possibile *fissare a priori* un certo grado di *massimo scostamento da  $\theta$* , quantità che si intende stimare, che sia garantito da una *prefissata* probabilità. In tal modo finiremo per qualificare il *grado di affidabilità* della stima stessa. Il tutto in probabilità.

## 5.1 Statistiche pivot

Cominciamo col dare la definizione di quella che diremo essere una *statistica pivot* e che giocherà un ruolo fondamentale nella costruzione di procedure inferenziali.

**Definizione 5.1.1** (Statistica pivot). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla generica distribuzione  $F_X(x; \theta)$ . Diremo *statistica pivot* la funzione  $Q(X_1, \dots, X_n; \theta)$  che soddisfa i seguenti requisiti:

- $Q$  è funzione del campione casuale  $(X_1, X_2, \dots, X_n)$  e del parametro su cui verte il problema inferenziale;
- $Q$  non contiene altri parametri ignoti al di fuori di  $\theta$ ;
- la distribuzione di  $Q$ ,  $F_Q(q)$  è *parameter free* ovvero è completamente nota;
- $Q$  è invertibile rispetto a  $\theta$ .

**Esempio 5.1.1.** Dato un campione casuale  $(X_1, X_2, \dots, X_n)$  da una distribuzione Normale  $N(\mu, \sigma^2)$ , è possibile costruire delle statistiche pivot per fare inferenza

sui parametri  $(\mu, \sigma^2)$  della distribuzione. In particolare, le seguenti statistiche soddisfano a quanto richiesto nella definizione 5.1.1:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad V_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2 \quad W_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1} \quad (5.1)$$

dove, ipotizzando la varianza  $\sigma^2$  della popolazione essere *nota*,  $Z_n$  restituisce una statistica pivot su cui basare l'inferenza sulla media  $\mu$  della popolazione mentre  $V_n$  è a sua volta una statistica pivot su cui basare l'inferenza sulla varianza  $\sigma^2$  della popolazione. Dimostreremo, inoltre, che  $W_n$  segue una distribuzione *t di Student* con  $\nu = n - 1$  gradi di libertà e che fornisce la statistica pivot per l'inferenza sulla media  $\mu$  della popolazione qualora la varianza  $\sigma^2$  della popolazione *non sia nota*. La scelta di stimare  $\sigma$  tramite  $S_n$  è giustificata dal fatto che  $S_n^2$  è uno stimatore *non distorto* e *consistente* di  $\sigma^2$  e dall'applicazione del Teorema 3.1.6 (Continuous mapping) che permette di concludere che  $S_n \xrightarrow{P} \sigma^2$  ergo  $S_n$  è a sua volta stimatore *consistente* di  $\sigma$ .

## 5.2 Campionamento da popolazione Normale

Supponiamo che il campione casuale  $(X_1, X_2, \dots, X_n)$  provenga da una popolazione avente distribuzione Normale di media  $\mu$  e varianza  $\sigma^2$ . Cominciamo con l'enunciare e dimostrare un teorema di cruciale importanza per quello che diremo nelle prossime pagine.

**Teorema 5.2.1** (Indipendenza di media e varianza campionaria). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione Normale  $N(\mu, \sigma^2)$  e siano  $\bar{X}_n$  e  $S_n^2$  rispettivamente media e varianza campionaria. Allora

a)  $\bar{X}_n$  e  $S_n^2$  sono indipendenti

b)  $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

c)  $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$

*Dimostrazione.* Dimostriamo nell'ordine le tre tesi del teorema

*Parte a)*

Senza perdita alcuna di generalità, tenuto conto del fatto che la distribuzione Normale appartiene a famiglia di locazione e scala, possiamo assumere  $\mu = 0$  e  $\sigma^2 = 1$ . Ora,

$$\begin{aligned} S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n-1} \left[ (X_1 - \bar{X}_n)^2 + \sum_{i=2}^n (X_i - \bar{X}_n)^2 \right] \\ &= \frac{1}{n-1} \left\{ \left[ \sum_{i=2}^n (X_i - \bar{X}_n) \right]^2 + \sum_{i=2}^n (X_i - \bar{X}_n)^2 \right\} \end{aligned}$$

poichè

$$0 = \sum_{i=1}^n (X_i - \bar{X}_n) = (X_1 - \bar{X}_n) + \sum_{i=2}^n (X_i - \bar{X}_n)$$

e perciò

$$(X_1 - \bar{X}_n) = - \sum_{i=2}^n (X_i - \bar{X}_n).$$

Di conseguenza,  $S_n^2$  può essere scritta come funzione delle sole

$$(X_2 - \bar{X}_n), (X_3 - \bar{X}_n), \dots, (X_n - \bar{X}_n)$$

e sarà sufficiente dimostrare che queste ultime sono indipendenti da  $\bar{X}_n$  per completare la dimostrazione della parte a) del teorema. La densità congiunta del campione casuale  $X_1, X_2, \dots, X_n$  è

$$f_{X_1, \dots, X_n}(x_1, \dots, y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x_1^2 + \sum_{i=2}^n x_i^2)}$$

e introducendo la trasformazione

$$y_1 = \bar{x}_n, \quad y_2 = (x_2 - \bar{x}_n), \quad y_3 = (x_3 - \bar{x}_n), \quad \dots, \quad y_n = (x_n - \bar{x}_n)$$

che è lineare e con determinante Jacobiano uguale a  $n$ , si ha

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \frac{n}{(2\pi)^{n/2}} e^{-\frac{1}{2}((y_1 - \sum_{i=2}^n y_i)^2 + \sum_{i=1}^n (y_1 + y_i)^2)}$$

essendo

$$x_1 = n\bar{x}_n - \sum_{i=2}^n x_i = ny_1 - \sum_{i=2}^n (y_1 + y_i) = ny_1 - \sum_{i=2}^n y_i + (n-1)y_1 = y_1 - \sum_{i=2}^n y_i$$

da cui

$$x_1^2 = \left( y_1 - \sum_{i=2}^n y_i \right)^2$$

e

$$x_i = y_1 + y_i, \quad \forall i \geq 2$$

perciò

$$\sum_{i=2}^n x_i^2 = \sum_{i=2}^n (y_1 + y_i)^2.$$

Ma

$$(y_1 - \sum_{i=2}^n y_i)^2 + \sum_{i=1}^n (y_1 + y_i)^2 = ny_1^2 + \left( \sum_{i=2}^n y_i^2 - \left( \sum_{i=2}^n y_i \right)^2 \right)$$

sicchè

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{n}{2}y_1^2} \cdot \frac{\sqrt{n}}{(2\pi)^{(n-1)/2}} e^{-\frac{1}{2}\left\{ \sum_{i=2}^n y_i^2 + (\sum_{i=2}^n y_i)^2 \right\}}$$

Poichè la densità congiunta di  $Y_1, Y_2, \dots, Y_n$  fattorizza, segue che  $Y_1$  è indipendente da  $(Y_2, Y_3, \dots, Y_n)$  e ciò comporta che  $Y_1 = \bar{X}_n$  è indipendente da  $S_n^2 = \frac{1}{n-1} \left\{ [\sum_{i=2}^n (X_i - \bar{X}_n)]^2 + \sum_{i=2}^n (X_i - \bar{X}_n)^2 \right\} = g(Y_2, \dots, Y_n)$  come affermato nella parte a) del teorema.

*Parte b)*

È sufficiente ricordare che la funzione generatrice dei momenti di  $\bar{X}_n$  è uguale a

$$\begin{aligned} M_{\bar{X}_n}(t) &= \mathbb{E} \left( e^{t\bar{X}_n} \right) = \mathbb{E} \left( e^{t\frac{1}{n} \sum_{i=1}^n X_i} \right) = \mathbb{E} \left( e^{\sum_{i=1}^n \frac{t}{n} X_i} \right) \\ &= \prod_{i=1}^n \mathbb{E} \left( e^{\frac{t}{n} X_i} \right) = \left( e^{\frac{t}{n}\mu + \frac{t^2}{2n^2}\sigma^2} \right)^n = e^{t\mu + \frac{t^2}{2}\frac{\sigma^2}{n}} \end{aligned}$$

in virtù dell'indipendenza e dell'identica distribuzione di  $(X_1, X_2, \dots, X_n)$  e riconoscendo nell'espressione finale la funzione generatrice dei momenti di una distribuzione  $N(\mu, \frac{\sigma^2}{n})$ , in ragione del legame che lega funzione generatrice (quando esiste, come in questo caso) e funzione di distribuzione di probabilità di una v.c., anche la parte b) del teorema risulta così dimostrata.

*Parte c)*

Siano  $\bar{X}_k$  e  $S_k^2$  rispettivamente la media campionaria e la varianza campionaria dei primi  $k$  elementi del campione casuale. Si può provare che

$$(n-1) S_n^2 = (n-2) S_{n-1}^2 + \left( \frac{n-1}{n} \right) (X_n - \bar{X}_{n-1})^2. \quad (5.2)$$

Fissiamo ora  $n = 2$ ; allora, per la (5.2), possiamo scrivere

$$(2-1)S_2^2 = (2-2)S_1^2 + \frac{2-1}{2} (X_2 - X_1)^2 = \frac{1}{2} (X_2 - X_1)^2$$

e, ricordando che  $X_i \sim N(\mu, \sigma^2)$ , si ha  $(X_2 - X_1) \sim N(0, 4\sigma^2)$  e dunque

$$\frac{(X_2 - X_1)}{2\sigma} \sim N(0, 1)$$

per cui

$$\frac{(2-1)S_2^2}{\sigma^2} = \frac{S_2^2}{\sigma^2} = \frac{(X_2 - X_1)^2}{2\sigma^2} \sim \chi_1^2.$$

Procedendo per induzione, assumiamo che per  $n = k$ ,

$$\frac{(k-1)S_k^2}{\sigma^2} \sim \chi_{k-1}^2.$$

Per  $n = k+1$ , dalla (5.2), otteniamo

$$kS_{k+1}^2 = (k-1)S_k^2 + \left( \frac{k}{k+1} \right) (X_{k+1} - \bar{X}_k)^2$$

e, in accordo con l'ipotesi di induzione,

$$\frac{(k-1) S_k^2}{\sigma^2} \sim \chi_{k-1}^2.$$

Ora, se riusciamo a dimostrare che

$$\left( \frac{k}{k+1} \right) \frac{(X_{k+1} - \bar{X}_k)^2}{\sigma^2} \sim \chi_1^2$$

ed è indipendente da  $S_k^2$ , per il Teorema 5.2.3 ovvero, in buona sostanza, per la proprietà di riproducibilità di cui gode la distribuzione Chi-quadrato, avremo provato che

$$\frac{k S_{k+1}^2}{\sigma^2} \sim \chi_k^2$$

e il teorema sarà così dimostrato.

L'indipendenza di  $(X_{k+1} - \bar{X}_k)^2$  e  $S_k^2$  segue da un noto teorema (funzioni di v.c. indipendenti sono, a loro volta, v.c. indipendenti). Il vettore  $(X_{k+1}, \bar{X}_k)$  è indipendente da  $S_k^2$  e così pure ogni funzione del vettore stesso quale, per esempio,  $(X_{k+1} - \bar{X}_k)^2$ . Inoltre,  $(X_{k+1} - \bar{X}_k)$  segue una distribuzione Normale di media 0 e varianza

$$\text{Var}(X_{k+1} - \bar{X}_k) = \frac{k+1}{k} \sigma^2$$

e di conseguenza

$$\left( \frac{k}{k+1} \right) \frac{(X_{k+1} - \bar{X}_k)^2}{\sigma^2} \sim \chi_1^2$$

sicchè anche la parte c) del teorema è dimostrata.

□

## 5.2.1 Distribuzioni campionarie

Nelle righe che seguono daremo alcuni dettagli su alcune distribuzioni associate a opportune statistiche costruite su campioni casuali provenienti da popolazioni Normali; a tali distribuzioni si dà comunemente il nome di *distribuzioni campionarie* e vedremo in seguito il ruolo da esse giocato nell'ambito della statistica matematica.

### 5.2.1.1 Distribuzione chi-quadrato

Abbiamo già avuto modo di incontrare questa distribuzione di probabilità e precisamente nell'Esempio 1.2.6 nell'ambito del teorema di trasformazione di v.c.

**Teorema 5.2.2** (Distribuzione chi-quadrato). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una popolazione avente distribuzione  $N(\mu, \sigma^2)$  e sia  $Z_i = \frac{(X_i - \mu)}{\sigma} \sim N(0, 1)$ . Allora,

$$V = \sum_{i=1}^n Z_i^2 \sim \chi_n^2 \tag{5.3}$$

la cui funzione di densità è data da

$$f_V(v; n) = \frac{1}{\Gamma(n/2)2^{n/2}} v^{n/2-1} e^{-(v/2)} \mathbb{1}_{\mathbb{R}^+}(v), \quad n \in \mathbb{N} \quad (5.4)$$

dove  $\nu = n$  sono i *gradi di libertà*, parametro che indica la distribuzione chi-quadrato.

*Dimostrazione.* Si veda Esempio 1.2.6 □

Ora,

- a) se  $V \sim \chi_\nu^2$  allora  $\mathbb{E}_\nu(V) = \nu = n$  e  $\text{Var}_\nu(V) = 2\nu = 2n$
- b)  $\chi_\nu^2$  è un caso particolare della distribuzione Gamma (con  $\alpha = \nu/2$  e  $\beta = 2$ )
- c) Vale il seguente

**Teorema 5.2.3** (Teorema di Cochran). Siano  $V_1 \sim \chi_{\nu_1}^2$  e  $V_2 \sim \chi_{\nu_2}^2$ , indipendenti. Allora

$$V_1 + V_2 \sim \chi_{\nu_1 + \nu_2}^2. \quad (5.5)$$

*Dimostrazione.* La dimostrazione segue agevolmente applicando la proprietà di riproducibilità (vedi Teorema 1.3.6). □

Il teorema 5.2.3 si può immediatamente estendere al caso di una somma di  $k$  v.c. chi-quadrato indipendenti.

### 5.2.1.2 Distribuzione $t$ di Student

**Teorema 5.2.4** (Distribuzione  $t$  di Student). Sia  $Z$  una variabile casuale Normale standard e  $V$  una variabile casuale  $\chi_\nu^2$ . Inoltre, siano  $Z$  e  $V$  indipendenti. Allora la variabile casuale

$$T = \frac{Z}{\sqrt{\frac{V}{\nu}}} \quad (5.6)$$

segue una *distribuzione  $t$  di Student* con  $\nu \in \mathbb{N}$  gradi di libertà, la cui densità è data da

$$f_T(t; \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(\frac{1}{1+t^2\nu}\right)^{\frac{\nu+1}{2}}} \mathbb{1}_{\mathbb{R}}(t) \quad (5.7)$$

La distribuzione  $t$  di Student è *simmetrica* e ha *code più pesanti* della distribuzione Normale standard.

*Dimostrazione.* La funzione di densità congiunta di  $Z$  e  $V$ , essendo queste ultime indipendenti, è data dal prodotto delle rispettive densità, vale a dire

$$f_{Z,V}(z, v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \cdot \frac{1}{\Gamma(\frac{\nu}{2})} 2^{\frac{\nu}{2}} v^{\frac{\nu}{2}-1} e^{-\frac{v}{2}} \mathbb{1}_{\mathbb{R}}(z) \mathbb{1}_{\mathbb{R}^+}(v).$$

Operando la trasformazione

$$T = \frac{Z}{\sqrt{\frac{V}{\nu}}}, \quad W = V$$

il cui determinante Jacobiano è dato da  $(\frac{w}{\nu})^{\frac{1}{2}}$ , sostituendo  $z = t \sqrt{\frac{w}{\nu}}$  e  $v = w$  nella densità congiunta di  $Z$  e  $V$  e moltiplicato il tutto per il determinante Jacobiano poc'anzi ottenuto, si ottiene la funzione di densità congiunta di  $T$  e  $W$

$$f_{T,W}(t, w) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{\nu}{2})} \frac{1}{2^{\frac{\nu}{2}}} e^{-\frac{1}{2}t^2 \frac{w}{\nu}} w^{\frac{\nu}{2}-1} e^{-\frac{w}{2}} \left(\frac{w}{\nu}\right)^{\frac{1}{2}} \mathbb{1}_{\mathbb{R}}(t) \mathbb{1}_{\mathbb{R}^+}(w).$$

La funzione di densità marginale di  $T$  è data da

$$\begin{aligned} f_T(t) &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{\nu}{2})} \frac{1}{2^{\frac{\nu}{2}}} e^{-\frac{1}{2}t^2 \frac{w}{\nu}} w^{\frac{\nu}{2}-1} e^{-\frac{w}{2}} \left(\frac{w}{\nu}\right)^{\frac{1}{2}} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{\nu}{2})} \frac{1}{2^{\frac{\nu}{2}}} \int_0^\infty e^{-\frac{1}{2}t^2 \frac{w}{\nu}} w^{\frac{\nu}{2}-1} e^{-\frac{w}{2}} \left(\frac{w}{\nu}\right)^{\frac{1}{2}} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{\nu}{2})} \frac{1}{2^{\frac{\nu}{2}} \nu^{\frac{1}{2}}} \int_0^\infty w^{(\frac{\nu+1}{2})-1} e^{-\frac{1}{2}(1+\frac{t^2}{\nu})w} dw. \end{aligned}$$

Subito possiamo riconoscere nell'integrandi il *kernel* di una distribuzione Gamma di parametri

$$\alpha = \left(\frac{\nu+1}{2}\right) \text{ e } \beta = \frac{2}{\left(1 + \frac{t^2}{\nu}\right)}$$

il cui integrale su  $[0, \infty)$  restituisce la costante di normalizzazione

$$\Gamma(\alpha) \beta^\alpha = \Gamma\left(\frac{\nu+1}{2}\right) \left[\frac{2}{1 + \frac{t^2}{\nu}}\right]^{\frac{\nu+1}{2}}$$

sicché, infine, otteniamo

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{\nu}{2})} \frac{1}{2^{\frac{\nu}{2}} \nu^{\frac{1}{2}}} \Gamma\left(\frac{\nu+1}{2}\right) \left[\frac{2}{1 + \frac{t^2}{\nu}}\right]^{\frac{\nu+1}{2}} \mathbb{1}_{\mathbb{R}}(t) \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\pi\nu}} \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} \mathbb{1}_{\mathbb{R}}(t). \end{aligned}$$

□

**Proposizione.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $N(\mu, \sigma^2)$ . La quantità  $T = (\bar{X}_n - \mu)/(S_n/\sqrt{n})$  segue una distribuzione  $t$  di Student di parametro  $\nu = n - 1$ .

Infatti, si può scrivere

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \left( \sqrt{\frac{S_n^2}{\sigma^2}} \right)^{-1} \quad (5.8)$$

e per il Teorema 5.2.4 si ha subito la tesi.

### 5.2.1.3 Distribuzione di Fisher-Snedecor

**Teorema 5.2.5** (Distribuzione di Fisher-Snedecor). Siano  $U$  e  $V$  due variabili casuali chi-quadrato indipendenti, rispettivamente con  $\nu_1$  e  $\nu_2$  gradi di libertà, dove  $\nu_1, \nu_2 \in \mathbb{N}$ . Allora la variabile casuale

$$W = \frac{U/\nu_1}{V/\nu_2}$$

ha *distribuzione F di Fisher-Snedecor* con  $\nu_1$  e  $\nu_2$  gradi di libertà, a cui corrisponde densità

$$f_W(w; \nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} w^{\frac{\nu_1}{2}-1} \frac{1}{\left[1 + \frac{\nu_1}{\nu_2} w\right]^{\frac{\nu_1+\nu_2}{2}}} \mathbb{1}_{\mathbb{R}^+}(w).$$

*Dimostrazione.* Dal momento che  $U$  e  $V$  sono variabili casuali *indipendenti*, la densità congiunta di  $(U, V)$  risulta uguale al prodotto delle loro densità, ossia

$$f_{U,V}(u, v; \nu_1, \nu_2) = \frac{1}{\Gamma\left(\frac{\nu_1}{2}\right) 2^{\frac{\nu_1}{2}}} u^{\frac{\nu_1}{2}-1} e^{-\frac{u}{2}} \cdot \frac{1}{\Gamma\left(\frac{\nu_2}{2}\right) 2^{\frac{\nu_2}{2}}} v^{\frac{\nu_2}{2}-1} e^{-\frac{v}{2}} \mathbb{1}_{\mathbb{R}^+}(u, v). \quad (5.9)$$

Dalla precedente si ricava facilmente la distribuzione della nuova variabile casuale  $Y = U/V$ ; ciò può essere fatto ricorrendo al teorema di trasformazione per vettori casuali (nel caso specifico, per vettori bidimensionali), ponendo  $Y = U/V$  e  $Z = V$  ottenendo  $U = Y \cdot Z$ ,  $V = Z$  mentre il determinante Jacobiano associato alla trasformazione risulta uguale a  $|J| = \begin{vmatrix} z & y \\ 0 & 1 \end{vmatrix} = z$ . Sostituendo ora  $u = yz$  e  $v = z$  in (5.9), ponendo per semplicità di notazione  $m = \frac{\nu_1}{2}$  e  $n = \frac{\nu_2}{2}$ , si ha

$$\begin{aligned} f_{Y,Z}(y, z; 2m, 2n) &= \frac{1}{\Gamma(m) \Gamma(n) 2^{m+n}} (yz)^{m-1} e^{-\frac{yz}{2}} z^{n-1} e^{-\frac{z}{2}} \cdot z \mathbb{1}_{\mathbb{R}^+}(y, z) \\ &= \frac{1}{\Gamma(m) \Gamma(n) 2^{m+n}} y^{m-1} z^{(m+n)-1} e^{-\frac{z}{2}(y+1)} \mathbb{1}_{\mathbb{R}^+}(y, z) \end{aligned} \quad (5.10)$$

sicché la densità (marginale) di  $Y$  potrà essere ottenuta integrando (5.10) rispetto  $z$ , ossia

$$f_Y(y; 2m, 2n) = \frac{y^{m-1}}{\Gamma(m) \Gamma(n) 2^{m+n}} \int_0^{+\infty} z^{(m+n)-1} e^{-\frac{z}{2}(1+y)} dz.$$

Effettuando la sostituzione  $t = \frac{z}{2}(1+y)$  nel precedente integrale, si ottiene

$$\begin{aligned} f_Y(y; 2m, 2n) &= \frac{y^{m-1}}{\Gamma(m) \Gamma(n) 2^{m+n}} \int_0^{+\infty} \frac{2^{(m+n)-1}}{(1+y)^{m+n-1}} t^{(m+n)-1} e^{-t} \frac{2}{(1+y)} dt \\ &= \frac{1}{\Gamma(m) \Gamma(n)} \frac{y^{m-1}}{(1+y)^{m+n}} \int_0^{+\infty} t^{(m+n)-1} e^{-t} dt \\ &= \frac{\Gamma(m+n)}{\Gamma(m) \Gamma(n)} y^{m-1} \frac{1}{(1+y)^{m+n}} \mathbb{1}_{\mathbb{R}^+}(y), \end{aligned} \quad (5.11)$$

poichè

$$\int_0^{+\infty} t^{(m+n)-1} e^{-t} dt = \Gamma(m+n)$$

ovvero, l'integrale in questione restituisce (per sua stessa definizione) la funzione Gamma valutata nel punto  $(m + n)$ .

Ricordando che  $Y = \frac{U}{V}$ , per completare la dimostrazione, rimane allora solo da effettuare un'ultima trasformazione in (5.11), ponendo  $W = \frac{n}{m} Y$  da cui  $y = \frac{m}{n} w$  e  $\frac{dy}{dw} = \frac{m}{n}$ ; in tal modo si ottiene la funzione di densità di  $W$  data da:

$$f_W(w; 2m, 2n) = \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} \left(\frac{m}{n}\right)^m w^{m-1} \frac{1}{\left[1 + \frac{m}{n}w\right]^{m+n}}.$$

Ricordando che, per semplicità di notazione, abbiamo posto  $m = \frac{\nu_1}{2}$  e  $n = \frac{\nu_2}{2}$ , possiamo infine riscrivere la precedente nel seguente modo

$$f_W(w; \nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} w^{\frac{\nu_1}{2}-1} \frac{1}{\left[1 + \frac{\nu_1}{\nu_2}w\right]^{\frac{\nu_1+\nu_2}{2}}} \mathbb{1}_{\mathbb{R}^+}(w)$$

a riprova del fatto che

$$W = \frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1, \nu_2}.$$

□

La distribuzione di Fisher-Snedecor risulta particolarmente utile nella definizione di una statistica pivot per il rapporto tra varianze di popolazioni diverse. Abbiamo già mostrato che, nell'ambito di campionamento da distribuzione Normale (vedi Teorema 5.2.1 - punto c), la *statistica pivot*  $V = \frac{(n-1)S_n^2}{\sigma^2}$  segue una distribuzione  $\chi_{n-1}^2$ .

Date dunque due variabili casuali *indipendenti*

$$V_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2 \quad \text{e} \quad V_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi_{n_2-1}^2 \quad (5.12)$$

per il teorema 5.2.5 possiamo dire che il rapporto tra  $V_1$  e  $V_2$  divise per i rispettivi gradi di libertà, segue la distribuzione di Snedecor-Fisher con  $(n_1 - 1, n_2 - 1)$  gradi di libertà, ossia

$$W = \frac{V_1/(n_1 - 1)}{V_2/(n_2 - 1)} \sim F_{n_1-1, n_2-1} \quad (5.13)$$

**Proposizione.** La distribuzione di Fisher-Snedecor gode delle seguenti proprietà:

1.  $W \sim F_{v_1, v_2}$  allora  $\frac{1}{W} \sim F_{v_2, v_1}$ ;
2.  $q_{v_1, v_2; 1-\alpha} = \frac{1}{q_{v_2, v_1; \alpha}}$ .

Quest'ultima identità risulta di fondamentale importanza alla luce del fatto che la distribuzione  $F$  di Fisher-Snedecor è stata tabulata per *piccoli valori*  $\alpha$  dell'area sottesa nella coda di destra della distribuzione stessa; grazie a questa identità abbiamo modo di trovare, con una semplice operazione aritmetica, i quantili di *ordine elevato* (per esempio, di ordine  $(1 - \alpha)$ ) della distribuzione in questione, anche se non direttamente tabulati.

## 5.3 Intervalli di confidenza

La procedura per la costruzione di un intervallo di confidenza per  $\theta \in \Theta$ , detto anche *stimatore intervallare* di  $\theta$ , cui rimane associata un *prefissato livello di confidenza*  $(1 - \alpha)$ ,  $\alpha \in (0, 1)$  consiste in

- individuare una statistica pivot  $Q((X_1, X_2, \dots, X_n), \theta)$  che potremmo anche riscrivere in termini di  $Q(T_n, \theta)$  con  $T_n = T_n(X_1, X_2, \dots, X_n)$  sulla base di uno stimatore  $T_n$  di  $\theta$  che gode di proprietà desiderabili (quali non distorsione e consistenza, per esempio)
- Fissare in  $(1 - \alpha)$  il *livello di confidenza* dell'intervallo che copre il vero (ma incognito) valore di  $\theta$
- Determinare l'intervallo (di ampiezza minima)  $[q_1, q_2]$  all'interno del quale è compresa la statistica pivot con probabilità pari al *livello di confidenza*  $(1 - \alpha)$  fissato, vale a dire

$$P[q_1 \leq Q(T_n, \theta) \leq q_2] = (1 - \alpha) \quad (5.14)$$

- invertire la relazione  $q_1 \leq Q(T_n, \theta) \leq q_2$  rispetto a  $\theta$  in modo da ricavare l'intervallo casuale che copre  $\theta$  con probabilità pari a  $(1 - \alpha)$  ossia

$$P[h(T_n, q_1, a) \leq \theta \leq h(T_n, q_2, a)] = (1 - \alpha) \quad (5.15)$$

con  $a$  vettore di costanti *note* mentre  $h(T_n, q_1, a)$  e  $h(T_n, q_2, a)$  rappresentano l'estremo inferiore e l'estremo superiore dell'intervallo casuale di livello  $(1 - \alpha)$ .

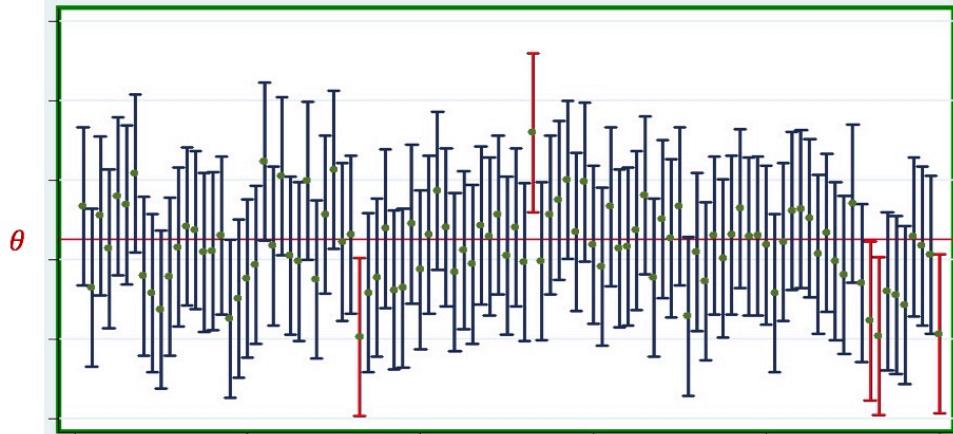
- infine, l'*intervallo di confidenza* cercato per  $\theta$  è quello restituito da (5.1), valutato in una specifica determinazione del campione casuale  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ . Potremmo anche scrivere

$$IC_\theta(1 - \alpha) : [h(t_n, q_1, a), h(t_n, q_2, a)] \quad (5.16)$$

dove, come si può notare direttamente dalla (5.14),  $q_1 = q_{1;\alpha/2}$  e  $q_2 = q_{2;(1-\alpha/2)}$  sono i quantili di ordine  $\alpha/2$  e  $(1 - \alpha/2)$  della distribuzione della statistica pivot  $Q(T_n, \theta)$ .

Vale la pena osservare che

- nella pratica, fissato il livello di confidenza, si costruisce *un solo* intervallo di confidenza tra gli  $M$  possibili che si possono ricavare sulle  $M$  determinazioni del campione casuale (considerate che  $M$  potrebbe anche essere infinito). Sicché, non essendo possibile stabilire se l'informazione fornita da tale (singolo) intervallo sia corretta o meno (e in definitiva se l'intervallo ricavato sia affidabile), si ragiona *in media* sullo spazio dei possibili campioni e si afferma che l'intervallo calcolato copre il vero valore di  $\theta$  nell' $(1 - \alpha)$  100% dei casi.



**Figura 5.1:**  $n = 100$  Intervalli di confidenza per  $\theta$  di livello  $(1 - \alpha) = 0.95$ . Il 5% di questi, evidenziati in rosso nella figura, non copre il valore di  $\theta$ .

b) il prefissato livello di confidenza  $(1 - \alpha)$  può essere raggiunto

1. per qualsiasi valore finito di  $n$

$$P[h_1(T_n, q_1, a) \leq \theta \leq h_1(T_n, q_2, a)] = (1 - \alpha) \quad (5.17)$$

2. *asintoticamente*, vale a dire per  $n \rightarrow \infty$

$$\lim_{n \rightarrow \infty} P[h_1(T_n, q_1, a) \leq \theta \leq h_1(T_n, q_2, a)] = (1 - \alpha) \quad (5.18)$$

- c) l'intervallo di confidenza cresce in ampiezza al crescere del livello di confidenza  $(1 - \alpha)$  mentre, al contrario, decresce al crescere di  $n$  (ampiezza campionaria); inoltre, esso cresce al crescere della variabilità della popolazione (ovvero della varianza del campione) essendo in questo caso necessario un intervallo più ampio per garantire la prefissata probabilità di copertura del valore di  $\theta$ .

### 5.3.1 Intervalli di confidenza esatti

Gli esempi che seguono si riferiscono alla costruzione di intervalli di confidenza *esatti*, vale a dire a partire dalla distribuzione *esatta* della statistica pivot su cui si basano, per diversi aspetti di interesse della popolazione (media, varianza, proporzione,...).

**Esempio 5.3.1** (IC per la media della popolazione ( $\sigma^2$  noto)). Consideriamo un campione casuale  $(X_1, \dots, X_n)$  da una popolazione con distribuzione  $N(\mu, \sigma^2)$  con  $\sigma^2$  noto. Vogliamo determinare l'intervallo di confidenza  $IC_\mu(1 - \alpha)$ , ovvero l'intervallo che copre  $\mu$  con probabilità pari a  $(1 - \alpha)$ . Abbiamo già detto che, noto  $\sigma$ , la media campionaria opportunamente standardizzata  $Z$  segue una distribuzione Normale standard. Quindi, al fine di individuare l'intervallo di confidenza di livello  $(1 - \alpha)$  dobbiamo identificare  $q_1, q_2$  tali che

$$P\left(q_1 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq q_2\right) = 1 - \alpha \quad (5.19)$$

Questa condizione è soddisfatta da  $q_1 = z_{\alpha/2} = -z_{1-\alpha/2}$  e  $q_2 = z_{1-\alpha/2}$  (per simmetria della distribuzione Normale standard), valori che si possono ricavare dalla tavole della distribuzione Normale standard. Possiamo ora determinare l'intervallo di confidenza prima calcolando

$$\begin{aligned} 1 - \alpha &= P \left( q_1 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq q_2 \right) \\ &= P \left( -z_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \right) \\ &= P \left( \bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \end{aligned} \quad (5.20)$$

e poi, valutato l'intervallo casuale  $\left[ \bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$  ottenuto in (5.20) nella determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale  $(X_1, X_2, \dots, X_n)$ , si ottiene l'intervallo di confidenza di livello  $(1 - \alpha)$  cercato per  $\mu$ , dato da:

$$IC_\mu(1 - \alpha) = \left[ \bar{x}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad (5.21)$$

**Esempio 5.3.2** (IC per la media della popolazione ( $\sigma^2$  non noto)). Consideriamo un campione casuale  $(X_1, \dots, X_n)$  da una popolazione avente distribuzione  $N(\mu, \sigma^2)$  con  $\sigma^2$  non noto. Vogliamo costruire un intervallo di confidenza di livello di confidenza  $(1 - \alpha)$  per  $\mu$ , vale a dire  $IC_\mu(1 - \alpha)$ . Per quanto visto nell'Esempio 5.1.1 sappiamo che  $W_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$  è una statistica pivot per  $\mu$ . Fissato il livello di confidenza  $(1 - \alpha)$ , risolviamo l'equazione in  $q_1$  e  $q_2$ ,

$$P \left( q_1 \leq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \leq q_2 \right) = 1 - \alpha \quad (5.22)$$

che è necessariamente soddisfatta da  $q_1 = t_{n-1, (1-\alpha)/2} = -t_{n-1, \alpha/2}$  e  $q_2 = t_{n-1, \alpha/2}$ , valori che sono tabulati. Di conseguenza, dalla seguente identità

$$P \left( \bar{X}_n - t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}} \right) = 1 - \alpha \quad (5.23)$$

possiamo ricavare l'intervallo casuale  $\left[ \bar{X}_n - t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}} \right]$  che, valutato nella determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale  $(X_1, X_2, \dots, X_n)$ , restituisce l'intervallo di confidenza cercato di livello  $(1 - \alpha)$  per  $\mu$

$$IC_\mu(1 - \alpha) = \left[ \bar{x}_n - t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}} \right] \quad (5.24)$$

Vale la pena osservare che, per  $n$  grande (convenzionalmente  $n > 30$ ) la v.c.  $t$  di Student converge in distribuzione a una v.c. Normale standard ossia

$$W_n = \frac{(\bar{X}_n - \mu)}{S_n/\sqrt{n}} \underset{a}{\sim} N(0, 1)$$

e quindi, sostituendo i quantili  $-t_{n-1, \alpha/2}$  e  $t_{n-1, \alpha/2}$  rispettivamente con  $-z_{1-\alpha/2}$  e  $z_{1-\alpha/2}$  in (5.132) si ottiene un intervallo di confidenza approssimato (poiché poggiato sulla distribuzione approssimata o asintotica della statistica pivot  $W_n$ ) per  $\mu$  di livello  $(1 - \alpha)$ .

**Esempio 5.3.3** (IC per la varianza  $\sigma^2$  della popolazione). Consideriamo un campione casuale  $(X_1, \dots, X_n)$  proveniente da una popolazione avente distribuzione  $N(\mu, \sigma^2)$ . Vogliamo determinare l'intervallo di confidenza  $IC_{\sigma^2}(1 - \alpha)$  di livello  $(1 - \alpha)$  per  $\sigma^2$ . Sappiamo che  $V_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$  è una statistica pivot per  $\sigma^2$ ; imponiamo dunque che

$$P\left(q_1 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq q_2\right) = 1 - \alpha \quad (5.25)$$

sicché otterremo  $q_1 = \chi_{n-1, (1-\alpha/2)}^2$  e  $q_2 = \chi_{n-1, \alpha/2}^2$ , valori entrambi tabulati. Isolato  $\sigma^2$  nella (5.25), otteniamo

$$P\left(\frac{(n-1)S_n^2}{\chi_{n-1, 1-\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi_{n-1, \alpha/2}^2}\right) = 1 - \alpha \quad (5.26)$$

sicché dalla precedente si ha l'intervallo casuale

$$\left[\frac{(n-1)S_n^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)S_n^2}{\chi_{n-1, 1-\alpha/2}^2}\right] \quad (5.27)$$

che, valutato nella determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale  $(X_1, X_2, \dots, X_n)$ , restituisce l'intervallo di confidenza cercato di livello  $(1 - \alpha)$  per  $\sigma^2$

$$IC_{\sigma^2}(1 - \alpha) = \left[\frac{(n-1)s_n^2}{\chi_{n-1, \alpha/2}^2}, \frac{(n-1)s_n^2}{\chi_{n-1, 1-\alpha/2}^2}\right] \quad (5.28)$$

**Esempio 5.3.4** (IC per il parametro  $\beta$  della distribuzione esponenziale). Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da una popolazione avente distribuzione  $Exp(\beta)$ ,  $\beta > 0$  e supponiamo di voler determinare  $IC_{\beta}(1 - \alpha)$ , intervallo di confidenza di livello  $(1 - \alpha)$  per  $\beta$ . Per farlo, possiamo utilizzare diverse statistiche pivot per  $\beta$ .

Una prima opzione è data da

$$W = \sum_{i=1}^n X_i \sim \mathcal{G}(n, \beta) \quad (5.29)$$

Tale statistica tuttavia non è pivot in quanto la sua distribuzione dipende da  $\beta$  (non è, come si dice, *parameter free*). Ma è certamente una statistica pivot per  $\beta$

$$V = \frac{1}{n} \frac{W}{\beta} \sim \mathcal{G}\left(n, \frac{1}{\beta}\right) \quad (5.30)$$

la cui distribuzione è ottenuta applicando il Teorema (di trasformazione) 1.2.2; come pure, sempre applicando il medesimo teorema, si ottiene

$$V_1 = nV = \frac{W}{\beta} \sim \mathcal{G}(n, 1) \quad \text{e} \quad V_2 = 2V_1 = 2nV = \frac{2W}{\beta} = \frac{2n\bar{X}_n}{\beta} \sim \mathcal{G}\left(\frac{2n}{2}, 2\right) = \chi_{2n}^2$$

che risultano tutte essere *statistiche pivot* per il parametro  $\beta$ . Avendo libertà di scegliere, conviene usare la statistica pivot  $V_2$  in quanto la sua distribuzione è tabulata. Imponiamo dunque

$$P\left(q_1 \leq \frac{2n\bar{X}_n}{\beta} \leq q_2\right) = 1 - \alpha \quad (5.31)$$

da cui, svolgendo gli usuali calcoli, si ottiene il seguente intervallo di confidenza di livello  $(1 - \alpha)$  per  $\beta$ :

$$IC_\beta(1 - \alpha) = \left[ \frac{2n\bar{x}_n}{\chi^2_{2n, \frac{\alpha}{2}}}, \frac{2n\bar{x}_n}{\chi^2_{2n, 1 - \frac{\alpha}{2}}} \right] \quad (5.32)$$

La statistica  $W = \sum_{i=1}^n X_i$  sintetizza efficacemente l'informazione sul parametro  $\beta$  contenuta nel campione casuale. La cosa non deve stupire: di fatto, la famiglia delle distribuzioni esponenziali di parametro  $\beta$  costituisce una *famiglia esponenziale* a  $k = 1$ -parametri e, come abbiamo visto, è sufficiente una sola statistica *naturale* quale  $W$  per realizzare tale sintesi.

Funzioni *monotone* di  $W$  conducono a *inferenze equivalenti* su  $\beta$ ; in altre parole, potevamo giungere allo stesso  $IC_\beta(1 - \alpha)$  (5.32) anche partendo da  $V_1 = \frac{W}{\beta} \sim \mathcal{G}(n, 1)$  o da  $\frac{W}{n\beta} \sim \mathcal{G}(n, 1/n)$ .

### 5.3.2 Intervalli di confidenza approssimati

La popolazione da cui proviene il campione potrebbe *non essere Normale* o la distribuzione del riassunto campionario alla base della statistica pivot potrebbe essere assai difficile (se non addirittura impossibile) da calcolare; tutto ciò può avere pesanti riflessi sulla distribuzione della statistica pivot su cui basare la costruzione dell'intervallo di confidenza  $IC_\theta(1 - \alpha)$  per il parametro di interesse  $\theta$ .

In casi siffatti possiamo lasciar perdere l'idea di procedere con la determinazione della *distribuzione esatta* della *statistica pivot* e piuttosto accontentarci di *approssimarne* la distribuzione; e nel far questo ci può venire in aiuto proprio il *teorema limite centrale*.

**Esempio 5.3.5.** Supponiamo di voler stimare la difettosità e il peso dei pezzi prodotti in un processo produttivo e supponiamo che da un campione casuale di  $n = 60$  elementi, si siano ottenute le seguenti statistiche:

$$(i) \quad \sum_{i=1}^{60} x_i = 840; \quad \sum_{i=1}^{60} x_i^2 = 12300 \quad (5.33)$$

$$(ii) \quad \sum_{i=1}^{60} y_i = 25$$

dove con  $X_i$  si indica la variabile casuale associata al peso dell' $i$ -esimo elemento del campione mentre  $Y_i \sim b(1, p)$  indica la sua difettosità; in altre parole,  $Y_i$  è una v.c. *indicatore di difettosità* sicché  $Y_i = 1$  se il pezzo è difettoso,  $Y_i = 0$  altrimenti.

Supponiamo inizialmente che  $X \sim N(\mu, \sigma^2)$  e costruiamo un intervallo di confidenza  $IC_\mu(0.99)$  sotto questa ipotesi. Poiché il parametro  $\sigma^2$  non è noto, stimiamolo con  $S_n^2$  e ricorriamo alla statistica pivot  $t$  di Student al fine di costruire un intervallo di confidenza per la media  $\mu$  della popolazione. Ripetendo i calcoli svolti nell'Esempio 5.3.2, sostituendo i valori ottenuti sulla determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale, otteniamo:

$$IC_\mu(0.99) = \left[ \bar{x}_n - t_{59, 0.005} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{59, 0.005} \frac{s_n}{\sqrt{n}} \right] \quad (5.34)$$

Dai dati a disposizione di evince facilmente che  $\bar{x}_n = 14$ ; inoltre, il valore  $t_{59,0.005} = 2.66$  è tabulato. Rimane dunque da calcolare il valore della varianza campionaria:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}_n^2 = 9.15 \quad (5.35)$$

Sostituendo i valori ottenuti nell'espressione scritta precedentemente, si ottiene  $IC_\mu(0.99) = [12.96, 15.04]$ .

Vogliamo ora ripetere il calcolo in assenza dell'ipotesi di normalità della distribuzione da cui proviene il campione; in questo caso, utilizziamo un risultato di convergenza discusso precedentemente:

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{a} N(0, 1) \quad (5.36)$$

Sfruttando nuovamente i risultati degli esempi precedenti, otteniamo quindi che l'intervallo di confidenza (approssimato) è

$$IC_\mu(0.99) = \left[ \bar{x}_n - z_{1-\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2} \frac{s_n}{\sqrt{n}} \right] = [13; 15] \quad (5.37)$$

Costruiamo ora l'intervallo di confidenza  $IC_p(0.95)$  per la proporzione  $p$  dei pezzi difettosi della popolazione; per farlo, utilizzeremo lo stimatore *frequenza relativa campionaria*  $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n Y_i$  la cui *distribuzione asintotica* si ricava facilmente usando il *Teorema Limite Centrale*

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{a} N\left(p, \frac{p(1-p)}{n}\right) \quad (5.38)$$

Inoltre, la stima di  $p$  sui dati in nostro possesso è uguale a  $\hat{p}_n = 0.42$ .

Ai fini della costruzione dell'intervallo di confidenza per  $p$ , si pone il problema di stimare  $\frac{p(1-p)}{n}$ ; ma dall'Esempio 3.1.4 già sappiamo che

$$\frac{\hat{p}_n(1-\hat{p}_n)}{n} \xrightarrow{P} \frac{p(1-p)}{n} \quad (5.39)$$

sicché, standardizzando e richiamando il Teorema 3.2.3 (di Slutsky), otteniamo la *statistica pivot* con cui procedere a costruire l'intervallo di confidenza approssimato (o asintotico) per  $p$

$$\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \xrightarrow{a} N(0, 1) \quad (5.40)$$

il cui calcolo risulta agevole e produce

$$IC_p(0.95) = \left[ \hat{p}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right] = [0.25, 0.54] \quad (5.41)$$

Ricaviamo infine l'intervallo di confidenza  $IC_{\sigma^2}(0.99)$  per la varianza  $\sigma_n^2$ , assumendo l'ipotesi di Normalità sulla popolazione generante il campione casuale. Abbiamo già calcolato il valore di  $S_n^2$ , che è uguale a  $s_n^2 = 9.15$ . Utilizzando dunque la statistica pivot  $V_n \sim \chi_{n-1}^2$ , otteniamo

$$IC_{\sigma^2}(0.99) = \left[ \frac{(n-1)s_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2}, \frac{(n-1)s_n^2}{\chi_{n-1, (1-\frac{\alpha}{2})}^2} \right] = [5.951; 15.626] \quad (5.42)$$

### 5.3.3 Intervalli di confidenza per differenze

Spesso è necessario confrontare due distribuzioni rispetto a un qualche aspetto di interesse, per esempio, la media, la varianza o la proporzione di successi e così via.

Il concetto di intervallo di confidenza può risultare ancora utile e nelle pagine che seguono tratteremo della costruzione di intervalli di confidenza per differenze di medie, di varianze o di proporzioni.

#### 5.3.3.1 Intervalli di confidenza per la differenza di medie

Vogliamo confrontare due distribuzioni e decidiamo di sintetizzare la loro differenza tramite la *differenza delle loro medie*.

Consideriamo due popolazioni aventi rispettivamente distribuzione  $N(\mu_1, \sigma_1^2)$  e  $N(\mu_2, \sigma_2^2)$  dalle quali estraiamo i due campioni casuali  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  e  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$  e assumiamo che

- a)  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ : vale a dire le due popolazioni hanno la stessa varianza (ipotesi di omoschedasticità) sicché siamo in presenza di *location model* poiché le distribuzioni differiscono solo per la *locazione* (vale a dire, per la media)
- b) i campioni  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  e  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$  sono *indipendenti*

E' immediato calcolare medie e varianze campionarie per i due campioni casuali e ricordare che sotto le ipotesi che abbiamo assunto:

$$\begin{aligned}\bar{X} &\sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right) & \text{e} & \frac{(n_1 - 1)S_X^2}{\sigma^2} \sim \chi_{n_1-1}^2 \\ \bar{Y} &\sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right) & \text{e} & \frac{(n_2 - 1)S_Y^2}{\sigma^2} \sim \chi_{n_2-1}^2\end{aligned}\tag{5.43}$$

sicché

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_1 - \mu_2, \frac{(n_1 + n_2)\sigma^2}{n_1 \cdot n_2}\right)\tag{5.44}$$

poiche, per un noto teorema, se  $(X_1, X_2, \dots, X_{n_1})$  e  $(Y_1, Y_2, \dots, Y_{n_2})$  sono *indipendenti* allora lo sono anche  $g_1(X_1, X_2, \dots, X_{n_1})$  e  $g_2(Y_1, Y_2, \dots, Y_{n_2})$  sicché

$$\mathbb{V}ar(\bar{X} - \bar{Y}) = \mathbb{V}ar(X) + \mathbb{V}ar(Y) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \frac{(n_1 + n_2)\sigma^2}{n_1 \cdot n_2}$$

e la (5.44) segue dalla *proprietà di riproducibilità* (Teorema 1.3.6). La dimostrazione della (5.44) è lasciata per esercizio.

La quantità  $\mathbb{V}ar(\bar{X} - \bar{Y}) = \frac{(n_1 + n_2)\sigma^2}{n_1 \cdot n_2}$  deve ora essere stimata e per far ciò abbiamo bisogno di uno stimatore della varianza *comune* alle due popolazioni (stando alle ipotesi assunte)  $\sigma^2$ . Quest'ultima può essere stimata ricorrendo alla media ponderata di  $S_X^2$  e  $S_Y^2$  data da

$$\begin{aligned}S_P^2 &= \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} S_X^2 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} S_Y^2 \\ &= \frac{n_1 - 1}{n - 2} S_X^2 + \frac{n_2 - 1}{n - 2} S_Y^2\end{aligned}\tag{5.45}$$

dove  $n_1 + n_2 = n$ . Quest'ultima quantità prende il nome di *pooled variance* e vale

$$\begin{aligned}\mathbb{E}(S_P^2) &= \mathbb{E}\left[\frac{n_1 - 1}{n - 2} S_X^2\right] + \mathbb{E}\left[\frac{n_2 - 1}{n - 2} S_Y^2\right] \\ &= \frac{n_1 - 1}{n - 2} \sigma^2 + \frac{n_2 - 1}{n - 2} \sigma^2 = \sigma^2, \quad \forall \sigma^2 > 0\end{aligned}\tag{5.46}$$

sicché  $S_P^2$  è uno stimatore *non distorto* di  $\sigma^2$ . Allora possiamo stimare

$$\mathbb{V}ar(\bar{X} - \bar{Y}) = \frac{(n_1 + n_2)}{n_1 \cdot n_2} \sigma^2$$

con il suo stimatore *plug-in* dato da

$$\frac{(n_1 + n_2)}{n_1 \cdot n_2} S_P^2$$

Si può immediatamente osservare che  $S_P^2$  è anche uno stimatore (quadraticamente) *consistente* di  $\sigma^2$ ; infatti

$$\begin{aligned}\mathbb{V}ar(S_P^2) &= \mathbb{V}ar\left[\frac{n_1 - 1}{n - 2} S_X^2 + \frac{n_2 - 1}{n - 2} S_Y^2\right] \\ &= \frac{(n_1 - 1)^2}{(n - 2)^2} \mathbb{V}ar(S_X^2) + \frac{(n_2 - 1)^2}{(n - 2)^2} \mathbb{V}ar(S_Y^2) \\ &= \frac{(n_1 - 1)^2}{(n - 2)^2} \frac{2\sigma^4}{n_1 - 1} + \frac{(n_2 - 1)^2}{(n - 2)^2} \frac{2\sigma^4}{n_2 - 1} \\ &= \frac{2\sigma^4}{n - 2} \xrightarrow[n \rightarrow \infty]{} 0\end{aligned}\tag{5.47}$$

Di conseguenza, la *statistica pivot* per l'inferenza su  $\Delta = (\mu_1 - \mu_2)$  sulla quale basare la costruzione dell'intervallo di confidenza per la differenza delle medie  $\Delta$  è data da

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}} \sim t_{n-2} \text{ con } n = n_1 + n_2\tag{5.48}$$

e risulta dal rapporto tra due v.c. indipendenti ovvero una v.c. Normale standard

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}}$$

e la radice quadrata di una v.c. chi-quadrato divisa per i suoi gradi di libertà

$$\left[ \frac{(n - 2)S_P^2}{\sigma^2} \cdot \frac{1}{n - 2} \right]^{\frac{1}{2}}$$

Infine, l'intervallo di confidenza esatto di livello  $(1 - \alpha)$  per  $\Delta = (\mu_1 - \mu_2)$  è

$$IC_{\Delta}(1 - \alpha) : \left[ (\bar{x} - \bar{y}) - t_{n-2, 1-\frac{\alpha}{2}} \frac{S_P}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}}, (\bar{x} - \bar{y}) + t_{n-2, 1-\frac{\alpha}{2}} \frac{S_P}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}} \right]\tag{5.49}$$

Vale la pena osservare che

a) se  $n_1 > 30$  e  $n_2 > 30$  allora possiamo ottenere un'approssimazione del  $IC_{\Delta}(1-\alpha)$  sfruttando il Teorema di Slutsky e di conseguenza sostituendo nella (5.49) i quantili della distribuzione (esatta)  $t$  di Student con i quantili della distribuzione (asintotica) Normale standard., ottenendo

$$IC_{\Delta}(1-\alpha) : \left[ (\bar{x} - \bar{y}) - z_{1-\frac{\alpha}{2}} \frac{s_p}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}}, (\bar{x} - \bar{y}) + z_{1-\frac{\alpha}{2}} \frac{s_p}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}} \right] \quad (5.50)$$

Di questo modo di procedere si può avere formale giustificazione: posto  $V = \bar{X} - \bar{Y}$  e  $\mu = \mu_1 - \mu_2$  e ricordando che  $S_p^2 \xrightarrow{P} \sigma^2$  (ovvero ne è stimatore consistente), possiamo scrivere

$$\begin{aligned} T &= \frac{V - \mu}{S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}} = \frac{V}{S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}} - \frac{\mu}{S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}} \\ &= \frac{1}{S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}} V - \frac{1}{S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}} \mu \\ &= B_n V - A_n \end{aligned} \quad (5.51)$$

ma  $B_n \xrightarrow{P} \frac{1}{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}} \sigma}$ ,  $(V - \mu) \xrightarrow{D} N \left( 0, \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \sigma^2 \right)$  e  $A_n \xrightarrow{P} \frac{\mu}{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}} \sigma}$  sicché per il Teorema di Slutsky

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}} \xrightarrow{a} N(0, 1) \quad (5.52)$$

il che giustifica quanto scritto in (5.50).

b) se le distribuzioni da cui provengono i campioni  $(X_1, \dots, X_{n_1})$  e  $(Y_1, \dots, Y_{n_2})$  non sono Normali (pur sempre sotto ipotesi di omoschedasticità) e  $n_1 > 30$  e  $n_2 > 30$ , possiamo procedere a ricavare l'intervallo di confidenza approssimato per  $\Delta = (\mu_1 - \mu_2)$  come fatto al precedente punto a).

### 5.3.3.2 Intervallo di confidenza per la differenza di proporzioni

Consideriamo due campioni casuali  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  e  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$  indipendenti e provenienti da due popolazioni bernoulliane per cui  $X_i \sim b(1, p_1)$  mentre  $Y_i \sim b(1, p_2)$ .

Vogliamo confrontare le due popolazioni bernoulliane e decidiamo di sintetizzare la loro differenza tramite la differenza delle loro proporzioni. Per far ciò costruiamo l'intervallo di confidenza approssimato di livello  $(1 - \alpha)$   $IC_{p_1-p_2}(1 - \alpha)$ .

Asumiamo che  $n_1 > 30$  e che  $n_2 > 30$ . Ora, partendo dai due estimatori  $\hat{p}_1$  e  $\hat{p}_2$  rispettivamente di  $p_1$  e di  $p_2$  e che sappiamo già essere non distorti e consistenti per le quantità che stimano, invocando il Teorema Limite Centrale, possiamo scrivere

$$\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \underset{a}{\sim} N \left( p_1, \frac{p_1(1-p_1)}{n_1} \right) \quad \hat{p}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i \underset{a}{\sim} N \left( p_2, \frac{p_2(1-p_2)}{n_2} \right) \quad (5.53)$$

e pertanto possiamo subito definire la seguente *statistica pivot* per l'inferenza sulla *differenza*  $\Delta = (p_1 - p_2)$  delle due proporzioni

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \stackrel{a}{\sim} N(0, 1) \quad (5.54)$$

Ripetendo un procedimento del tutto simile a quanto visto nella costruzione dell'intervallo di confidenza per  $\Delta = (\mu_1 - \mu_2)$ , si ricava agevolmente l'intervallo di confidenza per  $\Delta = (p_1 - p_2)$  che è dato da

$$IC_{\Delta}(1 - \alpha) : \left[ (\hat{p}_1 - \hat{p}_2) - z_{1-\frac{\alpha}{2}} SE, (\hat{p}_1 - \hat{p}_2) + z_{1-\frac{\alpha}{2}} SE \right] \quad (5.55)$$

dove

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad (5.56)$$

è lo *standard error* (ovvero, la deviazione standard) di  $(\hat{p}_1 - \hat{p}_2)$ .

### 5.3.3.3 Intervallo di confidenza per il rapporto di varianze

Consideriamo due popolazioni aventi rispettivamente distribuzione  $N(\mu_1, \sigma_1^2)$  e  $N(\mu_2, \sigma_2^2)$  dalle quali estraiamo i due campioni casuali  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  e  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$ , *indipendenti*.

E' immediato calcolare medie campionarie ( $\bar{X}$  e  $\bar{Y}$ ) e varianze campionarie ( $S_X^2$  e  $S_Y^2$ ) per i due campioni casuali e ricordare che sotto le ipotesi che abbiamo assunto,

$$\begin{aligned} \bar{X} &\sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) & \text{e} & \quad W_X = \frac{(n_1 - 1)S_X^2}{\sigma_1^2} \sim \chi_{n_1-1}^2 \\ \bar{Y} &\sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) & \text{e} & \quad W_Y = \frac{(n_2 - 1)S_Y^2}{\sigma_2^2} \sim \chi_{n_2-1}^2 \end{aligned} \quad (5.57)$$

Vogliamo ora costruire l'intervallo di confidenza  $IC_{\sigma_1^2/\sigma_2^2}(1 - \alpha)$  per il *rapporto delle varianze* delle due popolazioni. Per fare ciò, definiamo la seguente *statistica pivot* per l'inferenza su  $\sigma_1^2/\sigma_2^2$ :

$$W = \frac{W_X/(n_1 - 1)}{W_Y/(n_2 - 1)} = \frac{\sigma_2^2}{\sigma_1^2} \frac{S_X^2}{S_Y^2} \sim F_{n_1-1, n_2-1} \quad (5.58)$$

che sappiamo seguire la distribuzione di Fisher-Snedecor con  $(n_1 - 1)$  e  $(n_2 - 1)$  gradi di libertà in quanto *rapporto* di due variabili casuali *chi-quadrato indipendenti* divise per i rispettivi gradi di libertà (vedi Teorema 5.2.5). Pertanto, fissato il livello di confidenza in  $(1 - \alpha)$  e isolato il rapporto  $\frac{\sigma_1^2}{\sigma_2^2}$  dalla statistica pivot, si ha

$$P\left(q_1 \leq \frac{\sigma_2^2 S_X^2}{\sigma_1^2 S_Y^2} \leq q_2\right) = P\left(\frac{S_X^2/S_Y^2}{q_{n_1-1, n_2-2; \frac{\alpha}{2}}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_X^2/S_Y^2}{q_{n_1-1, n_2-2; 1-\frac{\alpha}{2}}}\right) = 1 - \alpha \quad (5.59)$$

dove  $q_2 = q_{n_1-1, n_2-1; \alpha/2}$  e

$$q_1 = q_{n_1-1, n_2-2; 1-\frac{\alpha}{2}} = \frac{1}{q_{n_2-1, n_1-1; \frac{\alpha}{2}}}$$

sono, rispettivamente, i quantili di ordine  $(1 - \alpha/2)$  e  $\alpha/2$  della distribuzione  $F$  di Fisher-Snedecor con  $(n_1 - 1) \geq (n_2 - 1)$  gradi di libertà,  $F_{(n_1-1), (n_2-1)}$ . Pertanto,

$$IC_{\sigma_1^2/\sigma_2^2}(1 - \alpha) = \left[ \frac{s_X^2/s_Y^2}{q_{n_1-1, n_2-1; \frac{\alpha}{2}}}, \frac{s_X^2/s_Y^2}{1/q_{n_2-1, n_1-1; \frac{\alpha}{2}}} \right] \quad (5.60)$$

**Esempio 5.3.6.** Supponiamo di avere le determinazioni di due campioni casuali, rispettivamente di ampiezza  $n_1 = 21$  e  $n_2 = 16$ , provenienti da due distribuzioni Normali  $N(\mu_1, \sigma_1^2)$  e  $N(\mu_2, \sigma_2^2)$ . Ora, sulla base dei dati a nostra disposizione abbiamo che  $s_X^2 = 1600$  e  $s_Y^2 = 1225$ . Fissato  $\alpha = 0.05$ , l'intervallo di confidenza per il *rapporto delle varianze* è dato da:

$$IC_{\sigma_1^2/\sigma_2^2}(0.95) = \left[ \frac{s_X^2/s_Y^2}{q_{n_1-1, n_2-1; \frac{0.05}{2}}}, \frac{s_X^2/s_Y^2}{1/q_{n_2-1, n_1-1; \frac{0.05}{2}}} \right] \quad (5.61)$$

dove  $q_{21-1, 16-1; 0.05/2} = q_{20, 15; 0.025} = 2.76$  e

$$q_{21-1, 16-1; 1-0.05/2} = q_{20, 15; 0.975} = \frac{1}{q_{15, 20; 0.025}} = \frac{1}{2.57} = 0.389$$

sicché

$$IC_{\sigma_1^2/\sigma_2^2}(0.95) = \left[ \frac{1600/1225}{2.76}, \frac{1600/1225}{0.389} \right] = [0.473, 3.36] \quad (5.62)$$

**Esempio 5.3.7.** Un gruppo di ricercatori vuole studiare le *abitudini predatorie* di due specie di ragni accomunati dall'essere "lanciatori" di ragnatela, *Deinopis* e *Menneus*, specie che coesistono nell'Australia dell'Est, sulla base della *lunghezza* (in mm) delle prede abitualmente intrappolate nelle loro ragnatele.

Lo studio viene condotto sulla base di due campioni casuali,  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  e  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$ , uno per ciascuna specie di ragno, con l'obiettivo di rispondere alle seguenti domande:

- i) c'è evidenza di una qualche (significativa) *differenza nella lunghezza media* delle prede delle due popolazioni di ragni?
- ii) c'è evidenza di una qualche (significativa) *differenza nella varianza delle lunghezze delle prede* delle due popolazioni di ragni? Quesito, quest'ultimo, piuttosto interessante: in presenza di una *piccola differenza* tra le varianze, potremmo supporre che le prede dei due gruppi di ragni siano abbastanza *omogenee* mentre una *grande differenza* tra le varianze porterebbe a supporre una forte *disomogeneità* tra le lunghezze delle prede dei due gruppi di ragni, aprendo alla possibilità a ipotizzare una qualche *specializzazione predatoria* da parte di *Deinopis* e *Menneus*.

Come è lecito attendersi, la soluzione al quesito i) consiste nel calcolo dell'intervallo di confidenza di livello  $(1 - \alpha)$  per la differenza delle medie delle due popolazioni di ragni ovvero  $IC_{\mu_1 - \mu_2}(1 - \alpha)$  mentre la soluzione al quesito ii) poggerà sul calcolo dell'intervallo di confidenza per il rapporto delle varianze della lunghezza delle prede delle due specie di ragni  $IC_{\sigma_1^2/\sigma_2^2}(1 - \alpha)$  (e sulla sua interpretazione)

In virtù del fatto che i ragni componenti il campione siano stati scelti casualmente, possiamo assumere che le misurazioni delle lunghezze delle prede che compongono i due campioni siano *indipendenti* (sia all'interno dei due gruppi che *tra* i due gruppi). Assumiamo, inoltre, che le lunghezze delle prede delle due specie di ragni siano *normalmente distribuite*,  $X \sim N(\mu_1, \sigma_1^2)$  e  $Y \sim N(\mu_2, \sigma_2^2)$ .

i) **Caso A):**  $\sigma_1^2 = \sigma_2^2$

In questo caso, in base a quanto già visto, si ha

$$IC_{\mu_1-\mu_2}(1-\alpha) : \left[ (\bar{x} - \bar{y}) - t_{n-2,1-\frac{\alpha}{2}} \frac{s_P}{\sqrt{\frac{n_1 n_2}{n_1+n_2}}}, (\bar{x} - \bar{y}) + t_{n-2,1-\frac{\alpha}{2}} \frac{s_P}{\sqrt{\frac{n_1 n_2}{n_1+n_2}}} \right]$$

Sulla base delle osservazioni disponibili,  $n_1 = 20$ ,  $\bar{x} = 10.320$ ,  $s_X^2 = 6.239$  mentre  $n_2 = 16$ ,  $\bar{y} = 9.025$ ,  $s_Y^2 = 3.515$ ; usando le informazioni disponibili, inoltre, si ottiene la stima della *pooled variance*

$$s_P^2 = \frac{(20-1)6.239 + (16-1)3.515}{(20+16)-2} = 5.037$$

Ora, fissato  $\alpha = 0.05$  per cui  $t_{(20+16)-2; 0.05/2} = t_{34; 0.025} = 2.032$  si ha

$$IC_{\mu_1-\mu_2}(0.95) : [-0.235, 2.825]$$

Poiché l'intervallo di confidenza appena calcolato contiene lo zero, non possiamo concludere che la *differenza tra le lunghezze medie* delle prede sia, alla luce dell'informazione disponibile e al livello di confidenza fissato, significativamente diversa da zero. In altre parole, le due specie di ragno, *Deinopis* e *Menneus*, non manifestano differenza nella dimensione media delle loro prede.

ii) **Caso B):**  $\sigma_1^2 \neq \sigma_2^2$

In questo caso la *statistica pivot* per la costruzione dell'intervallo di confidenza per la differenza delle medie delle due popolazioni è data da

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} \quad (5.63)$$

e segue una *distribuzione di Welch*, distribuzione piuttosto complicata; lo stesso Welch (1938), in un articolo pubblicato su *Biometrika*, propose una sua *approssimazione* della distribuzione di  $T$  dimostrando che che

$$T \underset{a}{\sim} t_\nu \quad (5.64)$$

con

$$\nu = \frac{\left( \frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left( \frac{s_X^2}{n_1} \right)^2 + \frac{1}{n_2-1} \left( \frac{s_Y^2}{n_2} \right)^2} \quad (5.65)$$

Le dimensioni dei due campioni nell'esempio che stiamo trattando,  $n_1 = 20$  e  $n_2 = 16$ , precludono il ricorso all'approssimazione (5.64) e usare direttamente la distribuzione di Welch esula dal livello di questo corso. Quindi, qui ci fermiamo.

Il confronto tra le medie di due popolazioni con *varianze differenti* è anche noto come *problema di Behrens-Fisher*.

ii) Provvediamo ora a ricavare l'intervallo di confidenza per il rapporto delle varianze delle due popolazioni. In questo caso, la statistica pivot è data da

$$W = \frac{\frac{(n_1-1)S_X^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_Y^2}{\sigma_2^2}/(n_2-1)} = \frac{\sigma_2^2}{\sigma_1^2} \frac{S_X^2}{S_Y^2} \sim F_{n_1-1, n_2-1} \quad (5.66)$$

e, fissato  $\alpha = 0.05$  si ha

$$q_{n_1-1, n_2-1; \alpha/2} = q_{19, 15; 0.025} = 2.77$$

e

$$q_{n_1-1, n_2-1; (1-\alpha/2)} = q_{19, 15; 0.975} = \frac{1}{q_{15, 19; 0.025}} = \frac{1}{2.62} = 0.38$$

quindi

$$\begin{aligned} IC_{\sigma_1^2/\sigma_2^2}(0.95) &= \left[ \frac{s_X^2/s_Y^2}{q_{n_1-1, n_2-1; \frac{\alpha}{2}}}, \frac{s_X^2/s_Y^2}{q_{n_2-1, n_1-1; \frac{\alpha}{2}}} \right] \\ &= \left[ \frac{6.239/3.515}{2.77}, \frac{6.239/3.515}{0.38} \right] \\ &= [0.640, 4.644] \end{aligned} \quad (5.67)$$

Poiché l'intervallo contiene il valore 1, sulla base dell'informazione contenuta nei dati a disposizione e del livello di confidenza fissato, *non vi è evidenza che le varianze delle lunghezze delle prede delle due specie di ragni siano diverse*, il che porterebbe a *escludere* la presenza di una qualche *specializzazione predatoria* delle due specie di ragno.



Unità campionarie	Deinopis		Menneus	
	$n_1=20$	$s_x^2$	$n_2=16$	$s_y^2$
1	12,9	166,4	10,2	104,0
2	10,2	104,0	6,9	47,6
3	7,4	54,8	10,9	118,8
4	7,0	49,0	11,0	121,0
5	10,5	110,3	10,1	102,0
6	11,9	141,6	5,3	28,1
7	7,1	50,4	7,5	56,3
8	9,9	98,0	10,3	106,1
9	14,4	207,4	9,2	84,6
10	12,3	151,3	8,8	77,4
11	14,3	204,5	10,2	104,0
12	7,8	60,8	6,7	44,9
13	7,1	50,4	6,2	38,4
14	13,3	176,9	10,3	106,1
15	9,7	94,1	10,7	114,5
16	12,3	151,3	10,1	102,0
17	9,8	96,0		
18		7,1	50,4	
19		11,7	136,9	
20		9,7	94,1	
	206,40	2248,58	144,40	1355,94
Media_camp	10,320		9,025	
Var_camp	6,239		3,515	
Pooled_Var	5,037			
$t_{(20+16-2); 0.025}$	2,032			
$q_{(18, 15); 0.975}$	2,77			
$q_{(18, 15); 0.025}$	0,38			

**Figura 5.2:** Osservazioni campionarie e misure di sintesi per le lunghezze delle prede delle due specie di ragni (Deinopis e Menneus)

## 5.4 Test di ipotesi

Stima puntuale e stima intervallare sono procedure inferenziali comuni e utili. Un altro tipo di inferenza spesso utilizzata riguarda i test per la verifica di ipotesi. La scienza moderna (e, analogamente, il ragionamento scientifico) procede per *ipotesi riconducibili a teorie o a esperimenti pregressi*. Ma

- a) quanto un'ipotesi è *realistica e compatibile* con l'informazione di cui disponiamo (o con la conoscenza che abbiamo) relativamente all'aspetto di interesse del fenomeno a cui essa si riferisce?  
[*credibilità di un'ipotesi*]
- b) esiste un *ragionamento oggettivo*, basato su ragioni matematiche, che permetta di *inferire dall'informazione contenuta nel campione* qualcosa circa la *verità postulata dall'ipotesi in questione*?  
[*verità di un'ipotesi alla luce dell'informazione campionaria*]
- c) riusciamo in qualche modo a *misurare* questa verità postulata dall'ipotesi?  
[*misura della credibilità di un'ipotesi*]

La teoria dei test per la verifica di ipotesi cerca di trovare delle risposte a queste domande.

**Definizione 5.4.1** (Ipotesi statistica). Si definisce *ipotesi statistica* una qualunque *congettura* relativa a un aspetto di interesse che può essere

- a) un parametro o una funzione del parametro nell'ambito di una distribuzione di cui si conosce la forma funzionale (*ipotesi parametrica*)
- b) un aspetto di una distribuzione di cui non conosciamo la sua forma funzionale se non per elementi del tutto generali (*ipotesi non parametrica*).

In questo capitolo, ci concentreremo in particolare su *ipotesi parametriche* (da intendesi, ipotesi relative a un parametro  $\theta \in \Theta \subseteq \mathbb{R}^k$ ,  $k \in \mathbb{N}$  o funzioni  $g(\theta)$  di esso), e quindi su quello che è noto come *approccio parametrico* alla teoria dei test statistici.

Le ipotesi parametriche possono a loro volta essere classificate in *semplici* o *composite (unilaterali o bilaterali)*

1. un'ipotesi *semplice* è un'ipotesi che fissa il parametro a un valore ben definito, identificando completamente la distribuzione;
2. un'ipotesi *composita* è un insieme di ipotesi semplici sul valore del parametro e pertanto non specifica completamente la distribuzione; un'ipotesi composta può essere *unilaterale* o *bilaterale*

**Esempio 5.4.1.** Supponiamo di avere formulato con un'ipotesi relativa alla media di una distribuzione Normale,  $N(\mu, \sigma^2)$ . Nel caso in cui la varianza  $\sigma^2 = \sigma_0^2$  sia nota, l'ipotesi  $H : \mu = \mu_0$  con  $\mu_0 \in \mathbb{R}$  valore fissato, è un'ipotesi *semplice* perché identifica univocamente la distribuzione che sotto ipotesi  $H$ , per l'appunto, è  $N(\mu_0, \sigma_0^2)$ . Diversamente, se la varianza non fosse nota, la medesima ipotesi  $H : \mu = \mu_0$  con  $\mu_0 \in \mathbb{R}$  finirebbe per non essere *semplice* poiché esistono infinite distribuzioni

Normali, una per ogni valore di  $\sigma^2 > 0$ ,  $N(\mu_0, \sigma^2)$ , compatibili con l'ipotesi  $H$  che non è bastante per specificare univocamente la distribuzione.

Sempre rimanendo nello stesso contesto, indipendentemente dal fatto che la varianza  $\sigma^2$  sia o meno nota, le ipotesi  $H : \mu > \mu_0$  al pari di  $H : \mu \geq \mu_0$  o  $H : \mu < \mu_0$ ,  $H : \mu \leq \mu_0$  o ancora  $H : \mu \neq \mu_0$  sono da ritenersi tutte *composite*. Nessuna, infatti, è in grado di *specificare completamente* la distribuzione.

Lo stesso discorso può essere fatto scambiando i ruoli di  $\mu$  e  $\sigma^2$  nell'ambito di ipotesi che vertono sulla varianza  $\sigma^2$ . Analogi discorsi possono essere fatti per una qualsiasi altra distribuzione.

### 5.4.1 Regola di decisione e potenza del test

Supponiamo di essere interessati a formulare e verificare ipotesi relative al parametro  $\theta$  che indica la distribuzione  $F_X(x; \theta), \theta \in \Theta$  di una v.c.  $X$  sulla base dell'informazione contenuta in un campione casuale  $(X_1, X_2, \dots, X_n)$ . Gli ingredienti di una siffatta procedura inferenziale sono i seguenti:

- a) il *modello statistico*:  $(\mathfrak{X}, \mathcal{F}_\theta)$  con  $\mathfrak{X}$  spazio campionario e  $\mathcal{F}_\theta = \{F_X(x; \theta), \theta \in \Theta\}$
- b) il *sistema di ipotesi* (parametriche)

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ vs. \\ H_1 : \theta \in \Theta_1 \end{cases} \quad \text{con } \Theta = \Theta_0 \cup \Theta_1 \text{ e } \Theta_0 \cap \Theta_1 = \emptyset \quad (5.68)$$

dove  $H_0$  prende il nome di *ipotesi nulla* ovvero ipotesi che rappresenta lo *status quo* e che si intende confutare sulla base dell'informazione campionaria mentre  $H_1$  prende il nome di *ipotesi alternativa* e rappresenta l'ipotesi di ricerca relativamente alla cui verità si vorrebbe avere il sostegno dei dati.

- c) la *regola di decisione*: sulla base dell'informazione contenuta nel campione casuale proveniente da  $F_X(x; \theta)$ , *decidere* (=costruire una regola di decisione) a favore della conservazione di  $H_0$  o del suo rifiuto a favore di  $H_1$ .

Nello stabilire la regola di decisione bisogna comunque tener conto del fatto che operiamo in un contesto di *informazione incompleta* poiché disponiamo solo di quella contenuta nel campione casuale sicché decideremo in merito a  $H_0$  o a  $H_1$  in termini probabilistici ossia, ogni possibile decisione non sarà priva di errore. riassumiamo nella tabella che segue i possibili casi

**Tabella 5.1:** Errori di I° e II° tipo

Decisione	se $H_0$ è vera	se $H_0$ è falsa
Rifiuto $H_0$	<b>Errore del I° tipo</b>	No Errore
Non rifiuto $H_0$	No Errore	<b>Errore del II° tipo</b>

Ora

- si commette un errore del I° tipo quando si *rifiuta*  $H_0$  quando essa è vera e a esso è legata la seguente probabilità  $\alpha$

$$\alpha = P_\theta(\text{Rifiutare } H_0 \mid H_0 \text{ vera})$$

spesso chiamata *livello di significatività del test* o, più semplicemente, *livello del test*

- si commette un errore del II° tipo quando si *non si rifiuta*  $H_0$  quando essa è *falsa* e a esso è legata la seguente probabilità  $\beta$

$$\beta = P_\theta(\text{Non rifiutare } H_0 \mid H_0 \text{ falsa})$$

Come vedremo, l'ipotesi nulla  $H_0$  (e la distribuzione ipotizzata sotto di essa) gioca un ruolo fondamentale nel determinare la *regola di decisione* mentre l'ipotesi alternativa  $H_1$  risulterà di fondamentale importanza per *misurare la bontà del test* (ovvero della *regola di decisione*).

La figura 5.3 mostra chiaramente che i due tipi di errori non possono essere *contemporaneamente* ridotti: una riduzione dell'errore di I° tipo comporta un'aumento della probabilità di errore di II° tipo e viceversa. Tradizionalmente, si preferisce avere sotto controllo il valore di  $\alpha$ , probabilità di errore di I° tipo, essendo questo tipo di errore decisamente importante perché relativo all'*erroneo rifiuto* di  $H_0$  (conoscenza consolidata e sicura) a favore di  $H_1$  (ipotesi sperimentale o di ricerca); è del tutto sensato volerne controllare la probabilità di commeterlo e, pertanto, fissare in  $\alpha$  il suo (massimo) valore. Fissato  $\alpha$ , sceglieremo la regola di decisione che *minimizza*  $\beta$ . Per avere ragione di quest'ultima scelta strategica, introduciamo il concetto di *potenza del test* di livello  $\alpha$ .

**Definizione 5.4.2** (Potenza del test). Dato un test di livello  $\alpha$  cui corrisponde la regione di rifiuto  $C_\alpha$  diciamo *potenza del test* la quantità

$$\eta_{C_\alpha} = 1 - \beta = 1 - P_\theta(\text{Non riifiutare } H_0 \mid H_0 \text{ falsa}) \quad (5.69)$$

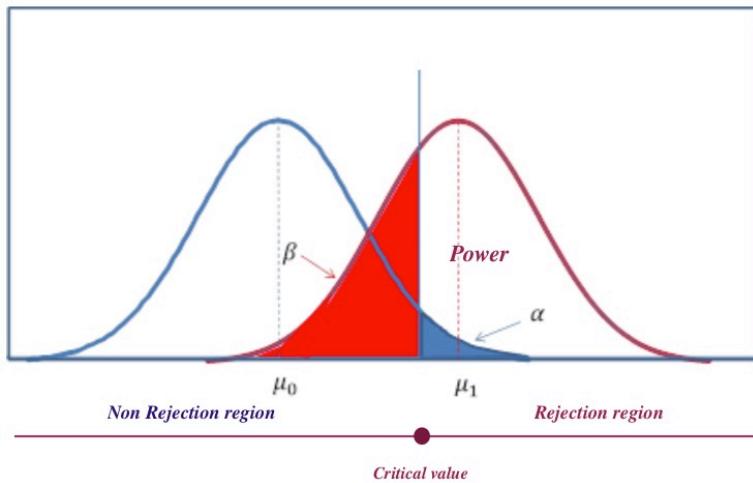
L'informazione fornita dalla funzione di potenza è relativa alla probabilità di rifiutare  $H_0$  a favore di  $H_1$  quando  $H_0$  è *falsa* o, in altri termini, non risulta verosimile alla luce dell'informazione su  $\theta$  contenuta nei dati a disposizione. Pertanto, *minimizzare* la probabilità dell'errore di II° tipo è equivalente a *massimizzare* la potenza del test ovvero la probabilità di prendere la decisione corretta riguardo  $H_0$ .

Indicato con  $\mathfrak{X}$  lo *spazio campionario*, un test di ipotesi consiste nell'individuare un suo sottospazio, che definiremo *regione critica*  $C_\alpha$ , compatibile con la probabilità  $\alpha$  di commettere un errore di I° tipo. Una volta determinato tale sottospazio, il test consisterà nell'applicare la seguente *regola di decisione*:

$$\begin{cases} \text{Rifiutare } H_0 & \text{se } (X_1, X_2, \dots, X_n) \in C_\alpha \\ \text{Non rifiutare } H_0 & \text{se } (X_1, X_2, \dots, X_n) \in \bar{C}_\alpha \end{cases} \quad (5.70)$$

Ma allora, operativamente, *come costruire una regola di decisione* riguardo a  $H_0$  e  $H_1$ ?

Per far ciò ci possono venire in aiuto ancora una volta opportune *statistiche pivot* tramite le quali stabilire i limiti (inferiore e/o superiore) della *regione critica*  $C_\alpha$  come mostra esempio che segue.



**Figura 5.3:** Probabilità di un Errore di I° ( $\alpha$ ) e di II° ( $\beta$ ) tipo e potenza del test (Power)

#### 5.4.1.1 Test per la media (campionamento da $N(\mu, \sigma^2)$ con $\sigma^2$ noto)

Sia  $X \sim N(\mu, \sigma^2)$  con  $\sigma^2$  parametro noto, diciamo  $\sigma^2 = \sigma_0^2$  e sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione di  $X$ . Si vuole verificare, sulla base dell'informazione circa  $\mu$  contenuta nella determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale, il seguente sistema di ipotesi:

$$\begin{cases} H_0 : \mu = \mu_0 \\ vs. \\ H_1 : \mu = \mu_1, \quad \mu_1 > \mu_0 \end{cases} \quad (5.71)$$

In questo caso  $\Theta_0 = \{\mu_0\}$  e  $\Theta_1 = \{\mu_1\}$  sicché  $\Theta = \{\mu_0, \mu_1\}$  e, come possiamo osservare, entrambe le ipotesi sono *semplici* in quanto ciascuna di esse *specifica completamente* la distribuzione da cui proviene il campione. Ora,

- a)  $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$  [sintesi dell'informazione in  $(X_1, X_2, \dots, X_n)$  circa  $\mu$ ]
- b) dalla precedente si ottiene facilmente la seguente *statistica test* che, essendo  $\sigma^2$  noto è, a tutti gli effetti una *statistica pivot*

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad (5.72)$$

- c) *regola di decisione*: è ragionevole rifiutare  $H_0$  quando  $\bar{X}_n > \bar{x}_c$  con  $\bar{x}_c$  valore critico ossia valore oltre il quale non sembra più essere ragionevole  $H_0$ ; ossia,

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } \bar{X}_n > \bar{x}_c \quad (5.73)$$

E' ragionevole pensare che  $\bar{x}_c$  dipenda sull'informazione contenuta nella realizzazione campionaria  $(x_1, x_2, \dots, x_n)$  e sulle assunzioni poste alla base del sistema di ipotesi da verificare.

Quello che, in definitiva, cerchiamo è una regola di decisione che sia compatibile con il requisito fissato sulla probabilità  $\alpha$  di commettere un errore del I° tipo. Detto questo, il valore di  $\bar{x}_c$  sarà soluzione della seguente equazione

$$P(\bar{X}_n > \bar{x}_c | H_0) = \alpha \quad (5.74)$$

Una volta ottenuto il valore di  $\bar{x}_c$  la *regola di decisione* (5.73) è determinata e operativa (vale a dire, possiamo usarla per prendere una decisione in merito a  $H_0$ , e conseguentemente a  $H_1$ ). Proviamo ora a formalizzare il ragionamento sviluppato nei punti precedenti.

Ricordando a), sotto  $H_0 : \mu = \mu_0$  si ha  $\bar{X}_n \underset{H_0}{\sim} N\left(\mu_0, \frac{\sigma^2}{n}\right)$  sicchè, fissato  $\alpha$ , l'equazione diventa

$$\alpha = P(\bar{X}_n > \bar{x}_c | H_0) = P\left(\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x}_c - \mu_0}{\sigma/\sqrt{n}}\right) = P(Z > z_c) \quad (5.75)$$

e dunque la soluzione dell'equazione  $P(Z > z_c) = \alpha$  è  $z_c = z_{1-\alpha}$ ; pertanto

$$\frac{\bar{x}_c - \mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha} \quad (5.76)$$

e infine con pochi e semplici passaggi algebrici, si ottiene il *valore critico del test per la media* di livello  $\alpha$

$$\bar{x}_c = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad (5.77)$$

che a sua volta costituisce è il limite inferiore della regione critica  $C_\alpha$  (da notare che su  $\bar{x}_c$  valgono le considerazioni fatte al punto c). Allora, la regione critica di livello  $\alpha$  sarà data da

$$C_\alpha = \left\{ (x_1, x_2, \dots, x_n) \in \mathfrak{X} \mid \bar{x}_n > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\} \quad (5.78)$$

e dunque, in accordo con la *regola di decisione* stabilita (5.73), possiamo riscrivere quest'ultima nella maniera seguente

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } z > z_{1-\alpha} \quad (5.79)$$

ovvero, ricordando che  $z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$ ,

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } \bar{x}_n > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad (5.80)$$

Calcoliamo infine la *potenza del test*. Ricordando che  $\beta = P_\theta(\text{Non rifiutare } H_0 | H_0 \text{ falsa})$ , per l'esempio in questione possiamo riscrivere  $\beta$  come

$$P(\bar{X}_n \leq \bar{x}_c | \mu = \mu_1)$$

e per la definizione 5.4.2 di potenza del test avere

$$\begin{aligned} \eta_{C_\alpha} &= 1 - \beta = 1 - P(\bar{X}_n \leq \bar{x}_c | \mu = \mu_1) \\ &= P(\bar{X}_n \geq \bar{x}_c | \mu = \mu_1) \\ &= P\left(\frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}} > \frac{\bar{x}_c - \mu_1}{\sigma/\sqrt{n}}\right) \\ &= P(Z > z_c) \\ &= 1 - \Phi_Z\left(\frac{\bar{x}_c - \mu_1}{\sigma/\sqrt{n}}\right) \end{aligned} \quad (5.81)$$

essendo, sotto  $H_1$ ,  $\bar{X}_n \underset{H_1}{\sim} N\left(\mu_1, \frac{\sigma^2}{n}\right)$ .

### 5.4.1.2 Test per la media (campionamento da $N(\mu, \sigma^2)$ con $\sigma^2$ non noto)

Il test che abbiamo appena trattato è un test *esatto* in quanto poggia sulla distribuzione *esatta* della statistica test (5.72). Di conseguenza possiamo dire *esatte* anche la regione critica  $C_\alpha$  e la potenza  $\eta_{C_\alpha}$ .

Assumere che la varianza della popolazione  $\sigma^2$  sia *nota*, come fatto nell'esempio precedente, può essere una forzatura che adesso vorremmo rimuovere ponendoci la seguente domanda: come costruire un test per la verifica di un'ipotesi in merito alla media della popolazione assumendo in campione provenire da una distribuzione Normale della quale però non si conosce né media né varianza? Inoltre, è ancora possibile costruire un test *esatto*?

La risposta è affermativa e, in particolare, tale test può essere poggiato sulla seguente *statistica pivot*

$$T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim t_{n-1} \quad (5.82)$$

con  $t_{n-1}$  distribuzione di *Student* con  $\nu = (n - 1)$  gradi di libertà.

Sia  $(X_1, X_2, \dots, X_n)$  un campione causuale da una distibuzione Normale  $N(\mu, \sigma^2)$  con  $\mu$  e  $\sigma^2$  quantità incognite e si voglia verificare il seguente sitema di ipotesi

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ vs. \\ H_1 : \mu > \mu_0 \end{cases} \quad (5.83)$$

che risulta costituito da ipotesi *composite unilaterali*.

Ora ripercorrendo i passi dell'esempio precedente, fissato in  $\alpha$  il livello di significatività del test e tenendo conto di (5.82), si arriva a individuare la regione critica del test di livello  $\alpha$

$$\begin{aligned} C_\alpha &= \left\{ (x_1, x_2, \dots, x_n) \in \mathfrak{X} \mid \frac{\bar{x}_n - \mu_0}{S_n / \sqrt{n}} \geq t_{n-1; \alpha} \right\} \\ &= \left\{ (x_1, x_2, \dots, x_n) \in \mathfrak{X} \mid \bar{x}_n \geq \mu_0 + t_{n-1; \alpha} \frac{S_n}{\sqrt{n}} \right\} \end{aligned} \quad (5.84)$$

dove  $t_{n-1; \alpha}$  rappresenta il *valore critico* ossia l'*estremo inferiore* della regione critica *esatta* per la statistica test  $T$ , valore che si ottiene risolvendo la seguente equazione in  $t_{crit}$

$$P(T > t_{crit} \mid \mu = \mu_0) = \alpha \quad (5.85)$$

cosa che può essere agevolmente fatta ricorrendo alle tavole della distribuzione  $t$  di Student. Di conseguenza,  $\mu_0 + t_{n-1; \alpha} \frac{S_n}{\sqrt{n}}$  rappresenta il *valore critico* ovvero l'*estremo inferiore* della regione critica del test in questione declinata però in termini di media campionaria  $\bar{X}_n$ . Ne segue la *regola di decisione* (*esatta*)

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } t > t_{n-1; \alpha} \quad (5.86)$$

ovvero, ricordando che  $t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$ ,

$$\text{...si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } \bar{x}_n > \mu_0 + t_{n-1;\alpha} \frac{s_n}{\sqrt{n}} \quad (5.87)$$

La funzione di potenza esatta

$$\begin{aligned} \eta_{C_\alpha}(\mu) &= 1 - \beta(\mu) = 1 - P(\bar{X}_n \geq \mu_0 + t_{n-1;\alpha} \frac{s_n}{\sqrt{n}} \mid \mu > \mu_0) \\ &= P\left(\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \geq \frac{\mu_0 + t_{n-1;\alpha} (s_n/\sqrt{n}) - \mu}{s_n/\sqrt{n}}\right) \\ &= P\left(\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \geq t_{n-1;\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{s_n}\right) \\ &= 1 - P\left(\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \leq t_{n-1;\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{s_n}\right) \end{aligned} \quad (5.88)$$

sicché  $\eta_{C_\alpha}(\mu) \rightarrow 1$  per  $\mu \rightarrow \infty$  e  $\eta_{C_\alpha}(\mu) \rightarrow 0$  per  $\mu \rightarrow -\infty$ .

Prima di lasciare l'esempio osserviamo ancora che:

a) Il *livello di significatività* del test si può anche scrivere come

$$\alpha = \max_{\mu \leq \mu_0} \eta_{C_\alpha}(\mu)$$

b) la funzione di potenza in questo caso è *strettamente monotona* in  $\mu$  sicché possiamo cambiare l'ipotesi nulla  $H_0 : \mu = \mu_0$  in  $H_0 : \mu \leq \mu_0$

c) analogo discorso, *mutatis mutandis*, può essere fatto per il sistema di ipotesi

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ vs. \\ H_1 : \mu < \mu_0 \end{cases} \quad (5.89)$$

Riflettete, come esercizio, sulla ridefinizione della regione critica  $C_\alpha$  (e del suo complemento,  $\bar{C}_\alpha$ , cui spesso si dà il nome di *regione di non rifiuto o di accettazione*) indotta dalla riformulazione (5.89) del sistema di ipotesi circa  $\mu$ .

d) Per un qualsiasi valore di  $\alpha$ , i *valori critici*  $t_{n-1;\alpha}$  sono più grandi dei valori critici  $z_{1-\alpha}$ : il test  $t$  è perciò più conservativo di  $H_0$  (vale a dire, rifiuta  $H_0$  in un numero minore di casi che il test approssimato che potremmo costruire basandoci sui quantili  $z_{1-\alpha}$  della distribuzione Normale standard, tenuto conto del fatto che la distribuzione t di Student converge in distribuzione proprio a quest'ultima, ossia  $T \xrightarrow{D} Z \sim N(0, 1)$ ).

Nonostante che già a partire da  $n > 30$  la distribuzione Normale standard garantisca una buona approssimazione della distribuzione t di Student, molti statistici preferiscono continuare a usare la distribuzione t anche quando  $n > 30$  sentendosi garantiti proprio dal suo essere maggiormente conservativa.

Quanto stabilito ai punti a) - c) ha valenza generale, quindi per procedure di test riguardanti un qualsiasi parametro o funzioni di esso (e non solo la media) di una qualsiasi distribuzione (non solo quella Normale cui l'esempio si riferiva).

### 5.4.1.3 Test per la proporzione (campionamento da $b(1, p)$ )

Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da una popolazione *dicotomica* con probabilità di successo  $p \in [0, 1]$  e formuliamo il seguente sistema di ipotesi (entrambe composite):

$$\begin{cases} H_0 : p \geq p_0 \\ vs. \\ H_1 : p < p_0 \end{cases} \quad (5.90)$$

Per prima cosa, individuiamo la statistica che meglio riesca a riassumere l'informazione contenuta nel campione circa la probabilità di successo  $p$  ed essendo  $X_i \sim b(1, p)$ ,  $i = 1, 2, \dots, n$ , la statistica cercata è

$$W_n = \sum_{i=1}^n X_i \sim b(n, p) \quad (5.91)$$

che rappresenta il numero di successi nelle  $n$  replicazioni della prova bernoulliana sotto identiche condizioni.

A partire dalla statistica  $W_n$  si può costruire una ragionevole *regola di decisione* che tenga conto della struttura del sistema di ipotesi (5.90) stabilendo che

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } W_n \leq K \quad (5.92)$$

Allo scopo di ottenere il valore di  $K$ , fissiamo il valore di  $\alpha$  ottenendo

$$\alpha = P(\text{rifiutare } H_0 \mid H_0 \text{ vera}) = P(W_n \leq K \mid p = p_0) = \sum_{w=0}^K \binom{n}{w} p_0^w (1 - p_0)^{n-w} \quad (5.93)$$

Quest'ultima per  $\alpha$  fissato costituisce un'equazione in  $K$  la cui soluzione restituisce l'*estremo superiore*  $k_c$  (o *valore critico*) della regione critica  $C_\alpha$  di livello  $\alpha$  data da

$$C_\alpha = \{(x_1, x_2, \dots, x_n) \in \mathfrak{X} \mid W_n \leq k_c\} \quad (5.94)$$

ergo la giustificazione della regola di decisione poc'anzi stabilita e adottata.

Supponiamo ora  $n = 20$  e  $p_0 = 0.7$  sicché, sotto  $H_0$ ,  $W_n \sim b(n = 20, p = 0.7)$  e fissiamo  $\alpha = 0.15$ . Imponiamo tale condizione sull'errore di I° tipo e determiniamo il valore critico  $k_c$ :

$$0.15 = \sum_{w=0}^{w_c} \binom{20}{w} (0.7)^w (1 - 0.7)^{20-w} \quad (5.95)$$

Si verifica computazionalmente che tale condizione è soddisfatta da un valore compreso tra:  $k_c' = 11$  e  $k_c'' = 12$ . Per scegliere il valore di  $k_c$ , confrontiamo le due seguenti probabilità condizionate:

$$P(W_n \leq 11 \mid p = 0.7) = 0.1133 \quad P(W_n \leq 12 \mid p = 0.7) = 0.227 \quad (5.96)$$

In ragione della maggior prossimità della rispettiva probabilità condizionata al valore  $\alpha = 0.15$  scegliamo come *valore critico*  $k_c = k_c' = 11$  e specifichiamo comunque la regione critica di livello  $\alpha = 0.15$

$$C_\alpha = \{(x_1, x_2, \dots, x_n) \in \mathfrak{X} \mid W_n \leq 11\} \quad (5.97)$$

ergo, la specificazione della *regola di decisione* recita:

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } W_n \leq 11 \quad (5.98)$$

Si noti che, in generale, se la distribuzione da cui proviene il campione casuale è discreta, si può avere in vincolo sui valori ammissibili per  $K$  (qui,  $K \in \{0, 1, \dots, n\}$ ). Ciò talvolta comporta l'impossibilità di trovare un valore critico (nel nostro caso  $K$ ) corrispondente al valore di  $\alpha$  fissato.

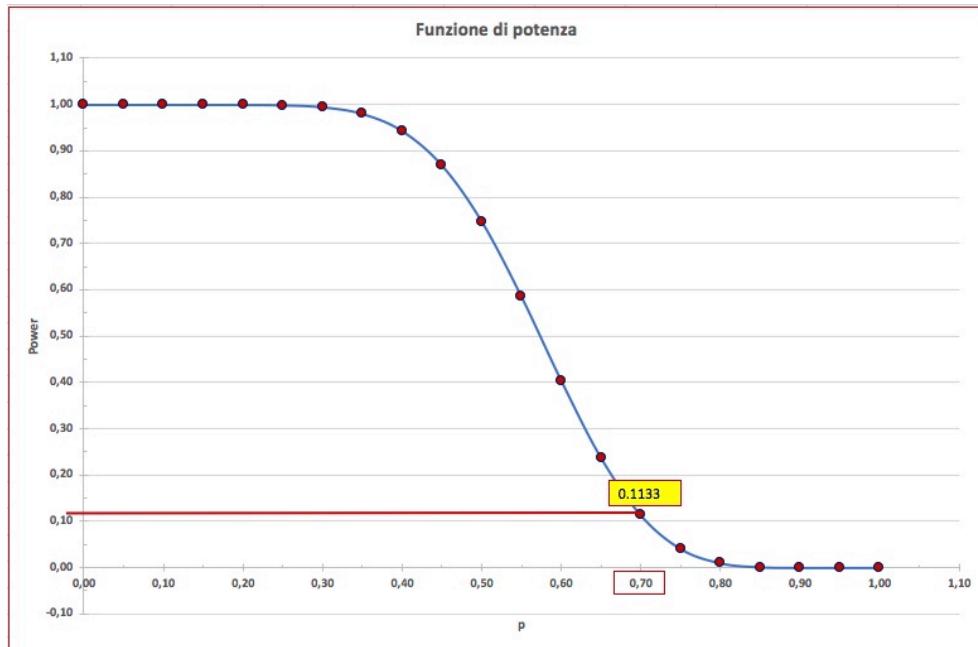
A questo punto, determiniamo la potenza del test. Ora più che di potenza, in questo caso, potremmo parlare di una *vera e propria funzione di potenza* al variare di  $p$  sotto  $H_1$  e restituita da

$$\eta_{C_\alpha}(p) = P(W_n \leq 11 | p < p_0) = \sum_{w=0}^{11} \binom{20}{w} p^w (1-p)^{20-w} \quad (5.99)$$

che si può valutare per valori di  $p \leq 0.7$

**Tabella 5.2:** Valori della funzione di potenza per valori di  $p \leq 0.7$

$p$	0.70	0.65	0.60	0.55	0.50	0.45	0.40
$\eta_{C_\alpha}(p)$	0.113	0.238	0.404	0.586	0.748	0.869	0.943



**Figura 5.4:** Funzione di potenza per  $H_0 : p \geq 0.7$  vs.  $H_1 : p < 0.7$

La Figura 5.4 riporta il grafico della funzione di potenza  $\eta_{C_\alpha}(p)$  per l'esempio che stiamo trattando.

Qualche considerazione in merito alla funzione di potenza prima di procedere:

a) la funzione di potenza di un test è una *misura della reattività/sensibilità* del test stesso all'*allontanamento* da  $H_0$  e quindi, sotto questo aspetto, ne misura la *bontà*. Ora, potendo scegliere tra diversi test (= diverse regole di decisione), propenderemo per test con elevata potenza, possibilmente massima, poiché tali test mostrano un'*alta capacità di discernimento* tra  $H_0$  e  $H_1$ . L'individuazione di un test *ottimo* può essere condotta attraverso l'analisi delle funzioni di potenza associate ai diversi test competitori per un livello  $\alpha$  fissato; molto sbrigativamente, sia  $C_{\alpha;1}$  la regione critica associata al test  $T_1$  che diremo essere ottimo se

$$\eta_{C_{\alpha;1}}(\theta) > \eta_{C_\alpha}(\theta), \quad \forall \theta \in \Theta_1 \quad (5.100)$$

dove  $\eta_{C_\alpha}(\theta)$  è la funzione di potenza associata a un qualsiasi altro test competitore di  $T_1$  del medesimo livello  $\alpha$ . Torneremo tra non molto sull'argomento (*Lemma di Neyman-Pearson*) in maniera decisamente più esaustiva e formale.

b) in molti casi la *funzione di potenza* è *monotona* (crescente o decrescente a seconda della *forma dell'ipotesi alternativa*)

Possiamo anche ottenere un test *approssimato* per il sistema di ipotesi 5.90. E' sufficiente notare che, per  $n > 30$ , per il *Teorema Limite Centrale*

$$W_n = \sum_{i=1}^n X_i \underset{a}{\sim} N(np, np(1-p)) \quad (5.101)$$

Possiamo dunque costruire un test approssimato (o asintotico) poggiandolo, nella determinazione della corrispondente regione critica (ergo della regola di decisione) sulla statistica test

$$Z_n = \frac{W_n - np}{\sqrt{np(1-p)}} \underset{a}{\sim} N(0, 1) \quad (5.102)$$

In maniera analoga a quanto fatto nell'Esercizio, fissato  $\alpha$ , ricaviamo la regione critica (*asintotica*)

$$C_\alpha = \{\boldsymbol{x} \in \mathfrak{X} \mid z_n \leq -z_{1-\alpha}\} \quad (5.103)$$

da cui la *regola di decisione (asintotica)*

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ se } w_n \leq np_0 - z_{1-\alpha} \sqrt{np_0(1-p_0)} \quad (5.104)$$

La funzione di potenza (anch'essa *asintotica*) è

$$\eta_{C_\alpha}(p) = P(W_n \leq k_c \mid p < p_0) = P\left(\frac{w_n - np}{\sqrt{np(1-p)}} \leq \frac{k_c - np}{\sqrt{np(1-p)}}\right) \quad (5.105)$$

È facile verificare che la funzione di potenza *asintotica*  $\eta_{C_\alpha}$  è *monotona decrescente* e che  $\eta_{C_\alpha} \rightarrow 1$  per  $p \rightarrow 0$ , mentre  $\eta_{C_\alpha} \rightarrow 0$  per  $p \rightarrow 1$ ; di conseguenza, la funzione di potenza *asintotica* sarà dunque del tutto simile a quella illustrata in figura 5.4.

#### 5.4.1.4 Test unilaterale sulla media per grandi campioni

Supponiamo di avere un campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione  $F_X(x; \theta)$ ,  $\theta \in \Theta$  e supponiamo che  $n$  sia grande ( $n > 30$ ). Supponiamo di aver

formulato le seguenti *ipotesi composite* in merito alla media  $\mu$  della popolazione da cui proviene il campione casuale

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ vs. \\ H_1 : \mu > \mu_0 \end{cases} \quad (5.106)$$

In questo caso, non disponiamo di informazioni sulla forma distributiva di  $F_X(x; \theta)$  ma sappiamo che  $n > 30$ . Procediamo dunque come segue: stabiliamo di stimare il valore di  $\mu$  ricorrendo al suo stimatore *plug-in*  $\bar{X}_n$  e fissiamo il valore di  $\alpha$ . Sulla base del *Teorema Limite Centrale* approssimiamo la distribuzione di  $\bar{X}_n$ :

$$\bar{X}_n \underset{a}{\sim} N(\mu, \sigma^2/n). \quad (5.107)$$

Poiché non conosciamo il valore di  $\sigma^2$ , tenuto conto del fatto che  $S_n^2 \xrightarrow{P} \sigma^2$ , lo si può stimare proprio con  $S_n^2$ , ottenendo, per il *Teorema di Slutsky*, che sotto  $H_0$

$$Z_n = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}} \underset{a}{\sim} N(0, 1) \quad (5.108)$$

tenuto conto del fatto che possiamo sostituire la condizione  $\mu \leq \mu_0$  con  $\mu = \mu_0$  in quanto tale valore è il valore limite. Se la condizione vale per  $\mu_0$ , allora vale anche per tutti i valori minori di  $\mu_0$ . Risolvendo la seguente equazione in  $\bar{z}_c$

$$\alpha = P(Z_n > \bar{z}_c) \quad (5.109)$$

si ha  $\bar{z}_c = z_{1-\alpha}$ ; pertanto

$$\bar{z}_c = \frac{\bar{x}_c - \mu_0}{s_n / \sqrt{n}} = z_{1-\alpha} \quad (5.110)$$

e infine con pochi e semplici passaggi algebrici, si ottiene il *valore critico del test asintotico per la media* di livello  $\alpha$

$$\bar{x}_c = \mu_0 + z_{1-\alpha} \frac{s_n}{\sqrt{n}} \quad (5.111)$$

che a sua volta costituisce il limite inferiore della regione critica  $C_\alpha$ . Allora, la regione critica di livello  $\alpha$  per il test in questione sarà data da

$$C_\alpha = \left\{ (x_1, x_2, \dots, x_n) \in \mathfrak{X} \mid \bar{x}_n > \mu_0 + z_{1-\alpha} \frac{s_n}{\sqrt{n}} \right\} \quad (5.112)$$

e dunque, in accordo con la *regola di decisione* stabilita (5.73), possiamo riscrivere quest'ultima nella maniera seguente

$$\text{...si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } \bar{X}_n > \mu_0 + z_{1-\alpha} \frac{s_n}{\sqrt{n}} \quad (5.113)$$

La funzione di potenza (*asintotica*)

$$\begin{aligned} \eta_{C_\alpha}(\mu) &= 1 - \beta(\mu) = 1 - P(\bar{X}_n > \mu_0 + z_{1-\alpha} \frac{s_n}{\sqrt{n}} \mid \mu > \mu_0) \\ &= P\left(\frac{\bar{X}_n - \mu}{s_n / \sqrt{n}} > \frac{\mu_0 + z_{1-\alpha} (s_n / \sqrt{n}) - \mu}{s_n / \sqrt{n}}\right) \\ &= P\left(\frac{\bar{X}_n - \mu}{s_n / \sqrt{n}} > z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{s_n}\right) \\ &= 1 - \Phi_Z\left(z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{s_n}\right) \end{aligned} \quad (5.114)$$

sicché  $\eta_{C_\alpha}(\mu) \rightarrow 1$  per  $\mu \rightarrow \infty$  e  $\eta_{C_\alpha}(\mu) \rightarrow 0$  per  $\mu \rightarrow -\infty$ .

Prima di lasciare l'esempio osserviamo ancora che:

a) Il *livello di significatività (approssimato)* del test si può anche scrivere come

$$\alpha = \max_{\mu \leq \mu_0} \eta_{C_\alpha}(\mu)$$

b) questa funzione di potenza *approssimata* è *strettamente monotona* in  $\mu$  sicché possiamo *cambiare* l'ipotesi nulla  $H_0 : \mu = \mu_0$  in  $H_0 : \mu \leq \mu_0$

c) analogo discorso, *mutatis mutandis*, può essere fatto per il sistema di ipotesi

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ vs. \\ H_1 : \mu < \mu_0 \end{cases} \quad (5.115)$$

Riflettete, come esercizio, sulla ridefinizione della regione critica  $C_\alpha$  (e del suo complemento,  $\bar{C}_\alpha$ , cui spesso si dà il nome di *regione di non rifiuto o di accettazione*) indotta dalla riformulazione (5.115) del sistema di ipotesi circa  $\mu$ .

Quanto stabilito ai punti a) - c) ha *valenza generale*, quindi per procedure di test riguardanti un qualsiasi parametro o funzioni di esso (e non solo la media) di una qualsiasi distribuzione (non solo quella Normale cui l'esempio si riferiva).

#### 5.4.1.5 Test sulla varianza

Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Normale,  $N(\mu, \sigma^2)$  con  $\mu$  e  $\sigma^2$  parametri *non noti*. Vogliamo ora verificare il seguente sistema di ipotesi:

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ vs. \\ H_1 : \sigma^2 = \sigma_1^2, \quad \sigma_1^2 > \sigma_0^2 \end{cases} \quad (5.116)$$

Ora, sotto  $H_0$ ,

$$W = \frac{(n-1) S_n^2}{\sigma_0^2} \sim \chi_{n-1} \quad (5.117)$$

sicché, fissato  $\alpha$ , risolvendo la seguente equazione nell'incognita  $w_{crit}$ , *valore critico* per la statistica test  $W$ ,

$$P \left( \frac{(n-1) S_n^2}{\sigma_0^2} > w_{crit} \right) = \alpha \quad (5.118)$$

si ha  $w_{crit} = \chi_{n-1; \alpha}$  sicché si ricava immediatamente la regione critica di livello  $\alpha$  associata al test

$$C_\alpha = \left\{ x \in \mathfrak{X} : w = \frac{(n-1) s_n^2}{\sigma_0^2} > \chi_{n-1; \alpha} \right\} \quad (5.119)$$

e la conseguente regola di decisione

$$\text{...si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } w = \frac{(n-1)s_n^2}{\sigma_0^2} > \chi_{n-1;\alpha} \quad (5.120)$$

Per esempio, sia  $n = 25$ ,  $\sigma_0 = 15$  e  $\sigma_1 = 18.3$ ; inoltre,  $s_n^2 = 17.4$ . Fissato  $\alpha = 0.05$ , dalle tavole della distribuzione chi-quadrato si ha

$$w_{crit} = \chi_{25-1;0.05} = \chi_{24;0.05} = 36.415 \quad (5.121)$$

sicché

$$w = \frac{24 \cdot 17.4}{15} = 27.84 \quad (5.122)$$

e quunque, essendo  $w = 27.84 < 36.415$ , ovvero  $w \notin C_{0.05}$ , l'ipotesi nulla  $H_0 : \sigma = 15$  non può essere rifiutata.

Quanto detto a proposito del sistema di ipotesi (5.116) può essere esteso anche a sistemi di ipotesi *unilaterali* del tipo

$$H_0 : \sigma^2 \leq \sigma_0^2 \text{ vs. } H_1 : \sigma^2 > \sigma_0^2 \quad (5.123)$$

o

$$H_0 : \sigma^2 \geq \sigma_0^2 \text{ vs. } H_1 : \sigma < \sigma_0^2 \quad (5.124)$$

Per il sistema di ipotesi (5.123) la regione critica coincide con (5.119) mentre per il sistema di ipotesi (5.124) essa va ridefinita nella seguente maniera:

$$C_\alpha = \left\{ \boldsymbol{x} \in \mathfrak{X} : w = \frac{(n-1)s_n^2}{\sigma_0^2} < \chi_{n-1;(1-\alpha)} \right\} \quad (5.125)$$

#### 5.4.1.6 Test bilaterali

Ci sono situazioni in cui si ritiene una certa ipotesi nulla  $H_0 : \theta = \theta_0$  *inedeguata* ma vi è incertezza circa la *direzione* da assegnare nel riformularla, e in ultima analisi sulla *direzione* da assegnare all'ipotesi alternativa  $H_1$ . In questo caso è certamente più ragionevole formulare il sistema di ipotesi da sottoporre a verifica nella maniera seguente

$$\begin{cases} H_0 : \theta = \theta_0 \\ \text{vs.} \\ H_1 : \theta \neq \theta_0 \end{cases} \quad (5.126)$$

Come di può notare, l'ipotesi alternativa  $H_1$  risulta essere *bilaterale* e di *test bilaterali* dunque parleremo. Soffermiamoci per il momento su test *bilaterali* per la media  $\mu$  della popolazione.

#### 5.4.1.7 Test bilaterale sulla media (esatto)

Supponiamo di avere un campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione  $N(\mu, \sigma^2)$  con  $\sigma^2$  non noto. Abbiamo già osservato che assumere la varianza  $\sigma^2$  nota implica una irragionevole forzatura: se non conosciamo la media  $\mu$  della popolazione per quale ragione dovremmo conoscerne la varianza  $\sigma^2$ ?

Formuliamo dunque il seguente sistema di ipotesi

$$\begin{cases} H_0 : \mu = \mu_0 \\ vs. \\ H_1 : \mu \neq \mu_0 \end{cases} \quad (5.127)$$

Ora, fissiamo il valore di  $\alpha$  e dovendo individuare una statistica pivot per  $\mu$  è naturale pensare a

$$T = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}} \underset{H_0}{\sim} t_{n-1} \quad (5.128)$$

Troviamo la ragione critica di livello  $\alpha$  del test in questione, imponendo

$$\alpha = P(|T| \geq t_c \mid \mu = \mu_0) = P\left(\left|\frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}\right| \geq t_c\right) \quad (5.129)$$

e risolvendo questa equazione in  $t_c$  si ottiene  $t_c = t_{n-1; \alpha/2}$  sicché

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : t = \left| \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{s_n} \right| \geq t_{n-1; \alpha/2} \right\} \quad (5.130)$$

In questo caso, data la presenza del valore assoluto,  $C_\alpha = C'_\alpha \cup C''_\alpha$  e di conseguenza, la nostra regola di decisione sarà

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } t = \left| \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{s_n} \right| > t_{n-1; \alpha/2} \quad (5.131)$$

In questo caso, anche la funzione di potenza  $\eta_{C_\alpha}(\mu)$  associata al test bilaterale avrà una forma differente da quella vista in precedenza associata a ipotesi alternative unilaterali. In particolare, ci aspettiamo che  $\eta_{C_\alpha} \rightarrow 1$  per  $\mu \rightarrow \pm\infty$  (ovvero che reagisca all'allontanamento da  $H_0$ ), che

$$\alpha = \max_{\mu \neq \mu_0} \eta_{C_\alpha}(\mu) = \eta_{C_\alpha}(\mu_0)$$

e che abbia un andamento quale quello nella figura 5.5.

Un paio di considerazioni:

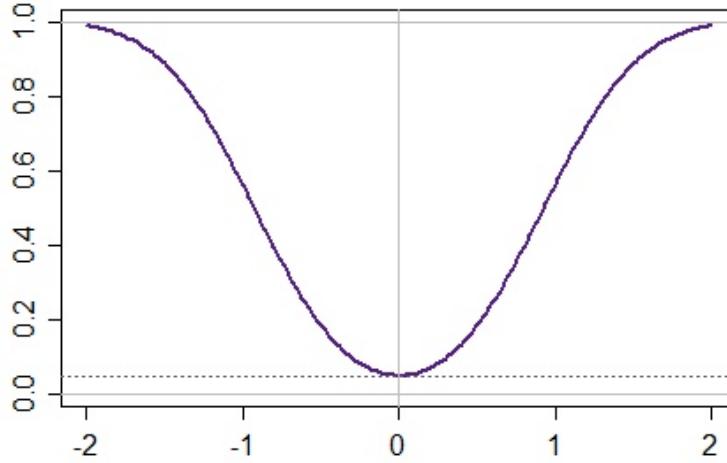
- a) come si può facilmente osservare dalla *forma del complemento  $\bar{C}_\alpha$*  (o *regione di non rifiuto*) della regione critica  $C_\alpha$ , esiste un'interessante relazione tra *test bilaterali e intervalli di confidenza*:  $\bar{C}_\alpha$  coincide con  $IC_\mu(1 - \alpha)$

$$\bar{C}_\alpha = \left[ \bar{x}_n - t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}} \right] \quad (5.132)$$

Ergo, laddove  $\mu_0 \in IC_\mu(1 - \alpha)$  non si hanno ragioni sufficientemente fondate per *rifiutare  $H_0 : \mu = \mu_0$* .

- b) laddove il *campione sia grande* (e in questo caso, non necessariamente da distribuzione Normale), in virtù del fatto che  $T \xrightarrow{D} N(0, 1)$ , sostituendo  $z_{1-\alpha/2}$  al posto di  $t_{n-1, \alpha/2}$  in (5.131) si ottiene la regione critica di un test bilaterale *asintotico* per la media  $\mu$  della popolazione, da cui la seguente *regola di decisione (asintotica)* che recita

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } z = \left| \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{s_n} \right| > z_{1-\alpha/2}$$



**Figura 5.5:** Funzione di potenza  $\eta_{C_\alpha}(\mu)$  per il test (esatto) bilaterale per  $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu \neq \mu_0$

#### 5.4.1.8 Test bilaterale per differenza di medie (esatto)

Consideriamo due popolazioni aventi rispettivamente distribuzione  $N(\mu_1, \sigma_1^2)$  e  $N(\mu_2, \sigma_2^2)$  dalle quali estraiamo i due campioni casuali  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  e  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$  e supponiamo di essere interessati a sottoporre a verifica il sistema di ipotesi

$$\begin{cases} H_0 : \mu_1 = \mu_2 & (\equiv \mu_1 - \mu_2 = 0) \\ \text{vs.} \\ H_1 : \mu_1 \neq \mu_2 & (\equiv \mu_1 - \mu_2 \neq 0) \end{cases} \quad (5.133)$$

Assumiamo che:

- a)  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ : vale a dire le due popolazioni hanno la *stessa varianza* (*ipotesi di omoschedasticità*) sicché siamo in presenza di *location model* poiché le distribuzioni differiscono solo per la *locazione* (vale a dire, per la media)
- b) i campioni  $\mathbf{X} = (X_1, X_2, \dots, X_{n_1})$  e  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{n_2})$  sono *indipendenti*

Come abbiamo osservato poc'anzi, grazie alla relazione che corre tra *regione di non rifiuto*  $\bar{C}_\alpha$  e intervallo di confidenza di livello  $(1 - \alpha)$ , e ricordando che da (5.49)

$$IC_\Delta(1 - \alpha) : \left[ (\bar{x} - \bar{y}) - t_{n-2, 1-\frac{\alpha}{2}} \frac{s_P}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}}, (\bar{x} - \bar{y}) + t_{n-2, 1-\frac{\alpha}{2}} \frac{s_P}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}} \right] \quad (5.134)$$

dove  $\Delta = \mu_1 - \mu_2$  e  $s_P$  è la radice quadrata della stima della *pooled variance*, si ha che

$$\bar{C}_\alpha = \left[ (\bar{x} - \bar{y}) - t_{n-2, 1-\frac{\alpha}{2}} \frac{s_P}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}}, (\bar{x} - \bar{y}) + t_{n-2, 1-\frac{\alpha}{2}} \frac{s_P}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}} \right] \quad (5.135)$$

da cui, ragionando per complemento, avremo la seguente regola di decisione

$$\text{...si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } t = \left| \frac{(\bar{x} - \bar{y})}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}} s_P} \right| > t_{n-2; \alpha/2} \quad (5.136)$$

Per maggiori dettagli relativi ai calcoli, si rinvia al paragrafo dedicato alla costruzione di intervalli di confidenza per la differenza delle medie di due popolazioni.

Resta ancora inteso che si può facilmente ottenere un test *approssimato* per la verifica dei sistemi di ipotesi (5.133), laddove i due campioni provenienti dalle due popolazioni (in questo caso, non necessariamente Normali) siano *grandi*, semplicemente sostituendo in (5.136)  $z_{1-\alpha/2}$  al posto di  $t_{n-2; \alpha/2}$ .

Lascio per esercizio ricavare e disegnare la *funzione di potenza* per questa classe di test; fin da subito, però, possiamo aspetterci un andamento assai simile a quello di Figura 5.5.

#### 5.4.1.9 Test per differenza di proporzioni

Supponiamo di essere interessati a verificare il seguente sistema di ipotesi:

$$\begin{cases} H_0 : p_1 = p_2 & (\equiv p_1 - p_2 = 0) \\ vs. \\ H_1 : p_1 \neq p_2 & (\equiv p_1 - p_2 \neq 0) \end{cases} \quad (5.137)$$

Tenuto conto del fatto che lo stimatore *plug-in*  $\hat{p}_n$  della proporzione altro non è che la media campionaria di un campione proveniente da una popolazione bernoulliana, ricordando la LDGN (o, il TLC), considerati due campioni casuali  $(X_1, X_2, \dots, X_{n_1})$  e  $(Y_1, Y_2, \dots, Y_{n_2})$  rispettivamente di ampiezza  $n_1$  e  $n_2$ , e calcolata su ciascuno di essi la frequenza relativa campionaria  $\hat{p}_X$  e  $\hat{p}_Y$  si ha che

$$W = \frac{(\hat{p}_X - \hat{p}_Y) - (p_1 - p_2)}{\sqrt{\text{Var}(\hat{p}_X - \hat{p}_Y)}} \underset{a}{\sim} N(0, 1) \quad (5.138)$$

sicché regione critica e regola di decisione saranno uguali a quelle ottenute nella declinazione asintotica dell'esempio precedente.

#### 5.4.1.10 Test per il confronto di varianze

Supponiamo di disporre di due campioni casuali indipendenti di cui il primo,  $(X_1, X_2, \dots, X_{n_1})$  segue la distribuzione  $N(\mu_1, \sigma_1^2)$  e il secondo,  $(Y_1, Y_2, \dots, Y_{n_2})$  segue la distribuzione  $N(\mu_2, \sigma_2^2)$ . Supponiamo inoltre che i parametri  $\mu$  e  $\sigma^2$  siano incogniti in entrambe le distribuzioni. Formuliamo ora le seguenti ipotesi in merito alle varianze delle due distribuzioni:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ vs. \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \quad (5.139)$$

Ora, fissiamo il valore di  $\alpha$  e scegliamo la seguente *statistica test* per il sistema di ipotesi in questione,

$$W = \frac{S_Y^2 / \sigma_2^2}{S_X^2 / \sigma_1^2} = \frac{S_Y^2}{S_X^2} \cdot \frac{\sigma_1^2}{\sigma_2^2} \quad (5.140)$$

Ora, sotto  $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$  si ha

$$W = \frac{S_Y^2}{S_X^2} \cdot \frac{\sigma^2}{\sigma^2} = \frac{S_Y^2}{S_X^2} \sim F_{(n_2-1), (n_1-1)} \quad (5.141)$$

Ricordando che  $S_X^2 \xrightarrow{P} \sigma_1^2$  e  $S_Y^2 \xrightarrow{P} \sigma_2^2$  e stimata  $W$  con  $w = \frac{s_Y^2}{s_X^2}$ , è del tutto ragionevole stabilire la seguente *regola di decisione*:

...si rifiuta  $H_0$  in favore di  $H_1$  quando  $w$  è lontano da 1 (5.142)

o in altri termini,

...si rifiuta  $H_0$  in favore di  $H_1$  quando  $w \notin [w'_{crit}, w''_{crit}]$  (5.143)

dove

$$w''_{crit} = w_{(n_2-1), (n_1-1); \alpha/2} \quad (5.144)$$

e

$$w'_{crit} = w_{(n_2-1), (n_1-1); (1-\alpha/2)} = \frac{1}{w_{(n_1-1), (n_2-1); \alpha/2}} \quad (5.145)$$

sono i quantili, rispettivamente di ordine  $\alpha/2$  e  $1 - \alpha/2$ , della distribuzione  $F$  di Fisher-Snedecor con  $(n_2 - 1), (n_1 - 1)$  gradi di libertà.

Per esempio, supponiamo  $n_1 = 14$  e  $n_2 = 10$ ,  $s_X^2 = 17.4$  e  $s_Y^2 = 37.9$  e fissiamo  $\alpha = 0.05$ . Allora, datte tavole della distribuzione di Fisher-Snedecor,

$$w_{(10-1), (14-1); 0.025} = 3.312 \quad (5.146)$$

e

$$w_{(10-1), (14-1); (1-0.025)} = \frac{1}{w_{(14-1), (10-1); 0.025}} = \frac{1}{3.87} = 0.26 \quad (5.147)$$

sicché  $\bar{C}_{0.05} = [0.26, 3.312]$ . Ora,

$$w = \frac{37.9}{17.4} = 2.178 \in \bar{C}_{0.05} = [0.26, 3.312] \quad (5.148)$$

sicché, in virtù della regola di decisione (5.143), l'ipotesi nulla  $H_0 : \sigma_1^2 = \sigma_2^2$  non può essere rifiutata.

Lascio per esercizio costruire le *regole di decisione* per i sistemi di ipotesi unilaterali

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 > \sigma_2^2 \quad (5.149)$$

e

$$H_0 : \sigma_1^2 \geq \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 < \sigma_2^2 \quad (5.150)$$

## 5.5 Test di Kolmogorov-Smirnov

Il test di Kolmogorov-Smirnov è un *test non parametrico* impiegato per la verifica di ipotesi sulla *forma* delle distribuzioni; esso è comunemente utilizzato per confrontare la distribuzione empirica con la distribuzione da cui si ipotizza provenga il campione e, in ultima analisi, per verificare ipotesi sulla distribuzione generante i dati. L'idea alla base del test di Kolmogorov-Smirnov è quella di confrontare la *distanza* tra la funzione di distribuzione empirica e quella ipotizzata sotto ipotesi nulla.

Supponiamo di avere un campione casuale  $(X_1, X_2, \dots, X_n)$  che proviene da una distribuzione  $F$  sconosciuta e di sottoporre a test il sistema di ipotesi

$$\begin{cases} H_0 : F = F_0 \\ vs. \\ H_1 : F \neq F_0 \end{cases} \quad (5.151)$$

con  $F_0$  una particolare (e *specificata*) distribuzione. Ricordiammo che

$$\hat{F}_n(x) = \hat{P}_n(X \leq x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i) \quad (5.152)$$

è la funzione di distribuzione *empirica*, stimatore *non distorto* di  $F(x) = P(X \leq x)$  (vale a dire,  $\mathbb{E}_{F(x)}[\hat{F}_n(x)] = 0$ ) e, per un qualsiasi valore  $x \in \mathbb{R}$  fissato, la LDGN implica che

$$\hat{F}_n(x) \xrightarrow{P} F(x) = P(X \leq x) \quad (5.153)$$

ossia, la proporzione campionaria nell'intervallo  $(-\infty, x]$  approssima la probabilità  $P(X \leq x)$  definita su quello stesso intervallo. In altre parole,  $\hat{F}_n(x)$  è uno stimatore *consistente* di  $F(x)$ ; e lo è anche *quadraticamente* poiché, per ogni  $x \in \mathbb{R}$  fissato,

$$\begin{aligned} MSE_{F(x)}(\hat{F}_n(x)) &= \mathbb{V}ar_{F(x)}(\hat{F}_n(x)) + \mathbb{E}_{F(x)}^2[\hat{F}_n(x)] \\ &= \frac{F(x)[1 - F(x)]}{n} + 0 \xrightarrow{n \rightarrow \infty} 0. \end{aligned} \quad (5.154)$$

Inoltre, per il TLC,

$$\sqrt{n} \left( \hat{F}_n(x) - F(x) \right) \underset{a}{\sim} N(0, F(x)[1 - F(x)]) \quad (5.155)$$

Il seguente teorema determina il *comportamento asintotico* di  $\hat{F}_n(x)$

**Teorema 5.5.1** (Glivenko-Cantelli (1933)). Siano  $X_1, X_2, \dots, X_n$  v.c. indipendenti e identicamente distribuite aventi distribuzione  $F(x)$ . Sia

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i) \quad (5.156)$$

la funzione di distribuzione empirica che *approssima* l'ignota  $F(x)$ . Sia

$$D_n = \sup_x |\hat{F}_n(x) - F(x)| \quad (5.157)$$

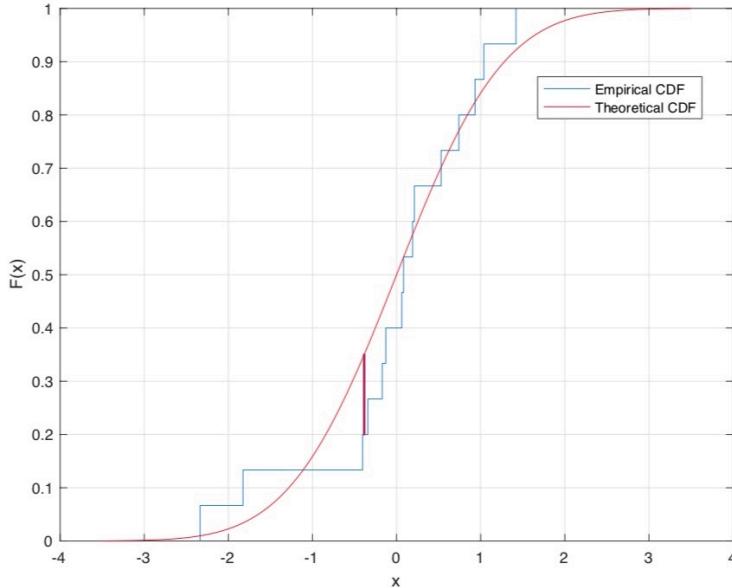
la massima deviazione di  $\hat{F}_n(x)$  da  $F(x)$ . Allora

$$D_n \xrightarrow{P} 0 \quad \equiv \quad \lim_{n \rightarrow \infty} P \left[ \sup_x | \hat{F}_n(x) - F(x) | \right] = 0 \quad (5.158)$$

o, equivalentemente, la successione delle distribuzioni empiriche  $\{\hat{F}_n(x)\}_{n \in \mathbb{N}}$  converge per  $n \rightarrow \infty$  uniformemente (ossia  $\forall x \in \mathbb{R}$ ) a  $F(x)$  con probabilità 1.

*Dimostrazione.* Per la dimostrazione vedere... □

In merito all'andamento della funzione di distribuzione empirica abbiamo già avuto modo di vedere (grafico 2.1 del paragrafo 3.1.1); in particolare,  $D_n$  è restituita dal *massimo scostamento*, qui rappresentato dal segmento in bordeaux, tra la curva in rosso ( $F(x)$ ) e la spezzata in blu ( $\hat{F}_n(x)$ ) che ne è l'approssimante.



**Figura 5.6:**  $\hat{F}_n(x)$ ,  $F(x)$  e massimo scostamento  $D_n$  (segmento bordeaux)

Prendiamo ora in considerazione il seguente teorema.

**Teorema 5.5.2.** Sia  $F(x)$  una funzione di distribuzione continua e sia  $X_1, X_2, \dots, X_n$  una  $n$ -pla di v.c. indipendenti e identicamente distribuite con funzione di distribuzione  $F(x)$ . Allora,

- a) la distribuzione sotto ipotesi nulla di

$$D_n = \sup_x | \hat{F}_n(x) - F(x) | \quad (5.159)$$

non dipende da  $F(x)$  ma solamente da  $n$ .

- b) la distribuzione di  $\sqrt{n}D_n$  converge in distribuzione alla distribuzione di una v.c. di Kolmogorov-Smirnov la cui funzione di distribuzione è

$$Q(x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2} \quad (5.160)$$

*Dimostrazione.* Ci limitiamo a dimostrare la tesi di cui al punto a). La dimostrazione del punto b) esula dal livello di questo corso e ci limiteremo qui ad alcune considerazioni.

*Punto a):* definiamo l'inversa di  $F$  nella maniera seguente

$$F^{-1}(y) = \min\{x : F(x) \geq y\} \quad (5.161)$$

ed effettuando il cambio di variabili  $y = F(X)$  e  $x = F^{-1}(y)$  possiamo scrivere

$$P\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \leq t\right) = P\left(\sup_{0 \leq y \leq 1} |\hat{F}_n(F^{-1}(y)) - y| \leq t\right). \quad (5.162)$$

Usando la definizione di funzione di distribuzione empirica si ha

$$\hat{F}_n(F^{-1}(y)) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, F^{-1}(y)]}(X_i) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, y]}(U_i) \quad (5.163)$$

essendo  $F(X_i) = U_i \sim \mathcal{U}(0, 1)$  dal momento che la funzione di distribuzione cumulata di  $F(X_i)$  è

$$P(F(X_i) \leq t) = P(X_i \leq F^{-1}(t)) = F(F^{-1}(t)) = t. \quad (5.164)$$

e perciò

$$P\left(\sup_{0 \leq y \leq 1} |\hat{F}_n(F^{-1}(y)) - y| \leq t\right) = P\left(\sup_{0 \leq y \leq 1} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, y]}(U_i) - y \right| \leq t\right). \quad (5.165)$$

Essendo  $(X_1, X_2, \dots, X_n)$  indipendenti, per un noto teorema, le v.c.  $U_i = F(X_i)$  sono anch'esse indipendenti e hanno distribuzione Uniforme su  $[0, 1]$ . Abbiamo così dimostrato che

$$P\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \leq t\right) = P\left(\sup_{0 \leq y \leq 1} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, y]}(U_i) - y \right| \leq t\right). \quad (5.166)$$

che è chiaramente indipendente da  $F(x)$ .

*Punto b):* In merito al punto b) del precedente teorema ci limitiamo a osservare che nel caso in cui  $n$  non tenda a infinito (leggi *per piccoli campioni*) non è più vero che  $D_n \xrightarrow{D} Q(x)$  e pertanto la distribuzione di  $D_n$  deve essere ricavata per ogni  $n$  (e per ogni  $\alpha$  onde ricavare gli opportuni quantili che fungono da valori critici per il test di Kolmogorov-Smirnov). In generale, si usano *metodi Monte Carlo* che consistono nel

- i) generare un numero elevato di campioni casuali di dimensione  $n$
- ii) calcolare  $D_n$  per ciascun campione generato e costruire la sua distribuzione empirica
- iii) ricavare i quantili  $D_{n; \alpha}$  dalla distribuzione empirica di  $D_n$ .

□

Quanto stabilito dal precedente Teorema 5.5.2 ha cruciale conseguenza: considerato il sistema di ipotesi (5.151), siamo ora in grado di costruire la *regola di decisione* in merito a  $H_0$ . Infatti possiamo subito osservare che, se  $H_0$  è vera, la funzione di distribuzione empirica  $\hat{F}_n(x)$  convergerà in probabilità a  $F_0(x)$  comportando un piccolo valore di  $D_n$ ; al contrario, se  $H_0$  è falsa, la funzione di distribuzione empirica  $\hat{F}_n(x)$  convergerà in probabilità a  $F(x) \neq F_0(x)$  e ciò comporterà un grande valore di  $D_n$ .

### 5.5.0.1 Test di Kolmogorov-Smirnov per un campione

Supponiamo di avere un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da una distribuzione  $F$  sconosciuta e di voler sottoporre a test il sistema di ipotesi (5.151).

Ora, in virtù di quanto fin qui detto, se  $H_0$  è vera, la funzione di distribuzione empirica  $\hat{F}_n(x)$  convergerà in probabilità a  $F_0(x)$  comportando un piccolo valore di  $D_n$ ; al contrario, se  $H_0$  è vera, la funzione di distribuzione empirica  $\hat{F}_n(x)$  convergerà in probabilità a  $F(x) \neq F_0(x)$  e ciò comporterà un grande valore di  $D_n$ .

Formalmente, fissato  $\alpha$ , la soluzione dell'equazione in  $D_{crit}$

$$P(D_n \geq D_{crit} | H_0 : F(x) = F_0(x)) = \alpha \quad (5.167)$$

restituirà  $D_{crit} = D_{n;\alpha}$ , quantile di ordine  $\alpha$  della distribuzione di  $D_n$  valore che è stato opportunamente tabulato (Tavola della distribuzione di Kolmogorov-Smirnov). Allora, indicato con  $D_n^{oss}$  il valore di  $D_n$  calcolato sui dati a disposizione, la *regola di decisione* sarà dunque

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } D_n^{oss} > D_{n;\alpha} \quad (5.168)$$

Il test di Kolmogorov-Smirnov trova frequente applicazione nella verifica della *ipotesi di normalità* della distribuzione che ha prodotto la determinazione  $(x_1, x_2, \dots, x_n)$  del campione casuale  $(X_1, X_2, \dots, X_n)$ ; in questo caso, la distribuzione sotto ipotesi nulla di normalità ( $F_0$ ) sarà  $N(\mu_0, \sigma_0^2)$

**Esempio 5.5.1.** Consideriamo un insieme di  $n = 100$  osservazioni di cui interessa stabilire se esso può essere la determinazione di un campione casuale da una distribuzione Normale standard; pertanto, siamo interessati a verificare il seguente sistema di ipotesi

$$H_0 : F = N(0, 1) \quad \text{vs.} \quad H_1 : F \neq N(0, 1) \quad (5.169)$$

Nelle tabelle 5.7 e 5.8 riportiamo le 100 osservazioni (ordinate in senso crescente) e i corrispondenti valori della funzione di distribuzione empirica

Nelle due tabelle 5.9 e 5.10 troviamo i valori della funzione di distribuzione ipotizzata sotto  $H_0$ , vale a dire  $N(0, 1)$ , e i valori assoluti delle differenze tra le funzioni di distribuzione ( $\hat{F}_n - F(x)$ ).

Dai dati contenuti nell'ultima tabella siamo in grado di determinare il valore osservato della statistica  $D_n$  che in questo caso vale  $D_n^{oss} = 0.092$ . Fissato  $\alpha = 0.05$ , dalle tavole della distribuzione di Kolmogorov-Smirnov si ricava il valore critico corrispondente al quantile di ordine  $\alpha = 0.05$ ,  $D_{100;0.05} = 0.136$ , sicché essendo

$$D_n^{oss} = 0.092 < 0.136 = D_{100;0.05} \quad (5.170)$$

non si ha ragione di rifiutare  $H_0$  ovvero non vi è sufficiente evidenza nei dati da portare al rifiuto dell'*ipotesi di Normalità*.

Riportiamo, infine, in Figura 5.11 il grafico delle due distribuzioni, quella empirica e quella ipotizzata sotto  $H_0$ :

-3,68	-2,28	-1,97	-1,94	-1,69	-1,68	-1,60	-1,53	-1,52	-1,48
-1,41	-1,38	-1,28	-1,25	-1,23	-1,16	-1,11	-1,02	-1,00	-0,88
-0,85	-0,79	-0,75	-0,68	-0,47	-0,40	-0,37	-0,35	-0,35	-0,33
-0,32	-0,26	-0,26	-0,25	-0,24	-0,23	-0,23	-0,19	-0,19	-0,17
-0,17	-0,17	-0,16	-0,13	-0,13	-0,12	-0,09	-0,08	-0,06	-0,06
-0,04	-0,04	-0,04	-0,03	-0,01	0,01	0,02	0,07	0,09	0,13
0,14	0,15	0,19	0,20	0,21	0,22	0,25	0,28	0,29	0,30
0,32	0,41	0,47	0,50	0,52	0,56	0,58	0,60	0,62	0,63
0,65	0,67	0,71	0,74	0,76	0,78	0,80	0,80	0,85	0,89
1,06	1,15	1,29	1,30	1,32	1,92	2,18	2,29	2,40	3,08

**Figura 5.7:** Valori delle  $n = 100$  osservazioni (ordinate in senso crescente)

0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090	0,100
0,110	0,120	0,130	0,140	0,150	0,160	0,170	0,180	0,190	0,200
0,210	0,220	0,230	0,240	0,250	0,260	0,270	0,280	0,290	0,300
0,310	0,320	0,330	0,340	0,350	0,360	0,370	0,380	0,390	0,400
0,410	0,420	0,430	0,440	0,450	0,460	0,470	0,480	0,490	0,500
0,510	0,520	0,530	0,540	0,550	0,560	0,570	0,580	0,590	0,600
0,610	0,620	0,630	0,640	0,650	0,660	0,670	0,680	0,690	0,700
0,710	0,720	0,730	0,740	0,750	0,760	0,770	0,780	0,790	0,800
0,810	0,820	0,830	0,840	0,850	0,860	0,870	0,880	0,890	0,900
0,910	0,920	0,930	0,940	0,950	0,960	0,970	0,980	0,990	1,000

**Figura 5.8:** Valori di  $\hat{F}_n(x)$  per le  $n = 100$  osservazioni

0,000	0,011	0,024	0,026	0,046	0,046	0,055	0,063	0,064	0,069
0,079	0,084	0,100	0,106	0,109	0,123	0,133	0,154	0,159	0,189
0,198	0,215	0,227	0,248	0,319	0,345	0,356	0,363	0,363	0,371
0,374	0,397	0,397	0,401	0,405	0,409	0,409	0,425	0,425	0,433
0,433	0,433	0,436	0,448	0,448	0,452	0,464	0,468	0,476	0,476
0,484	0,484	0,484	0,488	0,496	0,504	0,508	0,528	0,536	0,552
0,556	0,560	0,575	0,579	0,583	0,587	0,599	0,610	0,614	0,618
0,626	0,659	0,681	0,691	0,698	0,712	0,719	0,726	0,732	0,736
0,742	0,749	0,761	0,770	0,776	0,782	0,788	0,788	0,802	0,813
0,855	0,875	0,901	0,903	0,907	0,973	0,985	0,989	0,992	0,999

**Figura 5.9:** Valori di  $\Phi_Z(x)$ 

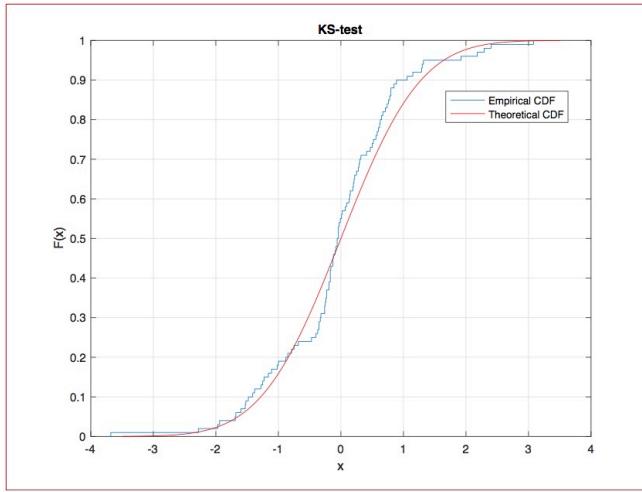
0,010	0,009	0,006	0,014	0,004	0,014	0,015	0,017	0,026	0,031
0,031	0,036	0,030	0,034	0,041	0,037	0,037	0,026	0,031	0,011
0,012	0,005	0,003	-0,008	-0,069	-0,085	-0,086	-0,083	-0,073	-0,071
-0,064	-0,077	-0,067	-0,061	-0,055	-0,049	-0,039	-0,045	-0,035	-0,033
-0,023	-0,013	-0,006	-0,008	0,002	0,008	0,006	0,012	0,014	0,024
0,026	0,036	0,046	0,052	0,054	0,056	0,062	0,052	0,054	0,048
0,054	0,060	0,055	0,061	0,067	0,073	0,071	0,070	0,076	0,082
0,084	0,061	0,049	0,049	0,052	0,048	0,051	0,054	0,058	0,064
0,068	0,071	0,069	0,070	0,074	0,078	0,082	0,092	0,088	0,087
0,055	0,045	0,029	0,037	0,043	-0,013	-0,015	-0,009	-0,002	0,001

**Figura 5.10:** Valori di  $D_n$ 

### 5.5.0.2 Test di Kolmogorov-Smirnov per due campioni

Supponiamo di avere due campioni casuali, il primo  $(X_1, X_2, \dots, X_{n_1})$  da una distribuzione  $F(x)$  e il secondo  $(Y_1, Y_2, \dots, Y_{n_2})$  da una distribuzione  $G(x)$ . Vogliamo testare il seguente sistema di ipotesi:

$$\begin{cases} H_0 : F = G \\ vs. \\ H_1 : F \neq G \end{cases} \quad (5.171)$$



**Figura 5.11:** Grafico di  $\hat{F}_n(x)$  (in blu) e della funzione di distribuzione  $N(0, 1)$  (in bordeaux)

Se  $\hat{F}_{n_1}$  e  $\hat{G}_{n_2}$  sono ripetutivamente le distribuzioni empiriche dei due campioni casuali poc'anzi considerati, la statistica

$$D_{n_1, n_2} = \sup_{x \in \mathbb{R}} |\hat{F}_{n_1} - \hat{G}_{n_2}| \quad (5.172)$$

tenuto conto del seguente teorema (di cui non diamo dimostrazione), permette di costruire un test per la verifica del sistema di ipotesi 5.171 in maniera analoga a quanto fatto nel caso di un solo campione

**Teorema 5.5.3.** Siano  $(X_1, X_2, \dots, X_{n_1})$  e  $(Y_1, Y_2, \dots, Y_{n_2})$   $n$ -ple di v.c. indipendenti e identicamente distribuite rispettivamente secondo  $F$  e  $G$  e siano  $\hat{F}_{n_1}$  e  $\hat{G}_{n_2}$  le distribuzioni empiriche corrispondenti ai due campioni. Sia Inoltre

$$D_{n_1, n_2} = \sup_{x \in \mathbb{R}} |\hat{F}_{n_1} - \hat{G}_{n_2}| \quad (5.173)$$

Allora,

$$\lim_{n_1, n_2 \rightarrow \infty} P \left( \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} D_{n_1, n_2} \leq t \right) = Q(t) \quad (5.174)$$

dove

$$Q(t) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 t^2} \quad (5.175)$$

*Dimostrazione.* Omessa. □

**Esempio 5.5.2.** Consideriamo le determinazioni di due campioni rispettivamente di ampiezza  $n_1 = 10$  e  $n_2 = 8$  riportate nella tabella 5.12 e verifichiamo il sistema di ipotesi 5.171 in merito alle distribuzioni che le hanno generati.

Costruiamo le corrispondenti distribuzioni empiriche che riportiamo nella prima e nella seconda riga della tabella 5.13 e riportiamo le loro differenze in valore assoluto nella terza riga della medesima tabella:

1	2	3	4	5	6	7	8	9	10
1,2	1,4	1,9	3,7	4,4	4,8	9,7	17,3	21,1	28,4
5,6	6,5	6,6	6,9	9,2	10,4	10,6	19,3		

**Figura 5.12:** Determinazioni dei due campioni casuali provenienti, rispettivamente, dalle distribuzioni  $F$  e  $G$

0,100	0,200	0,300	0,400	0,500	0,600	0,600	0,600	0,600	0,600	0,700	0,700	0,700	0,800	0,800	0,900	1,000	
0,000	0,000	0,000	0,000	0,000	0,000	0,125	0,250	0,375	0,500	0,625	0,625	0,750	0,875	0,875	1,000	1,000	
0,100	0,200	0,300	0,400	0,500	0,600	0,475	0,350	0,225	0,100	0,025	0,075	0,050	0,175	0,075	0,200	0,100	0,000

**Figura 5.13:**  $\hat{F}_{n_1}(x)$ ,  $\hat{G}_{n_2}(y)$  e  $|\hat{F}_{n_1}(x) - \hat{G}_{n_2}(y)|$

Fissato  $\alpha = 0.05$ , dalle tavole della distribuzione di Kolmogorov-Smirnov a due campioni si ottiene il valore critico  $D_{10,8;0.05} = 48/80 = 0.6$  mentre il valore osservato della statistica test è  $D = \sqrt{\frac{10-8}{10+8}} D_{10,8} = 1.265$  sicché, essendo  $D = 1.265 > 0.6 = D_{10,8;0.05}$ , si ha ragione di *rifiutare* l'ipotesi nulla  $H_0 : F = G$ .

## 5.6 *p*-value

In molte delle applicazioni è richiesto di verificare l'aderenza di una certa ipotesi all'informazione disponibile o, in altre parole, la sua credibilità alla luce dei dati, spesso si ricorre a ciò che è conosciuto come *p*-value (o *livello di significatività osservato*); esso fornisce una misura di sintesi probabilistica che può essere utilizzata per trarre conclusioni sulla presenza nei dati di *evidenza contro l'ipotesi nulla*.

L'approccio ai test basato sul *p*-value si affianca a quello più prettamente *statistico* di cui ci siamo occupati nelle pagine precedenti ed è spesso preferito da molti ricercatori di discipline in ambito bio-medico e economico-sociale. Perciò, vale la pena conoscerlo.

Cominciamo col dare la definizione di *p*-value.

**Definizione 5.6.1** (*p*-value). Diremo *p*-value (e lo indicheremo con  $\alpha^*$ ) la probabilità di ottenere *sotto*  $H_0$  un valore della statistica test *più estremo* di quello effettivamente osservato.

In altre parole, il *p*-value è la probabilità di rifiutare  $H_0$  quando è vera *sulla base dell'informazione fornita dai dati* o ancora *una misura del sostegno* fornito dai dati all'ipotesi nulla  $H_0$ . Pertanto, è ragionevole *rifiutare  $H_0$  quando il valore del *p*-value è piccolo* e per avere un termine di paragone su che cosa intendiamo per *piccolo*, prendiamo a riferimento il valore del livello di significatività  $\alpha$  fissato che ora chiameremo *livello di significatività nominale*.

Allora, considerato un sistema di ipotesi del tipo,

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ vs. \\ H_1 : \theta \in \Theta_1 \end{cases} \quad \text{con } \Theta = \Theta_0 \cup \Theta_1 \text{ e } \Theta_0 \cap \Theta_1 = \emptyset \quad (5.176)$$

in merito a  $H_0$ , arriviamo a formulare la seguente regola di decisione:

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } \alpha^* < \alpha \quad (5.177)$$

con  $\alpha$ , come ricordavamo poc'anzi, *livello di significatività nominale*, valore fissato dal ricercatore e ritenuto accettabile come probabilità di compiere un errore di I° tipo.

Il *p-value*, al pari della *forma* della regione critica del test, dipende dal tipo di *ipotesi alternativa* considerata (*unilaterale* o *bilaterale*) e quindi, nel calcolarlo, bisognerà tener conto di ciò.

A scopo illustrativo riportiamo qui sotto un esempio riguardante il calcolo del *p-value* per sistemi di ipotesi sulla media  $\mu$  di una qualsiasi distribuzione; comunque, quanto detto vale per sistemi di ipotesi su qualsiasi altro aspetto di interesse della distribuzione (varianza, proporzione, ecc.).

**Esempio 5.6.1.** a) Si vuole verificare

$$H_0 : \mu = 10 \text{ vs. } H_1 : \mu > 10 \quad (5.178)$$

sulla base della determinazione di un campione casuale di  $n = 50$  osservazioni su cui si è osservato  $\bar{x}_n = 10.6$  e  $s_n^2 = 4$ . In virtù del TLC che permette di approssimare, sotto condizioni non troppo restrittive (i.i.d. e finitezza dei primi due momenti), la distribuzione di una somma di v.c. indipendenti e identicamente distribuite con una distribuzione Normale di opportuni parametri ( $\mu = \bar{X}_n$ ,  $\sigma^2 = s_n^2$ ), sotto  $H_0 : \mu = 10$ , si ha

$$\begin{aligned} \alpha^* &= P(\bar{X}_n > 10.6 \mid \mu = 10) \\ &= P\left(\frac{\bar{X}_n - 10}{2/\sqrt{50}} > \frac{10.6 - 10}{2/\sqrt{50}}\right) \simeq \Phi_Z(2.12) = 0.0170 \end{aligned} \quad (5.179)$$

Ora, il valore ottenuto di *p-value* è *piccolo* ovvero, stando all'informazione contenuta nei dati (e sintetizzata nella media e nella deviazione standard campionarie) vi è una *bassa probabilità* di osservare un valore della media campionaria valore pari a 10.6 quando la media della popolazione è ipotizzata essere uguale a 10; peraltro,  $\alpha^* < \alpha = 0.05$ , livello di significatività usualmente adottato per default sicché, con buona ragione, possiamo **rifiutare**  $H_0 : \mu = 10$  **a favore di**  $H_1 : \mu > 10$ .

Possiamo anche avere una rappresentazione grafica di quanto detto poc'anzi con il *p-value* restituito dall'area ombreggiata in Figura 5.14:

Laddove avessimo  $H_0 : \mu = 10$  vs.  $H_1 : \mu < 10$  dovremmo "capovolgere" la Figura 5.14, spostando l'area ombreggiata a sinistra del valore del parametro ipotizzato sotto  $H_0$ .

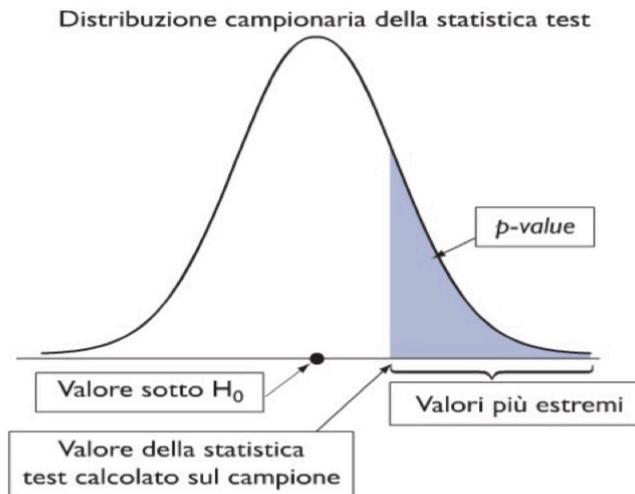
b) Con i medesimi dati a disposizione si vuole ora verificare il sistema di ipotesi

$$H_0 : \mu = 10 \text{ vs. } H_1 : \mu \neq 10 \quad (5.180)$$

Come abbiamo già visto, la regione di rifiuto di  $H_0$  in questo caso è anch'essa *bilatera* al pari dell'ipotesi  $H_1$  e pari a

$$C_\alpha = (-\infty, c_1) \cup (c_2, +\infty) \quad (5.181)$$

e la *distanza* in valore assoluto di  $\bar{x}_n$  dall'ipotesi nulla, meglio dal valore di  $\mu$  ipotizzato sotto  $H_0$ , è  $|\bar{x}_n - 10| = |10.6 - 10| = 0.6$ .



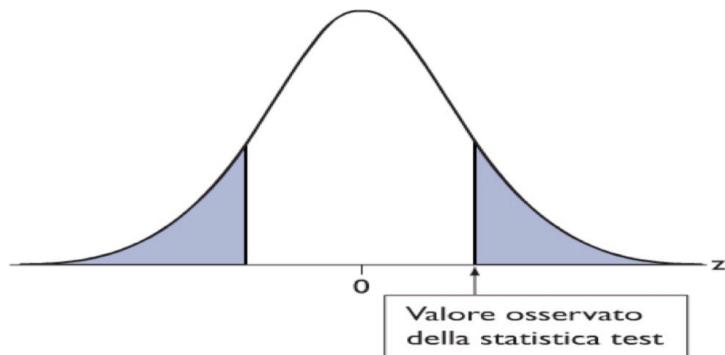
**Figura 5.14:**  $p$ -value per il sistema di ipotesi  $H_0 : \mu = \mu_0$  ve  $H_1 : \mu > \mu_0$

Il  $p$ -value, in questo caso, è

$$\begin{aligned}
 \alpha^* &= P((\bar{X}_n - 10) < -0.6) + P((\bar{X}_n - 10) > 0.6) \\
 &= 2P((\bar{X}_n - 10) < -0.6) \\
 &= 2P\left(\frac{(\bar{X}_n - 10)}{2/\sqrt{50}} < -\frac{0.6}{2/\sqrt{50}}\right) \\
 &\simeq \Phi_Z(-2.12) = 2 * 0.017 = 0.034
 \end{aligned} \tag{5.182}$$

Otteniamo ancora un valore di  $\alpha^*$  piccolo, e comunque più piccolo del valore nominale di  $\alpha$  che continuiamo a tenere fissato in 0.05; di conseguenza, anche in questo caso  $H_0$  è **rifiutata** a favore di  $H_1$ .

Come fatto al punto a), il  $p$ -value sarà rappresentato dall'area ombreggiata della seguente Figura 5.15:



**Figura 5.15:**  $p$ -value per il sistema di ipotesi  $H_0 : \mu = \mu_0$  ve  $H_1 : \mu \neq \mu_0$

# 6 Bootstrap

L'approccio tradizionale (parametrico) all'inferenza statistica poggia, come abbiamo visto, su *modelli idealizzati* (modelli statistici) e su *assunzioni* più o meno impegnative.

Spesso l'espressione matematica delle distribuzioni *esatte* associate a misure di precisione quali, per esempio, lo standard error risulta molto complicata, se non addirittura intrattabile, e pertanto si ricorre frequentemente a un largo uso della teoria asintotica; di conseguenza, non risulta disponibile per piccoli campioni dove il ricorso alla teoria asintotica è precluso.

## 6.1 Ricampionamento

Una soluzione al problema poc'anzi introdotto può essere fornita da procedure basate sul principio di ricampionamento. I metodi di ricampionamento (principalmente Jackknife e *bootstrap*) sono una delle più interessanti applicazioni inferenziali delle tecniche di simulazione stocastica e dei numeri (pseudo)-casuali. Essi si sono diffusi a partire dagli anni '60 del secolo scorso, in seguito del progresso delle tecnologie informatiche e dell'aumento di potenza computazionale e sono derivati concettualmente dal *metodo Monte Carlo*, ampiamente conosciuto in statistica matematica fin dagli anni '40 del secolo scorso. Ne riportiamo qui di seguito i passi fondamentali.

### Metodo Monte Carlo

- i) Supponiamo innanzitutto che  $F_\theta(x) = F(x; \theta)$  sia nota nella sua forma esatta o asintotica.
- ii) Generiamo  $B$  determinazioni di un campione casuale di ampiezza  $n$ ,  $(X_1, X_2, \dots, X_n)$  estratto da  $F_\theta(x)$   
$$(x_1, x_2, x_3, \dots, x_B)$$
- iii) Calcoliamo la stima dell'aspetto di interesse  $\eta = g(\theta)$ ,  $\hat{\eta}_n$  su ciascuna delle  $B$  determinazioni del campione casuale. iv) Ricaviamo la distribuzione empirica di  $\hat{\eta}_n^{(1)}, \hat{\eta}_n^{(2)}, \dots, \hat{\eta}_n^{(B)}$

$$\hat{F}_n(\hat{\eta}_n) = \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{(-\infty, \hat{\eta}_n)}(\hat{\eta}_n^{(b)}) \quad (6.1)$$

- v) L'equazione (6.1) fornisce una *stima* della *distribuzione campionaria* di  $\hat{\eta}_n$  che chiameremo *distribuzione Monte Carlo* di  $\hat{\eta}_n$ .

La *distribuzione Monte Carlo* di  $\hat{\eta}_n$  può essere usata per fare inferenza su  $\eta = g(\theta)$  o per esplorare e studiare le proprietà statistiche di  $\hat{\eta}_n$ ; o per confrontare  $\hat{\eta}_n$  con altri stimatori competitori nella stima di  $\eta = g(\theta)$ .

In buona sostanza, il *metodo Monte Carlo* è sostanzialmente basato su di una semplice *procedura di ricampionamento* dalla distribuzione  $F_\theta(x)$ .

Tratto comune di tutti i metodi di ricampionamento è quello di *ripetere semplici operazioni* un *numero elevato* di volte, generando numeri casuali da assegnare a variabili casuali o a campioni casuali. Essi richiedono maggior tempo macchina al crescere delle operazioni ripetute, sono molto semplici da implementare e, una volta implementati, sono automatici.

In statistica, esistono diversi ambiti di applicazione dei metodi di ricampionamento: *Jackknife*, *bootstrap*, *cross-validation*, test di permutazione: di questi considereremo in dettaglio il solo *bootstrap*.

## 6.2 Il bootstrap

Efron (1979) introdusse una tecnica statistica in grado di superare il problema legato alla non applicabilità dei metodi dell'inferenza tradizionale, a cui ha dato il nome di *bootstrap*. Questo nome deriva da quello del protagonista del romanzo *Adventures of Baron Münchhausen* di Rudolph E. Raspe, Karl Friedrich Hieronymus, Freiherr von Münchhausen (1720-1797), personaggio realmente esistito e noto per i suoi mirabolanti e inverosimili racconti tra cui, per restare in tema, quello in cui si salvò dal fango di una palude in cui era sprofondato con il proprio cavallo, afferrando i propri capelli e tirandosi su: atto in sè paradossale e tale deve anche essere sembrata a Efron la filosofia della tecnica statistica da lui inventata.



**Figura 6.1:** Il barone di Münchhausen nell'atto di (auto)trarsi fuori dai guai...

Di qui la frase *to pull oneself up by one's own bootstraps* ovvero ...*tirarsi su afferrando le lingue dei propri stivali* che vuol essenzialmente dire *riuscire a fare qualcosa facendo affidamento unicamente sulle proprie risorse*; in altre parole cavarsela da soli. Ed è proprio quello fa la tecnica *bootstrap* introdotta da Efron.

Il *bootstrap* è essenzialmente un metodo di ricampionamento *computer intensive* che risulta

- a) utile quando la distribuzione campionaria esatta (o asintotica) di una certa statistica di interesse è intrattabile o non disponibile
- b) largamente applicabile a molti contesti

### 6.2.1 La tecnica bootstrap

Supponiamo di avere a disposizione un campione casuale  $(X_1, X_2, \dots, X_n)$  da una distribuzione  $F_\theta(x)$  e di voler stimare il parametro  $\theta$  (o una qualche sua funzione  $\eta = g(\theta)$ ) tramite lo stimatore  $\hat{\theta}_n = T(X_1, X_2, \dots, X_n)$ .

Allo scopo di fare inferenza sul parametro  $\theta$  (o su un'altra caratteristica  $\eta = g(\theta)$ ) della distribuzione quale *standard error* o *distorsione* o altro ancora) dobbiamo disporre della distribuzione campionaria di  $\hat{\theta}_n$  (o di  $\hat{\eta}_n$ ) così da poter ricavare informazioni in merito all'accuratezza e alla precisione dello stimatore adottato, costruire intervalli di confidenza o regole di decisione per testare sistemi di ipotesi cui siamo interessati. L'intrattabilità della distribuzione campionaria di  $\hat{\theta}_n$  può rappresentare un serio ostacolo per i nostri interessi inferenziali su  $\theta$  o  $g(\theta)$ .

Ora,

- a) se la vera distribuzione  $F_\theta(x)$  è nota, possiamo applicare il *metodo Monte Carlo* per ricavare una *approssimazione* della distribuzione campionaria di  $\hat{\theta}_n$  e usarla per costruire le procedure inferenziali desiderate
- b) ma se così non è, ricampionare da  $F_\theta(x)$  è precluso, e la sola alternativa che rimane è quella di ricorrere alla tecnica *bootstrap*.

L'idea che sta alla base del *bootstrap* è che, in assenza di ogni altra informazione sulla distribuzione  $F(x; \theta)$  che ha generato i dati, la determinazione campionaria  $(x_1, x_2, \dots, x_n)$  contiene *tutta* l'informazione disponibile relativa alla distribuzione generante i dati e perciò *ricampionare (con riposizione)* da  $(x_1, x_2, \dots, x_n)$  è la miglior strategia per recuperare informazione su  $F(x; \theta)$  o su qualche suo aspetto interesse. Ciò implica il ricampionare dalla distribuzione empirica

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(X_i) \quad (6.2)$$

che può essere pensata come la distribuzione che mette *massa (probabilistica)* pari a  $1/n$  su ciascuna osservazione campionaria  $X_i$  e, laddove  $X_i$  ricorra più di una volta nel campione, la relativa massa risulterà essere un *multiplo* di  $1/n$ . Di conseguenza,  $\hat{F}_n(x)$  è una distribuzione discreta che insiste sullo spazio campionario costituito da  $(X_1, X_2, \dots, X_n)$ . Per ulteriori dettagli su  $\hat{F}_n(x)$  si vedani i paragrafi 2.1 e 5.5.

### 6.2.1.1 Distribuzione e inferenza bootstrap

Supponiamo di voler estrarre un campione casuale di ampiezza  $n$ ,  $(X_1^*, X_2^*, \dots, X_n^*)$ , dalla funzione di distribuzione empirica  $\hat{F}_n(x)$ . Stando a quanto appena detto,  $\hat{F}_n(x)$  colloca massa  $1/n$  su ciascun elemento  $X_i$ ; allora, possiamo descrivere il cuore della tecnica bootstrap alla base della generazione di  $(X_1^*, X_2^*, \dots, X_n^*)$  attraverso la seguente procedura (in due passi) di *ricampionamento (con riposizione)*.

#### Procedura di ricampionamento bootstrap

- i) in primo luogo, estraiamo gli indici  $i_1, i_2, \dots, i_n$ , indipendentemente l'uno dall'altro, dalla distribuzione Uniforme su  $\{1, 2, \dots, n\}$
- ii) successivamente, poniamo  $X_i^* = X_{i_k}$  ottenendo infine,  $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_n^*)$ .

In altre parole, ricampioniamo *con riposizione*  $(X_1^*, X_2^*, \dots, X_n^*)$  dall'originario campione casuale  $(X_1, X_2, \dots, X_n)$ . In questo modo riproduciamo il processo di generazione dei dati campionando dalla distribuzione emnpirica  $\hat{F}_n(x)$  che costituisce una buona *stima dell'incognita* distribuzione  $F_\theta(x)$ .

Ora, supponiamo di disporre di una determinazione  $(x_1, x_2, \dots, x_n)$  del campione causale  $(X_1, X_2, \dots, X_n)$  da una distribuzione  $F_\theta(x)$  e di essere interessati a fare inferenza sul parametro  $\theta$  (o su un'altra caratteristica  $\eta = g(\theta)$ ) della distribuzione quale *standard error* o *distorsione* o altro ancora); sia  $\hat{\theta}_n$  uno stimatore di  $\theta$  la cui distribuzione non è nota (è molto complicata da ricavare o non si conosce affatto). Quello che possiamo fare per averne una stima è ricorrere al *mondo bootstrap*; più specificatamente:

- a) ricorrendo alla **Procedura di ricampionamento bootstrap**, generiamo  $B$  campioni bootstrap indipendenti

$$\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_b^*, \dots, \mathbf{X}_B^*, \quad (6.3)$$

dove  $\mathbf{X}_b^* = (X_{b1}^*, X_{b2}^*, \dots, X_{bn}^*)$ ,  $b = 1, 2, \dots, B$ .

- b) generiamo le  $B$  repliche bootstrap di  $\hat{\theta}_n$  date da

$$\hat{\theta}_{nb}^* = T(\mathbf{X}_b^*), \quad b = 1, 2, \dots, B \quad (6.4)$$

- c) costruiamo infine la *distribuzione bootstrap* data dalla distribuzione empirica delle repliche bootstrap  $\hat{\theta}_{n1}^*, \hat{\theta}_{n2}^*, \dots, \hat{\theta}_{nB}^*$  e che indicheremo con

$$\hat{F}_n^*(w) = \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{(-\infty, w]}(\hat{\theta}_{nb}^*), \quad w \in \mathbb{R} \quad (6.5)$$

Quest'ultima fornisce naturalmente una *stima* della distribuzione campionaria di  $\hat{\theta}_n$  (che avevamo assunto essere *difficile da calcolare*).

Ottenuta la *distribuzione bootstrap*  $\hat{F}^*(\hat{\theta}_n)$ , possiamo costruire una serie di **procedure inferenziali bootstrap**, seguendo quanto detto nelle pagine precedenti. Infatti, grazie a questa, possiamo

- 1) ricavare immediatamente una *stima bootstrap* della deviazione standar (e quindi dello standard error) di  $\hat{\theta}_n^*$ , stimatore di  $\theta$ , cosa che inizialmente era preclusa dal fatto di non disporre della sua distribuzione campionaria. Ora, la *varianza bootstrap*

$$s_B^{*2} = \frac{1}{B} \sum_{b=1}^B \left[ \hat{\theta}_{nb}^* - \bar{\theta}_n^* \right]^2 \quad (6.6)$$

dove

$$\bar{\theta}_n^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_{nb}^* \quad (6.7)$$

fornisce immediatamente una *stima* della varianza campionaria di  $\hat{\theta}_n^*$  e la *stima bootstrap della deviazione standard* di  $\hat{\theta}_n^*$  è subito restituita dalla sua radice quadrata  $s_B^* = \sqrt{\frac{1}{B} \sum_{b=1}^B \left[ \hat{\theta}_{nb}^* - \bar{\theta}_n^* \right]^2}$ .

- 2) costruire un *intervallo di confidenza bootstrap* di livello  $(1 - \alpha)$ , noto come *percentile interval*, semplicemente ricavando i quantili  $\hat{\theta}_{n; \alpha/2}^*$  e  $\hat{\theta}_{n; (1-\alpha/2)}^*$  di ordine  $\alpha/2$  e  $(1 - \alpha/2)$  della distribuzione bootstrap e definendo

$$I_\theta^*(1 - \alpha) : \left[ \hat{\theta}_{n; \alpha/2}^*, \hat{\theta}_{n; (1-\alpha/2)}^* \right] \quad (6.8)$$

Talvolta si preferisce costruire anche il seguente intervallo di confidenza per  $\theta$ , noto come *Normal interval*

$$I_\theta^*(1 - \alpha) : \left[ \hat{\theta}_n - z_{1-\alpha/2} \sqrt{s_B^{*2}}, \hat{\theta}_n + z_{1-\alpha/2} \sqrt{s_B^{*2}} \right] \quad (6.9)$$

con  $z_{1-\alpha/2}$  quantile di ordine  $(1 - \alpha/2)$  della distribuzione Normale standard

- 3) individuare una *regola di decisione bootstrap* per un sistema di ipotesi del tipo

$$H_0 : \theta \leq \theta_0 \text{ vs. } H_1 : \theta > \theta_0 \quad (6.10)$$

Ricordando che il *p-value* va calcolato sotto  $H_0$ , dobbiamo rimodulare il campione originario in modo da tener conto di  $H_0$  (vale a dire, così come fosse stato estratto dalla distribuzione sotto  $H_0$ ).

Per esempio, supponiamo di essere interessati a ricavare il *p-value bootstrap* per il sistema di ipotesi sulla media della popolazione

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

Un modo immediato per rimodulare i dati in maniera da tener conto dell'ipotesi  $H_0 : \mu = \mu_0$ , consiste nel ridefinire ciascun elemento dell'originario campione nella maniera seguente

$$z_i = x_i - \bar{x}_n + \mu_0, \quad i = 1, 2, \dots, n$$

e poi campionare con riposizione da  $(z_1, z_2, \dots, z_n)$  e procedere nel solito modo per trovare le  $B$  repliche bootstrap  $\hat{\theta}_{nb}^*$  e la relativa distribuzione empirica  $\hat{F}_n^*(w) = \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{(-\infty, w]}(\hat{\theta}_{nb}^*)$ .

Il *p*-value *bootstrap* viene poi calcolato nella seguente maniera:

$$\alpha_b^* \simeq \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{[\hat{\theta}_n, +\infty)}(\hat{\theta}_{nb}^*) \quad (6.11)$$

Quindi, possiamo stabilire la seguente regola di decisione:

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } \alpha_b^* < \alpha \quad (6.12)$$

con  $\alpha$  livello di significatività *nominal*e fissato. Dal momento che stiamo lavorando con il *p*-value, siamo orientati a rifiutare  $H_0$  ogniqualvolta  $\alpha_b^*$  è *piccolo* ovvero i dati non forniscono supporto all'ipotesi nulla.

L'implementazione in R dei punti 1)-3) è estremamente facile. In appendice

## 6.3 Appendice

Riportiamo il codice sorgente (in un formato volutamente piuttosto grezzo ma, per converso, comprensibile) relativo alle diverse procedure inferenziali bootstrap che abbiamo presentato in questo capitolo.

# 7 Regressione lineare

Il modello di regressione lineare (semplice, multipla o multivariata), oltre a essere il *building block* per modelli più complessi e interessanti per applicazioni reali, costituisce un ambiente naturale per vedere in azione le procedure inferenziali fin qui studiate (stimatori puntuali e intervallari, test per la verifica di ipotesi).

## 7.1 Il modello

In termini generali, un modello cerca di *spiegare causalmente* una o più variabili ( $Y$ , nella notazione qui adottata) in termini di una o più altre variabili (indicate con  $x$ ); la sua costruzione si concretizza in tre fasi distinte

- specificazione** del modello: sulla base di teorie consolidate o evidenze sperimentali, si ipotizza la *forma della relazione causale* tra  $Y$  che chiamaremos *variabile risposta* (o anche variabile dipendente) e  $x$  cui daremo il nome di *regressore* (o variabile esplicativa o indipendente).

Diremo *semplice* il modello che considera una sola variabile risposta  $Y$  e un solo regressore  $x$  (ossia, spiega  $Y$  in termini della sola  $x$ ) e *lineare* se esso è *lineare nei parametri* in esso coinvolti come vedremo a breve.

- stima** del modello a partire dai dati disponibili su  $Y$  e  $x$
- verifica della bontà** del modello (significatività del modello)

Formalmente, dato un insieme di osservazioni  $\{(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)\}$  relative alla coppia  $(x, Y)$ , il *modello classico di regressione lineare semplice* è così definito:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (7.1)$$

La componente  $(\beta_0 + \beta_1 x_i)$  è detta *componente deterministica*, mentre  $\varepsilon_i$  si dice essere la *componente stocastica di errore*; tramite questa componente rappresentiamo tutti i possibili errori di misurazione, la casualità intrinseca nel fenomeno che stiamo cercando di modellare con la componente deterministica e, in ultima analisi, tutto quanto non spiegato da quest'ultima.

Nelle pagine che seguono ci soffermeremo principalmente sul modello di regressione lineare semplice; dedicheremo poi qualche cenno alla sua estensione al caso in cui vi sia più di un regressore ovvero al modello di regressione lineare multipla.

### 7.1.1 Le ipotesi "classiche"

Alla base del modello (7.1) vi è un set di ipotesi, noto come *ipotesi classiche*, che ora brevemente richiamiamo:

1. *assenza di errore sistematico*,  $\mathbb{E}(\varepsilon_i) = 0$  (un eventuale errore sistematico verrebbe incorporato nell'intercetta del modello)
2. gli *errori* sono *incorrrelati e omoschedastici* (ovvero che hanno tutti la stessa varianza), sicché:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & \text{se } i \neq j \\ \sigma^2 & \text{se } i = j \end{cases} \quad (7.2)$$

3. il *regressore*  $x_i$  è *incorrrelato* con  $\varepsilon_i$ , vale a dire  $\text{Cov}(x_i, \varepsilon_i) = 0$  ossia, le  $x_i$  non sono variabili aleatorie bensì quantità costanti e note; di fatto  $x$  rappresenta la *variabile di controllo*, vale a dire la variabile che è sotto il controllo del ricercatore che conosce o ne può determinare i diversi valori; è pacifico che, se così non fosse, non potremmo cercare di spiegare qualcosa ( $Y$ ) alla luce di qualcos'altro ( $x$ ) che non conosciamo o che non controlliamo
4. La dispersione, ovvero la *varianza* delle  $x_i$  deve rimanere *finita* e al di sotto di un valore limite, in modo da evitare di avere rette con pendenza nulla:  $0 < \text{Var}(x_i) \leq v$ .
5. La componente stocastica del modello (o *errore*) segue una *distribuzione Normale* di media zero e varianza  $\sigma^2 > 0$ , ovvero  $\varepsilon_i \sim N(0, \sigma^2)$ . Tale ipotesi è certamente rispettata, per  $n$  sufficientemente grande, per il Teorema del Limite Centrale.

### 7.1.2 Il metodo di stima dei minimi quadrati

Vogliamo ora individuare un metodo di *stima dei parametri*  $\beta_0$  e  $\beta_1$  del modello (7.1) *sotto le ipotesi* poc'anzi assunte. Cominciamo con scrivere la (7.1) in una maniera leggermente diversa

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i), \quad i = 1, 2, \dots, n \quad (7.3)$$

e riconoscere in questa quantità l'*errore* che si commette nello "stimare"  $Y_i$  tramite quella che abbiamo assunto essere la componente deterministica  $\beta_0 + \beta_1 x_i$  del modello e che incorpora la legge esplicativa che riteniamo leggi  $Y_i$  a  $x_i$ . Ed è del tutto naturale voler individuare, assumendo l'additività degli errori, tra tutti gli ammissibili valori di  $\beta_0$  e di  $\beta_1$  proprio quei valori che rendono *minima* la quantità

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n [(Y_i - (\beta_0 + \beta_1 x_i))^2] \quad (7.4)$$

che restituisce una *misura complessiva* dell'errore in cui si incorre "stimando" ciascun  $Y_i$  con  $\beta_0 + \beta_1 x_i$ ,  $i = 1, 2, \dots, n$ .

Si tratta dunque di risolvere il seguente problema di minimo:

$$\min_{\beta_0, \beta_1} Q(\beta_0, \beta_1) = \min_{\beta_0, \beta_1} \sum_{i=1}^n [(Y_i - (\beta_0 + \beta_1 x_i))^2] \quad (7.5)$$

Come siamo abituati a fare, imponiamo dunque che

$$\begin{cases} \frac{\partial}{\partial \beta_0} Q(\beta_0, \beta_1) = \sum_{i=1}^n [(Y_i - (\beta_0 - \beta_1 x_i))] = 0 \\ \frac{\partial}{\partial \beta_1} Q(\beta_0, \beta_1) = \sum_{i=1}^n [(Y_i - (\beta_0 - \beta_1 x_i))] x_i = 0 \end{cases} \quad (7.6)$$

la cui soluzione rispetto  $\beta_0$  e  $\beta_1$ , con qualche semplice passaggio algebrico, porta a

$$\begin{cases} b_0 = \bar{Y}_n - b_1 \bar{x}_n \\ b_1 = \frac{\text{Cov}(x, Y)}{\text{Var}(x)} \end{cases} \quad (7.7)$$

dove

$$\frac{\text{Cov}(x, Y)}{\text{Var}(x)} = \frac{\frac{1}{n-1} (\sum_{i=1}^n x_i Y_i - n \bar{x}_n \bar{Y}_n)}{\frac{1}{n-1} (\sum_{i=1}^n x_i^2 - n \bar{x}_n^2)} = \frac{(\sum_{i=1}^n x_i Y_i - n \bar{x}_n \bar{Y}_n)}{(\sum_{i=1}^n x_i^2 - n \bar{x}_n^2)} \quad (7.8)$$

mentre  $b_0$  e  $b_1$  stimano rispettivamente  $\beta_0$  e  $\beta_1$ .

Tale metodo di stima è noto come **metodo dei minimi quadrati**,  $b_0$  e  $b_1$  sono le *stimatori a minimi quadrati* dei parametri  $\beta_0$  e  $\beta_1$  del modello (7.1) e la (7.5) rende ragione di questo nome.

La *stima a minimi quadrati del modello* (7.1) è dunque data da

$$\hat{Y}_i = b_0 + b_1 x_i, \quad i = 1, 2, \dots, n. \quad (7.9)$$

e costituisce un valido strumento in ambito di *previsione* del fenomeno studiato.

Rimane il problema di *quantificare o misurare la bontà di adattamento* del modello ai dati. Questo può essere realizzato osservando che la *devianza totale* di  $Y$  (o *somma totale dei quadrati*, SSTO secondo l'acronimo in lingua inglese) è data da

$$\text{SSTO} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = (n-1) \text{Var}(Y) \quad (7.10)$$

e che quest'ultima può essere così decomposta (vedi Figura 7.1)

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \text{SSR} + \text{SSE} \end{aligned} \quad (7.11)$$

dove  $\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$  rappresenta la devianza *spiegata* dal modello e  $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  la devianza *residua*, vale a dire, *non spiegata* dal modello. Ora, dividendo ambo i membri della (7.11) per SSTO =  $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$  si ha

$$\begin{aligned} 1 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} + \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} \\ &= \frac{\text{SSR}}{\text{SSTO}} + \frac{\text{SSE}}{\text{SSTO}} \end{aligned} \quad (7.12)$$

individuando in  $\text{SSR}/\text{SSTO}$  la *quota* di devianza totale SSTO *spiegata* dal modello e in  $\text{SSE}/\text{SSTO}$  la *quota* di devianza totale SSTO *residua*, vale a dire, *non spiegata* dal modello.

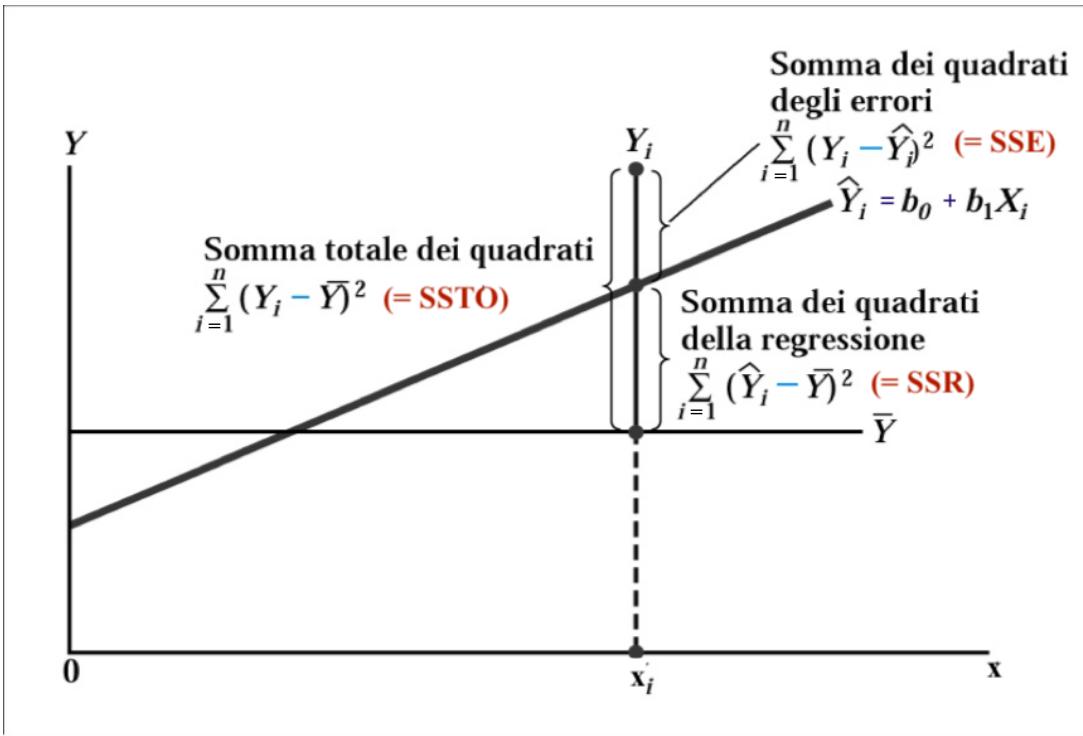


Figura 7.1: Scomposizione della devianza totale e bontà di adattamento

La quantità  $SSR/SSTO$  rappresenta il *guadagno relativo di regressione* che si ha nello stimare  $Y_i$  tramite  $\hat{Y}_i = b_0 + b_1 x_i$  piuttosto che tramite la media  $\bar{Y}_n$ , come si può facilmente osservare dalla Figura 7.1.

Ed è allora naturale pensare di costruire un indice di bontà di adattamento del modello ai dati (o *goodness-of-fit*) ricorrendo alla quantità

$$0 \leq R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \leq 1 \quad (7.13)$$

noto come *coefficiente di determinazione*. Ovviamente più  $R^2$  è vicino a 1 e meglio il modello si adatta ai dati. E' immediato osservare che un elevato valore di  $R^2$  equivale a un valore di  $SSR$  prossimo a  $SSTO$  e quindi alla capacità del modello di spiegare molta parte della variabilità della variabile dipendente  $Y$ .

Come abbiamo visto da (7.13) i *residui* di regressione  $e_i = (Y_i - \hat{Y}_i)$ ,  $i = 1, 2, \dots, n$  giocano un ruolo fondamentale nel *misurare* il grado di adattamento del modello ai dati, o secondo il temine inglese oramai entrato nell'uso comune, il *goodness-of-fit*. I residui  $e_i$ ,  $i = 1, 2, \dots, n$  godono di alcune importanti proprietà, utili nella comprensione formale delle cose che diremo nel seguito.

**Teorema 7.1.1** (Proprietà dei residui di regressione). Siano

$$e_i = (Y_i - \hat{Y}_i) = Y_i - (b_0 + b_1 x_i), \quad i = 1, 2, \dots, n \quad (7.14)$$

i *residui di regressione*. Allora valgono le seguenti proprietà:

- a)  $\sum_{i=1}^n e_i = 0$
- b)  $\sum_{i=1}^n x_i e_i = 0$
- c)  $\sum_{i=1}^n \hat{Y}_i e_i = 0$
- d)  $\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n) e_i = 0$

*Dimostrazione.* a) e b) seguono immediatamente dalle equazioni di stima

$$\begin{cases} \frac{\partial}{\partial \beta_0} Q(\beta_0, \beta_1) = \sum_{i=1}^n [(Y_i - (\beta_0 - \beta_1 x_i))] = 0 \\ \frac{\partial}{\partial \beta_1} Q(\beta_0, \beta_1) = \sum_{i=1}^n [(Y_i - (\beta_0 - \beta_1 x_i))] x_i = 0 \end{cases} \quad (7.15)$$

Infatti, sostituendo  $b_0$  e  $b_1$  al posto di  $\beta_0$  e  $\beta_1$  si ottiene

$$\sum_{i=1}^n (Y_i - \hat{Y}_i) = \sum_{i=1}^n (Y_i - (b_0 + b_1 x_i)) = \sum_{i=1}^n e_i = 0 \quad (7.16)$$

e

$$\sum_{i=1}^n (Y_i - \hat{Y}_i) x_i = \sum_{i=1}^n (Y_i - (b_0 + b_1 x_i)) x_i = \sum_{i=1}^n e_i x_i = 0 \quad (7.17)$$

La proprietà c) segue immediatamente dalla dimostrazione delle proprietà a) e b) poichè

$$\sum_{i=1}^n (Y_i - \hat{Y}_i) \hat{Y}_i = \sum_{i=1}^n (Y_i - \hat{Y}_i)(b_0 + b_1 x_i) = b_0 \sum_{i=1}^n (Y_i - \hat{Y}_i) + b_1 \sum_{i=1}^n (Y_i - \hat{Y}_i) x_i = 0 \quad (7.18)$$

e dunque il primo addendo è nullo per a) e il secondo lo è anch'esso per b) Infine la dimostrazione della proprietà d) segue immediatamente dalle dimostrazioni delle precedenti proprietà a) e c)

$$\sum_{i=1}^n (Y_i - \hat{Y}_i) (\hat{Y}_i - \bar{Y}_n) = \sum_{i=1}^n (Y_i - \hat{Y}_i) \hat{Y}_i - \bar{Y}_n \sum_{i=1}^n (Y_i - \hat{Y}_i) = 0 \quad (7.19)$$

□

## 7.2 Inferenza sul modello di regressione

Per fare inferenza sui parametri  $\beta_0$  e  $\beta_1$  del modello di regressione, vale a dire costruire intervalli di confidenza o verificare opportune ipotesi su di essi, bisogna disporre delle *distribuzioni campionarie* dei loro stimatori a minimi quadrati,  $b_0$  e  $b_1$ .

Il teorema che segue restituisce proprio queste ultime, sotto le ipotesi (*ipotesi classiche*) poste alla base del modello (7.1) stesso.

**Teorema 7.2.1.** [Distribuzione degli stimatori  $b_0$  e  $b_1$  di  $\beta_0$  e  $\beta_1$ ]

Siano  $b_0$  e  $b_1$  gli stimatori a minimi quadrati dei parametri  $\beta_0$  e  $\beta_1$  del modello di regressione lineare semplice (7.1) per il quale assumiamo valere le *ipotesi classiche*. Allora,

$$b_0 \sim N \left( \beta_0, \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2 + n\bar{x}_n^2}{n \sum_{i=1}^n (x_i - \bar{x}_n)^2} \cdot \sigma^2 \right)$$

e

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right)$$

*Dimostrazione.* Sia  $\{(x_1, Y_1), \dots, (x_n, Y_n)\}$  un insieme di  $n$  osservazioni relative alla coppia  $(x, Y)$  e assumiamo valga la seguente relazione, per  $i = 1, 2, \dots, n$ ,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (7.20)$$

in cui riconosciamo la componente deterministica  $\beta_0 + \beta_1 x_i$  che formalizza la relazione casuale che lega  $Y_i$  a  $x_i$  e quella stocastica (o di errore)  $\epsilon_i$  che riassume tutto quello che, relativamente a  $Y_i$ , sfugge alla componente deterministica (ovvero, non spiega). Gli stimatori a minimi quadrati dei parametri  $\beta_1$  e  $\beta_0$  del modello di regressione lineare (7.20) sono dati da

$$\begin{aligned} b_1 &\stackrel{d}{=} \frac{Cov(x, Y)}{Var(x)} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ &= \sum_{i=1}^n \frac{(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} Y_i \\ &= \sum_{i=1}^n u_i Y_i \end{aligned} \quad (7.21)$$

e da

$$\begin{aligned} b_0 &\stackrel{d}{=} \bar{Y}_n - \hat{\beta}_1 \bar{x}_n = \bar{Y}_n - \frac{Cov(x, Y)}{Var(x)} \bar{x}_n \\ &= \frac{1}{n} \sum_{i=1}^n Y_i - \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \bar{x}_n \\ &= \frac{1}{n} \sum_{i=1}^n Y_i - \sum_{i=1}^n \frac{(x_i - \bar{x}_n) Y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \bar{x}_n \\ &= \sum_{i=1}^n \left( \frac{1}{n} - \frac{(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \bar{x}_n \right) Y_i \\ &= \sum_{i=1}^n v_i Y_i \end{aligned} \quad (7.22)$$

Per le ipotesi sulla struttura degli errori  $\epsilon_i$ ,  $i = 1, 2, \dots, n$  (vedi ipotesi classiche), alla luce del fatto che  $\epsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, 2, \dots, n$  e in virtù della forma funzionale del modello di regressione semplice (7.20), si ottiene facilmente

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, 2, \dots, n$$

con  $(Y_1, Y_2, \dots, Y_n)$  variabili casuali a loro volta incorrelate (in questo caso, anche indipendenti).

Come si osserva in (7.21) e (7.22), gli stimatori a minimi quadrati di  $\beta_0$  e  $\beta_1$ ,  $b_0$  e  $b_1$ , risultano essere combinazioni lineari, rispettivamente con pesi  $v_i$  e  $u_i$ , delle v.c.

$Y_i$ ,  $i = 1, 2, \dots, n$ , incorrelate e normalmente distribuite; di conseguenza, per un noto teorema (vedi Teorema 1 - distribuzione Normale multivariata), anche gli estimatori  $b_0$  e  $b_1$  avranno distribuzione Normale di opportuna media e varianza che ora ricaveremo. Prima di procedere, però, vale la pena osservare che:

$$\sum_{i=1}^n u_i = \sum_{i=1}^n \frac{(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \sum_{i=1}^n (x_i - \bar{x}_n) = 0$$

e

$$\sum_{i=1}^n u_i x_i = \sum_{i=1}^n \frac{(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} x_i = \frac{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = 1.$$

Sfruttando questi due ultimi risultati, tenendo presente (7.21) e (7.22), si ha:

$$\mathbb{E}(b_1) = \mathbb{E}\left(\sum_{i=1}^n u_i Y_i\right) = \sum_{i=1}^n u_i \mathbb{E}(Y_i) = \sum_{i=1}^n u_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n u_i + \beta_1 \sum_{i=1}^n u_i x_i = \beta_1$$

e

$$\begin{aligned} \text{Var}(b_1) &= \text{Var}\left(\sum_{i=1}^n u_i Y_i\right) = \left(\sum_{i=1}^n u_i^2 \text{Var}(Y_i)\right) = \sigma^2 \sum_{i=1}^n \left[\frac{(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right]^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \end{aligned}$$

mentre

$$\mathbb{E}(b_0) = \mathbb{E}(\bar{Y}_n - b_1 \bar{x}_n) = \mathbb{E}(\bar{Y}_n) - \mathbb{E}(b_1 \bar{x}_n) = \mathbb{E}(\bar{Y}_n) - \bar{x}_n \mathbb{E}(b_1) = (\beta_0 + \beta_1 \bar{x}_n) - \beta_1 \bar{x}_n = \beta_0$$

e

$$\begin{aligned} \text{Var}(b_0) &= \text{Var}\left(\sum_{i=1}^n v_i Y_i\right) = \text{Var}\left(\sum_{i=1}^n \left(\frac{1}{n} - u_i \bar{x}_n\right) Y_i\right) \\ &= \sum_{i=1}^n \left(\frac{1}{n} - u_i \bar{x}_n\right)^2 \text{Var}(Y_i) \\ &= \sum_{i=1}^n \left(\frac{1}{n^2} - 2 u_i \bar{x}_n + u_i^2 \bar{x}_n^2\right) \sigma^2 \\ &= \sigma^2 \left(\frac{1}{n} - 2 \bar{x}_n \sum_{i=1}^n u_i + \bar{x}_n^2 \sum_{i=1}^n u_i^2\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right) \\ &= \sigma^2 \left(\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2 + n \bar{x}_n^2}{n \sum_{i=1}^n (x_i - \bar{x}_n)^2}\right). \end{aligned}$$

Ma allora, in conclusione, otteniamo

$$b_0 \sim N\left(\beta_0, \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2 + n \bar{x}_n^2}{n \sum_{i=1}^n (x_i - \bar{x}_n)^2} \cdot \sigma^2\right)$$

e

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right)$$

che è quanto dovevamo dimostrare.  $\square$

Sempre in merito agli stimatori a minimi quadrati  $b_0$  e  $b_1$  dei parametri  $\beta_0$  e  $\beta_1$  possiamo dire che essi

- i) risultano essere i *migliori* stimatori *lineari* e *non distorti* per i parametri del modello di regressione lineare semplice  $\beta_0$  e  $\beta_1$  dove con l'aggettivo migliore si intende quelli *più affidabili*, o tecnicamente i *più efficienti* in virtù di un noto teorema (*Teorema di Gauss-Markov*).

Avremo modo di tornare in seguito su questo argomento.

- ii) *senza l'assunzione di Normalità* sull'errore, sarebbero solo *asintoticamente* normalmente distribuiti.

### 7.2.1 Test di ipotesi e intervalli di confidenza per i parametri del modello di regressione

Supponiamo a questo punto di voler verificare se le nostre congetture in merito al *coefficiente angolare*  $\beta_1$  della retta di regressione siano fondate (possiamo procedere in modo analogo per verificare un'ipotesi sull'*intercetta*  $\beta_0$  della retta).

Costruiamo quindi un sistema di ipotesi del tipo:

$$\begin{cases} H_0 : \beta_1 = \beta_1^* \\ vs. \\ H_1 : \beta_1 \neq \beta_1^* \end{cases} \quad (7.23)$$

dove  $\beta_1^*$  è un valore che abbiamo fissato a priori, in base alla nostra conoscenza o aspettativa o necessità.

Per costruire la *regola di decisione* in merito al sistema di ipotesi 7.23 possiamo utilizzare la statistica test  $T$ , in quanto abbiamo dimostrato che il coefficiente angolare  $b_1$  della retta di regressione segue la distribuzione Normale in (7.23) di cui però non conosciamo la varianza  $\text{Var}(b_1)$  dipendendo quest'ultima da  $\sigma^2$ , varianza dell'errore anch'essa incognita. Però, possiamo stimare

$$\text{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \quad (7.24)$$

con

$$S_n^2(b_1) = \frac{\text{SSE}/(n-2)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \quad (7.25)$$

dove  $\text{SSE}/(n-2)$ , varianza dei residui; quest'ultima è uno stimatore *non distorto* di  $\sigma^2$  poiché  $\mathbb{E}(\text{SSE}) = (n-2)\sigma^2$  e anche *consistente* poiché  $\text{SSE}/(n-2) \xrightarrow{P} \sigma^2$ . Di conseguenza, possiamo concludere che

$$T^* = \frac{b_1 - \beta_1^*}{S_n(b_1)} \underset{H_0}{\sim} t_{n-2} \quad (7.26)$$

Scegliamo ora un valore del livello di significatività  $\alpha$  e imponiamo, come siamo abituati a fare, che la probabilità di effettuare un errore del primo tipo sia esattamente uguale ad  $\alpha$ , vale a dire

$$P(|T^*| > t_{crit} | H_0 : \beta_1 = \beta_1^*) = \alpha \quad (7.27)$$

sicché  $t_{crit} = t_{n-2; 1-(\alpha/2)}$ . Quindi ottenuto il valore osservato della statistica test  $T$ , dato da  $t_{oss} = \frac{b_1 - \beta_1^*}{s_n(b_1)}$ , possiamo formulare la seguente *regola di decisione* in merito a  $H_0 : \beta_1 = \beta_1^*$ :

$$\dots \text{ si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } |t_{oss}| > t_{n-2; 1-(\alpha/2)} \quad (7.28)$$

Laddove l'ipotesi nulla in (7.23) assume la forma  $H_0 : \beta_1 = 0$  contro l'alternativa  $H_1 : \beta_1 \neq 0$  si parla di **test di significatività per il parametro**  $\beta_1$  dove significatività di  $\beta_1$  va intesa come "... $\beta_1$  significativamente diverso da zero"; se dai dati a disposizione non emergessero evidenze contro l'ipotesi nulla  $H_0 : \beta_1 = 0$  e di conseguenza quest'ultima non potesse essere rifiutata, lo stesso modello (7.1) verrebbe meno riducendosi a  $Y_i = \beta_0, \forall i$ .

Se ora ricordiamo la relazione che corre tra *regione di non rifiuto*  $\bar{C}_\alpha$  di  $H_0$  in presenza di ipotesi alternativa bilaterale e intervallo di confidenza di livello  $(1 - \alpha)$ , si ha che  $IC_{\beta_1}(1 - \alpha) = \bar{C}_\alpha$  ovvero,

$$IC_{\beta_1}(1 - \alpha) : \left[ b_1 - t_{n-2; 1-(\alpha/2)} \sqrt{\frac{SSE/(n-2)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}, \quad b_1 + t_{n-2; 1-(\alpha/2)} \sqrt{\frac{SSE/(n-2)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}} \right] \quad (7.29)$$

Quanto finora detto in merito a  $\beta_1$  può essere **riproposto, mutatis mutandis**, per il **parametro**  $\beta_0$ .

In ambito applicativo, si preferisce spesso adottare un approccio alla verifica di ipotesi sui parametri basato sul  $p$ -value.

## 7.2.2 Test per la significatività del modello

Verifichiamo ora la *significatività globale* del modello (7.1) alla luce delle evidenze contenute nei dati.

Per fare ciò, riprendiamo i concetti di devianza totale (SSTO), devianza di regressione (SSR), devianza residua (SSE), rispettivamente date dalle quantità

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \quad SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad \text{e} \quad SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (7.30)$$

La costruzione del test di *significatività globale* poggia sul *coefficiente di determinazione*  $R^2$  che ricordiamo è definito come segue:

$$R^2 = \frac{SSR}{SSTO} = \mathbb{C}orr^2(X, Y) = \frac{\mathbb{C}ov^2(X, Y)}{\mathbb{V}ar(x) \mathbb{V}ar(Y)} \quad (7.31)$$

Si tratta dunque di sottoporre a verifica il seguente sistema di ipotesi

$$\begin{cases} H_0 : R^2 = 0 \\ vs. \\ H_1 : R^2 \neq 0 \end{cases} \quad (7.32)$$

sulla base della statistica test  $W$  così definita

$$W = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} = \frac{(n-p)}{(p-1)} \frac{\text{SSR}}{\text{SSE}} \underset{H_0}{\sim} F_{(p-1),(n-p)} \quad (7.33)$$

dove  $p$  è il numero di parametri del modello ( $p=2$  per il modello di regressione lineare semplice che stiamo qui trattando) mentre  $(p-1)$  e  $(n-p)$  sono i *gradi di libertà* rispettivamente del numeratore e del denominatore del rapporto che in (7.33) definisce la statistica test  $W$ , ovvero i parametri che indicizzano la famiglia delle distribuzioni  $F$  di Fisher-Snedecor. Si può infatti dimostrare che  $W$  è data dal rapporto di due v.c. *chi-quadrato indipendenti* rispettivamente divise per i propri gradi di libertà obbedendo così alla definizione che abbiamo dato di v.c. di Fisher-Snedecor..

Intuitivamente, si **rifiuta**  $H_0 : R^2 = 0$  per *grandi* valori di  $W$ ; coerentemente, fissato in  $\alpha$  il livello di significatività *nomina*le del test, possiamo definire la seguente *regola di decisione* per il sistema di ipotesi (7.32)

$$\dots \text{si rifiuta } H_0 \text{ in favore di } H_1 \text{ quando } w_{\text{oss}} > w_{\text{crit}} = F_{(p-1),(n-p);\alpha} \quad (7.34)$$

con  $w_{\text{oss}}$  valore della statistica test  $W$  osservato sui dati di cui si dispone.

Inoltre, per decidere in merito a (7.32) si potrebbe anche utilizzare l'approccio basato sul *p-value*, ovvero determinare  $\alpha^* = P(W > w_{\text{oss}} \mid H_0 : R^2 = 0)$  e laddove

$$\alpha^* < \alpha \quad (7.35)$$

*rifiutare* l'ipotesi  $H_0 : R^2 = 0$  concludendo quindi a favore della *significatività (globale)* del modello di regressione ipotizzato in (7.1).

Presentiamo di seguito un esempio numerico allo scopo di vedere all'opera concetti e tecniche inferenziali fin qui introdotte. L'esempio poggia volutamente su un numero assai limitato di osservazioni in maniera che i calcoli necessari possano essere eseguiti con carta e penna. Resta inteso che il medesimo modo di operare può essere esteso a un numero qualsiasi di osservazioni nel qual caso, l'aiuto di pacchetti, quali per esempio R o Matlab, può rivelarsi senz'altro utile.

**Esempio 7.2.1.** Ipotizziamo di aver misurato  $n = 8$  studenti con un *test di apprendimento* durante l'anno scolastico e di voler studiare la sua *relazione* con il *voto finale* della materia. I dati disponibili sono riportati nella seguente tabella:

Studente	Punteggio al test	Voto finale
1	12	8
2	10	7
3	14	8
4	9	5
5	9	6
6	13	9
7	11	7
8	8	5

Ora, dai dati in tabella, si ricavano le seguenti quantità di sintesi:

$$\sum_{i=1}^8 x_i = 86, \quad \sum_{i=1}^8 Y_i = 55, \quad \sum_{i=1}^8 x_i^2 = 956, \quad \sum_{i=1}^8 Y_i^2 = 393, \quad \sum_{i=1}^8 x_i Y_i = 611$$

La *correlazione* fra il punteggio del Test di apprendimento e il Voto finale è pari a

$$\text{Corr}(x, Y) = \frac{\text{Cov}(x, Y)}{\sqrt{\text{Var}(x) \text{Var}(Y)}} = \frac{\sum_{i=1}^8 x_i Y_i - n \bar{x}_n \bar{Y}_n}{\sqrt{(\sum_{i=1}^8 x_i^2 - n \bar{x}_n^2)(\sum_{i=1}^8 Y_i^2 - n \bar{Y}_n^2)}} = 0.91$$

Questa quantità, lo abbiamo già detto, non fornisce informazioni sul *tipo di relazione* esistente tra le due variabili in questione né sul ruolo da esse giocato (variabile dipendente o variabile indipendente).

Però, osservando i dati possiamo vedere che a *valori alti* nel punteggio al Test di apprendimento corrispondono *valori alti* del Voto finale e viceversa. Poiché la variabile Test di apprendimento e la variabile Voto sono separati nel tempo e successivi, è illogico pensare che il voto finale possa aver avuto un'azione retroattiva e aver influenzato il test, mentre è più logico immaginare che il risultato del test sia in relazione diretta con il voto. Ancora più logico è pensare che entrambe le variabili siano "influenzate" da altre variabili come il numero di ore passate a studiare, la facilità/difficoltà della materia, la predisposizione personale e così via.

A scopo didattico, partiamo dal presupposto che il punteggio al Test di apprendimento possa essere la *causa* del Voto finale.

Se rappresentiamo graficamente le due variabili, usando l'asse delle ascisse per i valori del punteggio al Test di apprendimento ( $x$ ) e l'asse delle ordinate per i valori del Voto finale ( $Y$ ), otterremo il grafico di Fig. 7.2

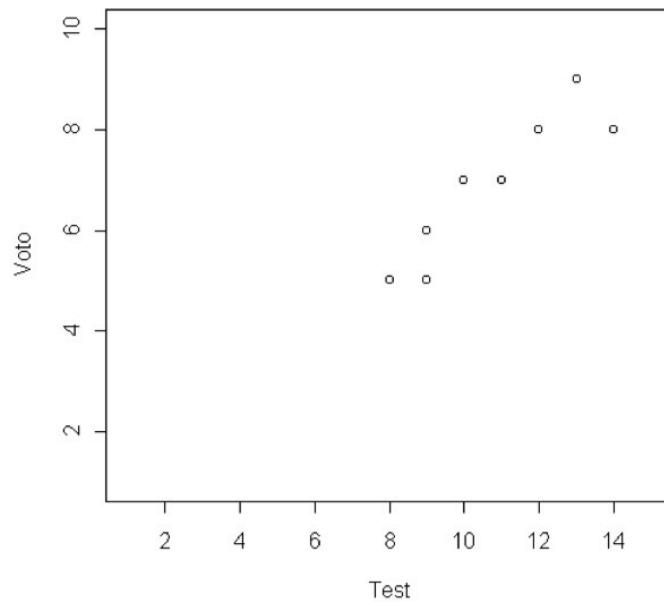


Figura 7.2: Scatterplot di Voto finale e Punteggio al test di apprendimento

Se lo osserviamo attentamente, possiamo immaginare una *linea retta* che passa, più o meno, in mezzo ai punti e che indica la *tendenza* della *relazione* fra le due variabili; potremmo quindi adottare un *modello lineare* per rappresentare analiticamente la *relazione causale* che corre tra  $x$  e  $Y$  e che *spiega*  $Y$  in termini di  $x$  secondo il modello (7.1).

Sicché usando le formule (7.9) otteniamo le *stime a minimi quadrati*

$$b_0 = 0.135 \quad \text{e} \quad b_1 = 0.627 \quad (7.36)$$

dei parametri  $\beta_0$  e  $\beta_1$  sicchè la retta di regressione sarà

$$\hat{Y}_i = 0.135 + 0.627 x_i \quad (7.37)$$

Ricorrendo a (7.10) e (7.13) è immediato calcolare

$$SSTO = (n - 1)\mathbb{V}ar(Y) = 14.875 \quad \text{e} \quad SSE = (1 - R^2)SSTO = 2.492 \quad (7.38)$$

e, per differenza,  $SSR = SSTO - SSE = 14.875 - 2.492 = 12.383$  sicché

$$R^2 = \frac{SSR}{SSTO} = \frac{12.383}{14.875} = 0.832 \quad (7.39)$$

La stima della varianza dell'errore,  $\sigma^2$  è data da

$$s^2 = \frac{SSE}{(n - 2)} = 0.415 \quad (7.40)$$

e di conseguenza, tenuto conto del fatto che  $\sum_{i=1}^n (x_i - \bar{x}_n)^2 = (n - 1) \mathbb{V}ar(x)$ , si ha

$$s_n(b_1) = \left( \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right)^{1/2} = \sqrt{\frac{0.415}{31.5}} = 0.115 \quad (7.41)$$

Verifichiamo subito se il parametro  $\beta_1$  è significativamente diverso da zero ricorrendo alla statistica test

$$T^* = \frac{b_1 - \beta_1^*}{S_n(b_1)} \stackrel{H_0}{=} \frac{b_1}{S_n(b_1)} \stackrel{H_0}{\sim} t_{n-2} \quad (7.42)$$

per sottoporre a verifica il sistema di ipotesi  $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$ . Fissato  $\alpha = 0.05$  per cui, dalle tavole della distribuzione  $t$  di Student si ricava  $t_{(6; 0.025)} = 2.447$  da cui la regione critica di livello 0.05 data dal complemento di  $C_{0.05} = [-2.447, 2.447]$  ovvero

$$C_{0.05} = (-\infty, -2.447) \cup (2.447, \infty) \quad (7.43)$$

per cui, essendo

$$t_{oss} = \frac{b_1}{S_n(b_1)} = \frac{0.627}{0.115} = 5.45 \in C_{0.05} \quad (7.44)$$

si ha ragione di rifiutare  $H_0$  in favore della *significatività* del parametro  $\beta_1$  ovvero, sulla base dei dati a disposizione, possiamo ritenere il parametro  $\beta_1$  *significativamente diverso da zero*. Per quanto abbiamo inoltre visto in (7.29)

$$IC_{\beta_1}(0.95) = [0.343, 0.908] \quad (7.45)$$

Infine possiamo testare la *significatività globale* del modello; per far questo dobbiamo ricavare il valore della statistica  $W$  osservato

$$w_{oss} = \frac{R^2/(p - 1)}{(1 - R^2)/(n - p)} = \frac{0.832/(2 - 1)}{(1 - 0.832)/(8 - 2)} = 29.814 \quad (7.46)$$

ed essendo  $w_{crit} = F_{(2-1),(8-2);0.05} = 5.99$ , ergo la regione critica di livello  $\alpha = 0.05$  associata al test uguale a  $C_{0.05} = (5.99, +\infty)$ , rifiuteremo l'ipotesi nulla  $H_0 : R^2 = 0$  poiché

$$w_{oss} = 29.814 > w_{crit} = 5.99 \text{ ovvero } w_{oss} = 29.814 \in C_{0.05} = (5.99, +\infty)$$

concludendo a favore della *significatività globale* del modello adottato.

La quasi totalità dei pacchetti statistici riporta tra i risultati restituiti di default dall'esecuzione delle routine legate alla stima del modello di regressione, la seguente tavola

Fonte	gdl	SS	MS	$w_{oss}$	p-value
Modello	1	12.383	12.383	29.814	0.0016
Residui	6	2.492	0.415		
Total	7	14.875			

Tabella 7.1: Tavola di ANoVA

detta *Tavola di ANoVA*, acronimo di *Analysis of Variance*; come si può vedere, la tabella reccoglie i valori di alcune note statistiche che si sono rivelate cruciali per la valutazione della *bontà di adattamento* del modello ipotizzato ai dati.

### 7.3 Un breve cenno al modello di regressione lineare multipla

Laddove i regressori siano  $p \geq 2$  è immediato scrivere il modello (7.1) nella maniera seguente:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (7.47)$$

ovvero

$$Y_i = \sum_{j=0}^p \beta_j x_{ij} + \varepsilon_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (7.48)$$

con  $x_{i0} = 1$  per ogni  $i = 1, 2, \dots, n$ . Potremmo anche scrivere il tutto in forma compatta usando l'algebra matriciale, ottenendo

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\epsilon}_{n \times 1} \quad (7.49)$$

con

$$\mathbf{X}_{n \times (p+1)} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad (7.50)$$

Premettiamo subito che, *mutatis mutandis*, valgono ancora le *ipotesi classiche* viste per il modello di regressione lineare semplice (7.1). In particolare ora, i regressori  $x_1, x_2, \dots, x_p$  si assumono essere incorrelati e il *rango* di  $\mathbf{X}$  pari a  $p+1 < n$ ; quest'ultima assunzione in merito al rango di  $\mathbf{X}$  implica che il numero di informazioni campionarie *genuine* (ossia non ridondanti perché esprimibili come combinazioni lineari delle altre) sia *almeno uguale* al numero di parametri da stimare. Intuitivamente, ciò vuol anche dire che ogni variabile esplicativa (o regressore che dir si voglia)  $x_i$  porta informazioni aggiuntive alla costruzione di un modello esplicativo del fenomeno rappresentato da  $Y$ .

Ora, l'espressione equivalente a (7.4) è

$$Q(\beta) = \varepsilon' \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}'\mathbf{Y} + \beta \mathbf{X}'\mathbf{X}\beta - 2\beta' \mathbf{X}'\mathbf{Y} \quad (7.51)$$

e dunque derivando rispetto le componenti del vettore dei parametri  $\beta$  e uguagliando a zero, si oltiene l'equazione

$$\frac{d}{d\beta} Q(\beta) = 2(\mathbf{X}'\mathbf{X})\beta - 2\mathbf{X}'\mathbf{Y} = 0 \quad (7.52)$$

la cui *unica* soluzione è data da

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \quad (7.53)$$

che sarà lo *stimatore a minimi quadrati* di  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ .

A questo proposito ricordiamo che, affinché la soluzione dell'equazione (7.52) *esista* e sia *unica* è necessario che la matrice quadrata di ordine  $p+1$ ,  $(\mathbf{X}'\mathbf{X})$  sia *invertibile*, cosa peraltro garantita se il suo rango vale  $p+1$  come di fatto è, essendo

$$\text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{X}) = p+1. \quad (7.54)$$

Inoltre, sfruttando gli elementi introdotti nel paragrafo 1.5.2 relativo alla distribuzione Normale multivaria, si può dimostrare che, in analogia con quanto visto nel Teorema 7.2.1, vale il seguente risultato

$$\mathbf{b} \sim N_{p+1} \left( \beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right) \quad (7.55)$$

dove  $\sigma^2 = \text{Var}(\varepsilon_i)$ ,  $i = 1, 2, \dots, n$ , è la varianza dell'errore.

A partire dalla distribuzione dello stimatore a minimi quadrati di  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$  si possono estendere tutte le procedure inferenziali studiate in ambito del modello di regressione lineare semplice al modello di regressione lineare multipla.

Ma qui ci fermiamo, lasciando come esercizio per gli interessati pensarci un po' su....

# 8 Principles of Data Reduction

Cominciamo col richiamare alcuni concetti che abbiamo già avuto modo di vedere in maniera più o meno diretta.

- a) Ogni statistica  $T_n : \mathfrak{X} \subseteq \mathbb{R}^n \mapsto \Theta \subseteq \mathbb{R}^k$  realizza un *processo di riduzione* della dimensionalità dei dati (o, in altre parole, di *sintesi* dell'informazione circa  $\theta \in \Theta$  contenuta nei dati), non fosse altro, per lo stesso fatto che  $k << N$ .

Per esempio, se  $x_i$  e  $x_j$  rappresentano due *realizzazioni* del campione casuale  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ , queste saranno *equivalenti in informazione* riguardo  $\theta \in \Theta$  se

$$T_n(x_1) = T_n(x_2) \quad (8.1)$$

nonostante i valori delle due realizzazioni  $x_i$  e  $x_j$  possano *differire*, vale a dire  $x_i \neq x_j$ .

- b) ogni statistica  $T_n$  induce una *partizione*  $\mathcal{T}$  su  $\mathfrak{X}$  costituita dagli elementi

$$A_{t_n} = \{\mathbf{x} \in \mathfrak{X} : T_n(\mathbf{x}) = t_n, \mathbf{x} \in \mathfrak{X}\} \quad (8.2)$$

Nell'esempio di poc'anzi,  $x_i$  e  $x_j$  appartengono allo stesso elemento  $A_{t_n}$  della *partizione*  $\mathcal{T}$ .

## 8.1 Verosimiglianza e principio di verosimiglianza

In questo paragrafo introdurremo e studieremo una speciale quanto importante statistica chiamata *funzione di verosimiglianza* che può essere usata per sintetizzare i dati.

Ci sono molti modi di usare la funzione di verosimiglianza, alcuni dei quali sono menzionati in questo paragrafo e altri nei capitoli successivi. Ma la conclusione più importante cui si arriva in questo paragrafo è che, se accettiamo alcuni principi, la funzione di verosimiglianza *deve* necessariamente essere usata come strumento per la riduzione dei dati.

Sia  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  un campione casuale da una popolazione avente funzione di distribuzione cumulata  $F_X(x; \theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}^k$  e sia  $f_X(x; \theta)$  la corrispondente funzione di densità o di massa probabilistica.

**Definizione 8.1.1** (Funzione di verosimiglianza). Osservato  $\mathbf{X} = \mathbf{x}$ , diremo la funzione di  $\theta$

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n f_{X_i}(x_i; \theta), \theta \in \Theta \subseteq \mathbb{R}^k \quad (8.3)$$

*funzione di verosimiglianza* associata alla distribuzione da cui proviene il campione casuale di cui abbiamo la determinazione  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , mentre diremo

$$\ell(\theta|\mathbf{x}) = \ln L(\theta|\mathbf{x}) = \ln \prod_{i=1}^n f_{X_i}(x_i; \theta) = \sum_{i=1}^n \ln f_{X_i}(x_i; \theta), \quad \theta \in \Theta \subseteq \mathbb{R}^k \quad (8.4)$$

*funzione di log-verosimiglianza.*

Vale la pena osservare e rimarcare che:

- a) la funzione di verosimiglianza  $L(\theta|\mathbf{x})$  è *funzione del solo* parametro  $\theta$  e attribuisce un *ordine di preferenza* ai diversi valori di  $\theta \in \Theta$  alla luce dei dati osservati  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ ; è per sottolineare proprio questo fatto che scriviamo  $L(\theta|\mathbf{x})$  piuttosto che limitarci a  $L(\theta)$ . Analogio discorso può essere fatto per la funzione di log-verosimiglianza  $\ell(\theta|\mathbf{x})$
- b) spesso la funzione di log-verosimiglianza  $\ell(\theta|\mathbf{x})$  ha un'espressione analitica più semplice di quella della funzione di verosimiglianza: ciò ne facilita la manipolazione matematica e i calcoli e, per queste ragioni, viene spesso preferita a  $L(\theta|\mathbf{x})$
- c) possiamo guardare a  $L(\theta|\mathbf{x})$  come funzione di densità o di massa *congiunta* di  $(X_1, X_2, \dots, X_n)$  valutata in  $(x_1, x_2, \dots, x_n)$  (sebbene non normalizzata).
- d) la funzione verosimiglianza  $L(\theta|\mathbf{x})$  induce una *partizione*  $\mathfrak{L}$  sullo spazio campionario  $\mathfrak{X}$ : tale partizione è detta *partizione di verosimiglianza* ed è *condivisa* con la funzione di log-verosimiglianza  $\ell(\theta|\mathbf{x})$

**Esempio 8.1.1.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  da  $b(1, p)$ , con  $p \in [0, 1]$ . Sia inoltre  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  una sua determinazione. La funzione di verosimiglianza

$$L(p|\mathbf{x}) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \mathbb{1}_{\{0;1\}}(x_i) = p^{\sum x_i} (1-p)^{n-\sum x_i} \prod_{i=1}^n \mathbb{1}_{\{0;1\}}(x_i) \quad (8.5)$$

induce una *partizione* (di verosimiglianza) dello spazio campionario  $\mathfrak{X}$  avente elementi

$$A_x = \left\{ \mathbf{x} \in \mathfrak{X} \mid \sum_{i=1}^n x_i = x \right\}, \quad \text{con } x \in \{0, \dots, n\} \quad (8.6)$$

Supponiamo ora che il campione casuale segua la distribuzione  $\mathcal{G}(\alpha, \beta)$ ,  $\alpha > 0, \beta > 0$ ; in questo caso

$$\begin{aligned} L(\alpha, \beta; \mathbf{x}) &= \prod_{i=1}^n \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-\frac{x_i}{\beta}} \mathbb{1}_{\mathbb{R}^+}(x_i) = \\ &= \left[ \frac{1}{\Gamma(\alpha)\beta^\alpha} \right]^n \left( \prod_{i=1}^n x_i \right)^{\alpha-1} e^{-\frac{\sum x_i}{\beta}} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \end{aligned} \quad (8.7)$$

induce una *partizione* (di verosimiglianza) dello spazio campionario  $\mathfrak{X}$  avente elementi

$$A_{t_1, t_2} = \left\{ \mathbf{x} \in \mathfrak{X} \mid \sum_{i=1}^n x_i = t_1 \wedge \prod_{i=1}^n x_i = t_2 \right\}, \quad \text{con } (t_1, t_2) \in \mathcal{T} \quad (8.8)$$

Coerentemente con quanto fin qui detto, possiamo guardare  $L(\theta|\mathbf{x})$  (o  $\ell(\theta|\mathbf{x})$ ) come uno strumento per estrarre informazione dai dati relativamente al valore più plausibile del parametro incognito  $\theta$  sul quale verte il problema inferenziale che si sta affrontando. Diamo allora la seguente definizione:

**Definizione 8.1.2** (Stimatore di massima verosimiglianza). Sia  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  la realizzazione di un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da una popolazione avente funzione di distribuzione  $F_X(x; \theta)$ ,  $\theta \in \Theta \subseteq \mathbb{R}^k$ . Diremo

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L(\theta|\mathbf{x}) = \arg \max_{\theta \in \Theta} \ell(\theta|\mathbf{x}) \quad (8.9)$$

stimatore di massima verosimiglianza di  $\theta$ .

Quanto appena stabilito nella definizione si può anche esprimere nei due seguenti modi equivalenti:

- a)  $\hat{\theta}_n$  è quel valore di  $\theta$  per il quale la determinazione  $\mathbf{x}$  del campione casuale è la più plausibile
- b)  $\hat{\theta}_n$  è quel valore di  $\theta$  cui corrisponde la massima probabilità che  $F_X(x; \theta)$  generi la realizzazione  $\mathbf{x}$  del campione casuale effettivamente poi osservata.

**Esempio 8.1.2.** Sia  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  l'esito della ripetizione, sotto identiche condizioni, di un esperimento bernoulliano la cui probabilità di successo è  $\theta \in [0, 1]$ ; in altre parole,  $X_i \sim b(1, \theta)$ ,  $i = 1, 2, \dots, n$ . Allora

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i) = \theta^x (1 - \theta)^{n-x} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(x_i) \quad (8.10)$$

con  $x = \sum_{i=1}^n x_i$ , numero di successi nelle  $n$  replicazioni indipendenti dell'esperimento bernoulliano; e analogamente,

$$\ell(\theta|\mathbf{x}) = \ln L(\theta|\mathbf{x}) = x \ln(\theta) + (n - x) \ln(1 - \theta) + \ln \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(x_i) \quad (8.11)$$

Ora,

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \ell(\theta|\mathbf{x}) \quad (8.12)$$

sicché, in questo caso,  $\hat{\theta}_n$  sarà soluzione dell'equazione

$$\frac{d}{d\theta} \ell(\theta|\mathbf{x}) = \frac{x}{\theta} - \frac{n - x}{1 - \theta} = 0 \quad (8.13)$$

dunque

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (8.14)$$

dal momento che

$$\frac{d^2}{d\theta^2} \ell(\theta|\mathbf{x}) < 0 \quad (8.15)$$

ovvero  $\hat{\theta}_n$  coincide con la frequenza relativa campionaria, stimatore plug-in di  $\theta$ .

**Esempio 8.1.3.** Sia  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  la realizzazione di un campione casuale proveniente da una distribuzione Uniforme su  $(0, \theta)$ ,  $\theta > 0$ . Allora,

$$\begin{aligned}
L(\theta|\mathbf{x}) &= \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}_{(0,\theta)}(x_i) \\
&= \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{(0,\theta)}(x_i) \\
&= \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \mathbb{1}_{(0,\theta)}(x_{(n)}) \\
&= \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \mathbb{1}_{(x_{(n)}, +\infty)}(\theta) \\
&= \frac{1}{\theta^n} \mathbb{1}_{(x_{(n)}, +\infty)}(\theta) \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i)
\end{aligned} \tag{8.16}$$

Ora la funzione di verosimiglianza  $L(\theta|\mathbf{x})$  è strettamente *monotona decrescente* sull'intervallo  $(x_{(n)}, +\infty)$  sicché

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L(\theta|\mathbf{x}) = \arg \max_{\theta \in \Theta} \left( \frac{1}{\theta^n} \mathbb{1}_{(x_{(n)}, +\infty)}(\theta) \right) = X_{(n)} \tag{8.17}$$

ossia, in questo caso, lo *stimatore di massima verosimiglianza* di  $\theta$  è restituito dal *massimo campionario*  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$  che, ancora una volta, coincide con lo *stimatore plug-in* di  $\theta$ .

In questo caso, procedere come in (8.13) non sarebbe stato possibile non essendo  $L(\theta|\mathbf{x})$  derivabile in  $\theta = 0$ . Ciò non toglie che l'argmax di  $L(\theta|\mathbf{x})$  possa essere comunque calcolato e quindi possa essere individuato lo *stimatore di massima verosimiglianza* di  $\theta$  secondo la definizione che ne poc'anzi abbiamo dato.

**Esempio 8.1.4.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $N(\mu, \sigma^2)$ , con  $\mu, \sigma^2$  non noti. La funzione di verosimiglianza è data da

$$\begin{aligned}
L(\mu, \sigma^2|\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\} \mathbb{1}_{\mathbb{R}}(x_i) \\
&= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_i)
\end{aligned} \tag{8.18}$$

ed è facile verificare che la funzione di log-verosimiglianza ad essa associata è la seguente:

$$l(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \tag{8.19}$$

Derivando la funzione di log-verosimiglianza rispetto  $\mu$  e  $\sigma^2$ , uguagliando le derivate prime a zero e formato il sistema delle equazioni di stima nelle due variabili  $\mu$  e  $\sigma^2$ ,

otteniamo

$$\begin{cases} \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = 0 \\ \frac{1}{2\sigma^4} \left[ \sum_{i=1}^n (x_i - \mu)^2 - n\sigma^2 \right] = 0 \end{cases} \quad (8.20)$$

la cui soluzione restituisce lo stimatore di massima verosimiglianza di  $(\mu, \sigma^2)$  dato da

$$\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n^2) = \left( \bar{X}_n, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right) \quad (8.21)$$

dove

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2 \quad (8.22)$$

### 8.1.1 Principio formale di verosimiglianza

La funzione di verosimiglianza  $L(\theta|x)$  sintetizza le informazioni contenute nella determinazione  $\mathbf{x}$  campione casuale  $\mathbf{X}$  con riferimento a uno specifico modello probabilistico  $F_X(x; \theta)$ ,  $\theta \in \Theta$  e l'inferenza che è poggiata su questa funzione è la conseguenza di principi generali che brevemente riassumono i seguenti principi:

- **Principio debole di verosimiglianza:** se  $\mathbf{x}$  e  $\mathbf{y}$  sono due determinazioni del campione casuale  $\mathbf{X}$  da una distribuzione  $F(x; \theta)$  tali per cui  $L(\theta|\mathbf{x})$  è proporzionale a  $L(\theta|\mathbf{y})$ , cioè esiste una costante  $C(\mathbf{x}, \mathbf{y})$  tale che

$$L(\theta|\mathbf{x}) = C(\mathbf{x}, \mathbf{y}) L(\theta|\mathbf{y}), \quad \forall \theta$$

allora le *conclusioni inferenziali* su  $\theta$  tratte da  $\mathbf{x}$  e  $\mathbf{y}$  devono essere le stesse.

- **Principio forte di verosimiglianza:** se  $\mathbf{x}$  e  $\mathbf{y}$  sono due determinazioni dei campioni casuali  $\mathbf{X}$  e  $\mathbf{Y}$  estratti rispettivamente dalle distribuzioni  $F(x; \theta)$  e  $G(y; \theta)$  tali per cui  $L_F(\theta; \mathbf{x})$  è proporzionale a  $L_G(\theta; \mathbf{y})$ , cioè esiste una costante  $C(\mathbf{x}, \mathbf{y})$  tale che

$$L_F(\theta; \mathbf{x}) = C(\mathbf{x}, \mathbf{y}) L_G(\theta; \mathbf{y}), \quad \forall \theta$$

allora le *conclusioni inferenziali* su  $\theta$  basate su  $\mathbf{x}$  e  $\mathbf{y}$  devono essere le stesse.

Il principio *debole* di verosimiglianza conduce alle stesse inferenze su  $\theta$  a parità di modello mentre il principio *forte* porta alle medesime conclusioni inferenziali anche con modelli differenti.

**Esempio 8.1.5.** Un comune vuole stimare l'ignota proporzione  $\theta$  dei veicoli alimentati a diesel che circolano in città e a tale scopo, incarica tra ricercatori della rilevazione dei dati e della stima di  $\theta$ . I tre ricercatori si collocano a un incrocio più trafficato della città per osservare un campione di veicoli:

- a) ricercatore *A*: fissa l'ampiezza del campione di osservazioni in  $n = 10$  e osserva dove  $D$  sta per *veicolo alimentato a diesel* e  $ND$  per *veicolo non alimentato a diesel*, ottenendo

$$L_A(\theta|\mathbf{x}) = \binom{10}{3} \theta^3 (1-\theta)^7 \quad (8.23)$$

1	2	3	4	5	6	7	8	9	10
ND	D	ND	ND	ND	D	ND	ND	ND	D

Tabella 8.1: Osservazioni rilevate dal ricercatore A

- b) ricercatore  $B$ : decide di rilevare le osservazioni nell'arco di 10 minuti e osserva gli stessi veicoli rilevato dal ricercatore  $A$ , ottenendo

$$L_B(\theta|\mathbf{x}) = (1 - \theta)\theta(1 - \theta) \cdots (1 - \theta)\theta = \theta^3(1 - \theta)^7 \quad (8.24)$$

- c) ricercatore C: decide di sospendere la rilevazione non appena avrà osservato il passaggio di tre veicoli diesel e rileva gli stessi veicoli di  $A$  e di  $B$

$$L_C(\theta|\mathbf{x}) = P_\theta \left( X_{10} = 1 \mid \sum_{i=1}^9 X_i = 2 \right) = \theta \binom{9}{2} \theta^2(1 - \theta)^7 = \binom{9}{2} \theta^3(1 - \theta)^7 \quad (8.25)$$

dove  $X_i = 1$  se l' $i$ -mo veicolo transitato all'incrocio è alimentato a diesel e 0 altrimenti.

E' indubbio che per tutti e tre i ricercatori il valore più plausibile di  $\theta$  sarà 3/10 ovvero quello che è *massimamente coerente* con il modello adottato alla luce di quanto osservato.

Consideriamo il modello (8.23) del ricercatore  $A$  e supponiamo però che la realizzazione del campione casuale sia quella data in Tabella 8.2.

1	2	3	4	5	6	7	8	9	10
D	D	ND	ND	ND	D	ND	ND	ND	ND

Tabella 8.2: Osservazioni rilevate dal ricercatore A

In tal caso la nuova funzione di verosimiglianza sarà comunque *proporzionale* a  $L_A(\theta|\mathbf{x})$  sicché le informazioni fornite dalle due realizzazioni del campione casuale sono *equivalenti* (*in informazione*) e porteranno entrambe alle *medesime* conclusioni inferenziali su  $\theta$  (in virtù del *Principio di verosimiglianza debole*).

Consideriamo ora le verosimiglianze associate ai due *diversi* modelli (8.23) e (8.25); dal momento che queste sono tra loro *proporzionali*, seppur riferite a due *modelli diversi*, condurranno alle *medesime* conclusioni inferenziali circa  $\theta$ . (in virtù del *Principio di verosimiglianza forte*).

## 8.2 Sufficienza

Il concetto di sufficienza è stata introdotto nel 1920 da R.A. Fisher, statistico, matematico e genetista britannico, e ripresa poi nel lavoro *On the mathematical foundations of theoretical statistics* pubblicato nel 1922. Fisher viene considerato

colui che ha trasformato la statistica da una raccolta di metodi e tecniche spesso *ad hoc* in una scienza moderna, individuando e definendo i concetti fondativi su cui si regge l'intero edificio della moderna statistica matematica e a cui ricondurne i principali risultati.

La nozione di sufficienza gioca un ruolo di assoluto rilievo in tutta l'inferenza statistica e ne costituisce uno dei pilastri di fondazione. Una volta introdotto, il concetto di sufficienza ci accompagnerà fino alla fine del corso.

**Definizione 8.2.1** (Statistica sufficiente). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da  $F_X(x; \theta)$ , con  $\theta \in \Theta \subseteq \mathbb{R}^k$  e sia  $T_n(X_1, X_2, \dots, X_n)$  una statistica. Diremo  $T_n$  *statistica sufficiente* per  $\theta$  se (e solo se) la distribuzione condizionata di  $(X_1, X_2, \dots, X_n)$  dato  $T_n = t_n$  non dipende da  $\theta$ .

**Esempio 8.2.1.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  da una distribuzione di Bernoulli  $b(1, \theta)$  e sia  $T_n(X_1, X_2, \dots, X_n) = \sum X_i$ . Allora  $T_n$  è statistica sufficiente per  $\theta$ . Infatti,

$$\begin{aligned} f_{\mathbf{X}|T_n}(x_1, \dots, x_n | t_n) &= \frac{f_{X_1, \dots, X_n, T_n}(x_1, \dots, x_n, t_n)}{f_{T_n}(t_n)} \\ &= \frac{\theta^{t_n} (1 - \theta)^{n-t_n}}{\binom{n}{t_n} \theta^{t_n} (1 - \theta)^{n-t_n}} \\ &= \frac{1}{\binom{n}{t_n}} \end{aligned} \tag{8.26}$$

Abbiamo perciò dimostrato che la distribuzione condizionata (8.26) non dipende dal parametro  $\theta$  e pertanto possiamo concludere che  $T_n = \sum X_i$  è una *statistica sufficiente* per  $\theta$ . Notiamo inoltre che questo giustifica la scelta fatta a suo tempo di stimare  $\theta$  con  $\hat{\theta}_n = T_n/n$ .

Se  $T(\mathbf{X})$  è una *statistica sufficiente* per  $\theta$ , allora ogni inferenza su  $\theta$  dipende dal campione  $\mathbf{X}$  soltanto attraverso il valore di  $T(\mathbf{X})$ . In altre parole, detti  $\mathbf{x}$  e  $\mathbf{y}$  due realizzazioni campionarie, se  $T(\mathbf{x}) = T(\mathbf{y})$  l'inferenza su  $\theta$  deve essere la medesima, sia che si osservi  $\mathbf{X} = \mathbf{x}$  o  $\mathbf{X} = \mathbf{y}$ . Tale principio è noto come *principio di sufficienza*.

**Esempio 8.2.2.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  da  $\mathcal{G}(2, \beta)$  e sia  $T_n(X_1, X_2, \dots, X_n) = \sum X_i \sim \mathcal{G}(2n, \beta)$ . Calcoliamo la funzione di densità di  $\mathbf{X}$  condizionato a  $T_n$ :

$$f_{X_1, \dots, X_n | T_n=t_n} = \frac{\left[ \frac{1}{\Gamma(2)\beta^2} \right]^n \exp \left\{ -\frac{1}{\beta} \sum_{i=1}^n x_i \right\} \prod_{i=1}^n x_i}{\frac{1}{\Gamma(2n)\beta^{2n}} \exp \left\{ -\frac{1}{\beta} \sum_{i=1}^n x_i \right\} t_n^{2n-1}} = \frac{\Gamma(2n) \prod_{i=1}^n x_i}{t_n^{2n-1}} \tag{8.27}$$

Notiamo che quest'ultima non dipende in alcun modo da  $\beta$  e pertanto  $T_n = \sum_{i=1}^n X_i$  è una *statistica sufficiente* per  $\beta$ .

**Esempio 8.2.3.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  da una distribuzione con densità

$$f_X(x; \theta) = e^{-(x-\theta)} \mathbb{1}_{[\theta, +\infty)}(x), \quad \theta > 0 \tag{8.28}$$

Dobbiamo ora dimostrare che  $X_{(1)} = \min(X_1, X_2, \dots, X_n)$  è statistica *sufficiente* per  $\theta$  (o per la famiglia di distribuzioni da esso indicizzata e avente densità (8.28)).

La funzione di densità di  $X_{(1)}$  si può ricavare ponendo  $m = 1$  in (2.29) e tenuto conto del fatto che  $\mathcal{S}_X = [\theta, +\infty)$  e che

$$F_X(x; \theta) = \int_{-\infty}^x e^{-(v-\theta)} \mathbb{1}_{[\theta, +\infty)}(v) dv = 1 - e^{-(x-\theta)} \quad (8.29)$$

si ha

$$f_{X_{(1)}}(x; \theta) = n e^{-(x-\theta)} \mathbb{1}_{[\theta, +\infty)}(x). \quad (8.30)$$

Ora la densità della distribuzione condizionata di  $(X_1, X_2, \dots, X_n)$  dato  $T_n(X_1, X_2, \dots, X_n) = X_{(1)}$  è

$$\begin{aligned} \frac{\prod_{i=1}^n e^{-(x_i-\theta)} \mathbb{1}_{[\theta, +\infty)}(x_i)}{n e^{-(x_{(1)}-\theta)} \mathbb{1}_{[\theta, +\infty)}(x_{(1)})} &= \frac{e^{-\sum_{i=1}^n x_i} e^{n\theta} \prod_{i=1}^n \mathbb{1}_{[\theta, +\infty)}(x_i)}{n e^{-(x_{(1)}-\theta)} e^{n\theta} \mathbb{1}_{[\theta, +\infty)}(x_{(1)})} \\ &= \frac{e^{-\sum_{i=1}^n x_i}}{n e^{-n x_{(1)}}} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \end{aligned} \quad (8.31)$$

poichè

$$\prod_{i=1}^n \mathbb{1}_{[\theta, +\infty)}(x_i) = \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \cdot \mathbb{1}_{[\theta, +\infty)}(x_{(1)}) \quad (8.32)$$

e dunque si conclude, per la Definizione 8.2.1, che  $X_{(1)} = \min(X_1, X_2, \dots, X_n)$  è statistica sufficiente per  $\theta$ .

Ora conviene osservare che

- i) ciò che la definizione di statistica sufficiente vuole dire è che se  $T_n$  è statistica sufficiente per  $\theta$  allora essa *contiene tutta l'informazione* presente nel campione riguardo  $\theta$ ; una volta che il valore di  $T_n$  è noto, non possiamo spremere ulteriore informazione da  $(x_1, x_2, \dots, x_n)$  su  $\theta$
- ii) come immediata conseguenza, a partire da  $T_n(x_1, x_2, \dots, x_n) = t_n$  è possibile *ricostruire* campioni *equivalenti in informazione* circa  $\theta$  a quello originario su cui  $t_n$  è stata calcolata tramite un meccanismo di *generazione casuale*. Giusto per fare un esempio, rimanendo nell'ambito dell'Esempio 8.2.1, notiamo che dato  $t_n$  è facile generare un campione equivalente a quello originale, semplicemente utilizzando un generatore di numeri casuali. Definito infatti  $\hat{\theta}_n = t_n/n$ , sarà infatti sufficiente estrarre  $n$  valori casuali  $u_i \in [0, 1]$  e, per ogni  $i \in \{1, \dots, n\}$ , definire un  $x_i^*$  come segue:

$$x_i^* = \begin{cases} 0 & \text{se } u_i \leq \hat{\theta}_n \\ 1 & \text{se } u_i > \hat{\theta}_n \end{cases} \quad (8.33)$$

Il campione  $(x_1^*, \dots, x_n^*)$  generato secondo questa procedura sarà *equivalente* all'originale  $(x_1, \dots, x_n)$  per *quantità di informazione* contenuta riguardo  $\theta$ ; e dal momento che il meccanismo di generazione casuale dei valori  $u_1, \dots, u_n$  non può certo produrre alcuna informazione su  $\theta$ , il contenuto informativo di  $(x_1^*, \dots, x_n^*)$  sarà stato veicolato dalla statistica  $T_n = \sum_{i=1}^n X_i$  e da ascriversi interamente a quest'ultima.

iii) spesso si dice che la statistica  $T_n$  è *sufficiente per la famiglia di distribuzioni*, indicizzata da  $\theta \in \Theta$ , da cui proviene il campione su cui è stata calcolata.

La definizione di sufficienza (vedi Definizione 8.2.1) *non* è per niente *costruttiva* vale a dire, una volta ipotizzata una statistica  $T_n$  essere sufficiente per  $\theta$ , offre un criterio per verificare che ciò effettivamente sia. In realtà, ciò che vorremmo invece avere è una strategia che, a partire dalla espressione analitica della distribuzione che ha generato il campione casuale, permetta di ricavare direttamente per via analitica la statistica sufficiente per  $\theta$ , senza necessità di supporne alcunché sulla sua forma. Il teorema che segue realizza proprio questo nostro desiderio, fornendo una conveniente caratterizzazione del concetto di statistica sufficiente.

**Teorema 8.2.1** (Teorema di fattorizzazione di Neyman). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una popolazione avente funzione di densità (o di massa)  $f_X(x; \theta)$  e sia  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  una sua realizzazione. Allora  $T_n(\mathbf{X})$  è statistica sufficiente per  $\theta$  se e solo se esistono due funzioni non negative  $h$  e  $g$  tali che

$$L(\theta|\mathbf{x}) = h(\mathbf{x}) \cdot g(\theta; t_n(\mathbf{x})), \quad \forall \mathbf{x} \in \mathfrak{X}, \forall \theta \in \Theta \quad (8.34)$$

dove  $h$  è funzione solo di  $\mathbf{x}$  e  $g$  è funzione di  $\theta$  e dipende da  $\mathbf{x}$  solo attraverso  $t_n$ .

*Dimostrazione.* Diamo qui la dimostrazione nel caso discreto; per distribuzioni continue, il procedimento è analogo.

Dimostriamo inizialmente l'*implicazione diretta*.

Sia  $T_n$  una statistica sufficiente e  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  una realizzazione del campione casuale  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ . La funzione di verosimiglianza  $L(\theta|\mathbf{x})$  si può sempre scrivere nella maniera che segue (per il Teorema del prodotto)

$$L(\theta|\mathbf{x}) = f_{\mathbf{X}, T_n}(\mathbf{x}, t_n) = f_{\mathbf{X}|T_n}(\mathbf{x}|t_n) \cdot f_{T_n}(t_n; \theta) \quad (8.35)$$

Ora, la funzione  $f_{\mathbf{X}|T_n}(\mathbf{x}|t_n)$  non può dipendere da  $\theta$  perché  $T_n$  è statistica sufficiente per ipotesi; inoltre  $f_{T_n}(t_n; \theta)$  dipenderà da  $\mathbf{x}$  solo tramite  $t_n$  oltre che da  $\theta \in \Theta$  in quanto è la funzione di massa di  $T_n$ . Pertanto la tesi è dimostrata per  $h(\mathbf{x}) = f_{\mathbf{X}|T_n}(\mathbf{x}|t_n)$  e  $g(t_n; \theta) = f_{T_n}(t_n; \theta)$ .

Dimostriamo ora l'*implicazione inversa*,

Sia

$$L(\theta|\mathbf{x}) = h(\mathbf{x}) \cdot g(t_n; \theta) \quad (8.36)$$

Allora,

$$\begin{aligned} f_{T_n}(t_n; \theta) &= \sum_{\mathbf{x} \in A_{t_n}} L(\theta|\mathbf{x}) \text{ con } A_{t_n} = \{\mathbf{x} \in \mathfrak{X} : T_n(\mathbf{x}) = t_n\} \\ &= \sum_{\mathbf{x} \in A_{t_n}} g(t_n; \theta) \cdot h(\mathbf{x}) \\ &= g(t_n; \theta) \sum_{\mathbf{x} \in A_{t_n}} h(\mathbf{x}) \end{aligned} \quad (8.37)$$

dove l'ultimo passaggio è giustificato dal fatto che  $g(t_n; \theta)$  assume lo stesso valore (ossia, è una quantità costante) sullo stesso  $\mathbf{x} \in A_{t_n}$  in quanto in ogni partizione, il valore di  $t_n$  è fissato. Perciò

$$f_{\mathbf{X}|T_n}(\mathbf{x}|t_n) = \frac{L(\theta|\mathbf{x})}{f_{T_n}(t_n)} = \frac{g(t_n; \theta)h(\mathbf{x})}{g(t_n; \theta) \sum h(\mathbf{x})} = \frac{h(\mathbf{x})}{\sum h(\mathbf{x})} \quad (8.38)$$

e dal momento che  $f_{\mathbf{X}|T_n}(\mathbf{x}|t_n)$  non dipende da  $\theta$ , dalla Definizione 8.2.1 segue la sufficienza della statistica  $T_n(X_1, X_2, \dots, X_n)$ .  $\square$

**Esempio 8.2.4.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  da alcune note distribuzioni e mostriamo che la corrispondente funzione di verosimiglianza può essere fattorizzata come richiesto nel Teorema 8.2.1 di fattorizzazione di Neyman, facilitando così l'individuazione della statistica sufficiente.

- a) Distribuzione di Bernoulli  $b(1, \theta)$ ,  $\theta \in [0, 1]$ : la funzione di verosimiglianza è data da:

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i) = \underbrace{\theta^{\sum x_i} (1-\theta)^{n-\sum x_i}}_{g(\theta; t_n)} \underbrace{\prod_{i=1}^n \mathbb{1}_{\{0,1\}}(x_i)}_{h(\mathbf{x})}$$

sicché  $T_n = \sum_{i=1}^n x_i$  è statistica sufficiente per  $\theta$

- b) Distribuzione Gamma  $\mathcal{G}(\alpha; \beta)$  con  $\alpha = 2n$  (noto),  $\beta > 0$ : la funzione di verosimiglianza è

$$L(\beta; \mathbf{x}) = \underbrace{\left(\frac{1}{\beta^{2n}}\right)^n \exp\left\{-\frac{1}{\beta} \sum_{i=1}^n x_i\right\}}_{g(\beta; t_n)} \underbrace{\left(\frac{1}{\Gamma(2n)}\right)^n \left(\prod_{i=1}^n x_i\right)^{2n-1} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i)}_{h(\mathbf{x})}$$

sicché  $T_n(\mathbf{X}) = \sum_{i=1}^n X_i$  è statistica sufficiente per  $\beta$ .

Laddove  $\alpha$  non fosse noto, riorganizzando i temini della precedente fattorizzazione della funzione di verosimiglianza, si ricava facilmente la statistica *congiuntamente* sufficiente per  $(\alpha, \beta)$  data da

$$(T_{n;1}(\mathbf{X}), T_{n;2}(\mathbf{X})) = \left( \sum_{i=1}^n X_i, \prod_{i=1}^n X_i \right) \quad (8.39)$$

- c) Distribuzione Normale  $N(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$ : la funzione di verosimiglianza è

$$\begin{aligned} L(\mu, \sigma^2 | \mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right\} \mathbb{1}_{\mathbb{R}}(x_i) \\ &= \underbrace{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{2\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2}\right\}}_{g(\mu, \sigma^2; t_n)} \underbrace{\prod_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_i)}_{h(\mathbf{x})} \end{aligned}$$

per cui la statistica *congiuntamente* sufficiente per  $(\mu, \sigma^2)$  è data da

$$(T_{n;1}(\mathbf{X}), T_{n;2}(\mathbf{X})) = \left( \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right) \quad (8.40)$$

Ora se  $\mu$  è *noto*, la statistica sufficiente per  $\sigma^2$  è data da

$$T_n(\mathbf{X}) = \sum_{i=1}^n (X_i - \mu)^2$$

mentre se  $\sigma^2$  è *noto*, la statistica sufficiente per  $\mu$  è restituita da

$$T_n(\mathbf{X}) = \sum_{i=1}^n X_i.$$

Quanto visto poc'anzi a proposito della distribuzione Gamma e di quella Normale, vale in generale: il concetto di sufficienza è *problem dependent* ovvero varia al variare dell'informazione disponibile sulla distribuzione generante il campione. Questa particolarità si rivela essere un punto di forza che conferisce flessibilità all'inferenza basata su statistiche sufficienti.

**Esempio 8.2.5.** Riprendiamo l'esempio 8.2.3 dove avevamo verificato, via Definizione 8.2.1, che  $X_{(1)} = \min(X_1, X_2, \dots, X_n)$  era in quel caso statistica sufficiente per  $\theta$ . Ora, disponendo del teorema di fattorizzazione le cose sono molto più semplici; basta infatti osservare che

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n e^{-(x_i-\theta)} \mathbb{1}_{[\theta, +\infty)}(x_i) = e^{\sum_{i=1}^n (x_i-\theta)} \prod_{i=1}^n \mathbb{1}_{[\theta, +\infty)}(x_i), \quad \theta > 0 \\ &= e^{\sum_{i=1}^n x_i + n\theta} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \mathbb{1}_{[0, x_{(1)}]}(\theta), \quad \theta > 0 \\ &= e^{\sum_{i=1}^n x_i} \underbrace{\prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i)}_{h(\mathbf{x})} \underbrace{e^{n\theta} \mathbb{1}_{[0, x_{(1)}]}(\theta)}_{g(\theta; t_n)} \end{aligned} \quad (8.41)$$

e dunque si conclude che  $T_n(\mathbf{X}) = X_{(1)} = \min(X_1, X_2, \dots, X_n)$  è statistica sufficiente per  $\theta$ .

Vediamo ora quest'altro esempio che riserverà qualche sorpresa.

**Esempio 8.2.6.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione  $U_\theta(\theta; \theta + 1)$  quindi con funzione densità

$$f_{X_i}(x_i; \theta) = \begin{cases} 1 & \text{se } \theta \leq x_i \leq \theta + 1 \\ 0 & \text{altrimenti} \end{cases} \quad (8.42)$$

Calcoliamone la funzione di verosimiglianza

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n \mathbb{1}_{(\theta, \theta+1)}(x_i) = \prod_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_i) \mathbb{1}_{(\theta, \theta+1)}(x_{(n)}) \mathbb{1}_{\{\theta, \theta+1\}}(x_{(1)}) = \\ &= \mathbb{1}_{(x_{(n)}-1, x_{(1)})}(\theta) \prod_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_i) = g(\theta; x_{(1)}, x_{(n)}) h(\mathbf{x}) \end{aligned} \quad (8.43)$$

sicché la statistica sufficiente per  $\theta$  in questo caso sarà data da  $(X_{(1)}, X_{(n)})$  e si può immediatamente notare che non corrisponde alla dimensione del parametro  $\theta$  (scalare) mentre questa è un vettore bidimensionale. Questa particolarità è dovuta alla non regolarità della famiglia delle distribuzioni Uniformi. Avremo modo di tornare sull'argomento.

Nel paragrafo 1.4 del secondo capitolo abbiamo introdotto un'importante famiglia di distribuzioni, la *famiglia esponenziale* a  $k$ -parametri ne abbiamo sottolineato alcune *proprietà di regolarità*. Una di queste era relativa al fatto che in presenza di un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da uno dei membri di una questa famiglia, erano sufficienti  $k$  quantità, cui avevamo dato il nome di *statistiche naturali*, per riassumere tutta l'informazione riguardo  $\theta$  contenuta nel campione (e quindi anche riguardo la distribuzione da cui proviene il campione visto che ci stiamo muovendo nell'ambito dell'approccio parametrico alla statistica).

Ebbene, le statistiche naturali altro non sono che statistiche sufficienti e il seguente teorema fornisce il legame tra *appartenenza a famiglia esponenziale a  $k$ -parametri* e *espressione analitica* delle statistiche sufficienti.

**Teorema 8.2.2** (Famiglie esponenziali e statistiche sufficienti). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione che ammette funzione di massa (o di densità)  $F_X(x; \theta)$ , con  $\theta \in \Theta$  e che appartiene a una famiglia esponenziale a  $k$  parametri, ossia con funzione di densità

$$f_X(x; \theta) = C(x)D(\theta) \exp \left\{ \sum_{m=1}^k \left[ A_m(\theta)B_m(x) \right] \right\} \mathbb{1}_{S_X}(x) \quad (8.44)$$

Allora la statistica

$$\begin{aligned} \mathbf{T}_n(\mathbf{X}) &= (T_1(\mathbf{X}), T_2(\mathbf{X}), \dots, T_k(\mathbf{X})) \\ &= \left( \sum_{i=1}^n B_1(X_i), \sum_{i=1}^n B_2(X_i), \dots, \sum_{i=1}^n B_k(X_i) \right) \end{aligned} \quad (8.45)$$

è congiuntamente sufficiente per  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$

*Dimostrazione.* La funzione di verosimiglianza corrispondente a una distribuzione appartenente a famiglia esponenziale a  $k$ -parametri è data da:

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n f_{x_i}(x_i; \theta) = C^*(\mathbf{x})D^*(\theta) \exp \left\{ \sum_{m=1}^k A_m(\theta) \sum_{i=1}^n B_m(x_i) \right\} \prod_{i=1}^n \mathbb{1}_{S_X}(x_i) \quad (8.46)$$

È immediato osservare che  $L(\theta | \mathbf{x}) = g(\theta, \mathbf{t}_n) \cdot h(\mathbf{x})$ , pertanto, per il teorema di fattorizzazione di Neyman, la statistica  $\mathbf{T}_n(\mathbf{X})$  è congiuntamente sufficiente per  $\theta = (\theta_1, \dots, \theta_n)$ .  $\square$

Tutti gli esempi che abbiamo finora trattato nell'ambito del teorema di fattorizzazione di Neyman hanno inequivocabilmente mostrato che quest'ultimo, a differenza della Definizione 8.2.1, è *costruttivo* nel senso che restituisce l'*espressione analitica* della statistica sufficiente senza che nulla su quest'ultima debba essere ipotizzato.

E' comunque immediato osservare che la statistica sufficiente *non è unica*: ogni trasformazione biunivoca di statistica sufficiente è anch'essa sufficiente dal momento che ne condivide con la prima la partizione indotta sullo spazio campionario  $\mathfrak{X}$ .

La *non unicità* delle statistiche sufficienti solleva immediatamente un problema di scelta tra le (infinite) alternative possibili; ovviamente, la scelta ottimale ricadrà sulla statistica caratterizzata dal *massimo grado di sintesi* compatibile con il non dover rinunciare ad alcuna delle informazioni rilevanti per la comprensione del problema di interesse: una tale statistica viene detta *sufficiente minimale*.

Formalmente potremmo anche dire che una statistica  $T_n(\mathbf{X})$  è detta *sufficiente minimale* se, per qualsiasi statistica sufficiente  $V_n(\mathbf{X})$  esiste una qualche funzione  $s$  tale che  $T_n = s(V_n)$  ovvero una statistica sufficiente minimale è funzione di ogni altra statistica sufficiente. E così dicendo, vogliamo rimarcare il fatto che la partizione indotta su  $\mathfrak{X}$  dalla statistica sufficiente minimale  $T_n$  è la *più grossolana possibile* compatibilmente con il non perdere informazione rilevante per il problema inferenziale trattato.

In altre parole, una statistica *sufficiente* ma *non minimale* contiene informazioni *superflue* sul parametro  $\theta$ , inutili al fine di ricostruire la legge congiunta del campione casuale  $f_{\mathbf{X}}(\mathbf{x}; \theta)$ . Per contro, una statistica *sufficiente minimale* fornisce la minima informazione necessaria per ricostruire il modello statistico (=fare inferenza su  $\theta$ ) a partire dai dati.

La definizione che segue tiene conto di tutto ciò.

**Definizione 8.2.2** (Statistica sufficiente minimale). Una statistica  $T_n(X_1, X_2, \dots, X_n)$  è detta *sufficiente minimale* per  $\theta$  se

- a) è sufficiente per  $\theta$
- b) assume valori distinti solamente in punti dello spazio campionario  $\mathfrak{X}$  a cui corrispondono verosimiglianze non equivalenti cioè se, per ogni  $\mathbf{x}_1, \mathbf{x}_2 \in \mathfrak{X}$ ,

$$T_n(\mathbf{x}_1) \neq T_n(\mathbf{x}_2) \iff L(\theta|\mathbf{x}_1) \neq L(\theta|\mathbf{x}_2) \quad (8.47)$$

Per convincersi che una statistica di questo tipo esiste, si consideri la partizione indotta dalla funzione di verosimiglianza, detta per l'appunto, *partizione di verosimiglianza*, i cui elementi sono gli insiemi formati dai punti  $\mathbf{x} \in \mathfrak{X}$  che conducono a verosimiglianze *equivalenti*: una qualsiasi statistica che assume valore costante sullo stesso elemento della partizione di verosimiglianza e valori distinti su elementi distinti di tale partizione è per costruzione *sufficiente minimale*.

Notiamo anche che una statistica sufficiente è minimale se induce su  $\mathfrak{X}$  una partizione che coincide con quella introdotta dalla funzione di verosimiglianza; inoltre, essa è *unica*, a meno di trasformazioni biunivoche di essa e che comunque condividono la medesima *partizione di verosimiglianza*.

**Esempio 8.2.7.** Consideriamo un campione casuale a  $n = 4$  elementi,  $\mathbf{X} = (X_1, X_2, X_3, X_4)$ , estratto da una distribuzione di Bernoulli,  $b(1, \theta)$ ; sulla base di questo, vogliamo stimare la probabilità di successo nella singola prova  $P_\theta(X = 1) = \theta$ . Consideriamo dunque le seguenti statistiche:

$$T_n(\mathbf{X}) = \sum_{i=1}^4 x_i \quad V_n(\mathbf{X}) = X_1 \quad W_n(\mathbf{X}) = (T_n, V_n) \quad (8.48)$$

Determiniamo ora la funzione di verosimiglianza

$$L(p; \mathbf{x}) = p^{\sum x_i} (1-p)^{n-\sum x_i} \prod_{i=1}^4 \mathbb{1}_{\{0,1\}}(x_i) \quad (8.49)$$

per capire, alla luce del teorema di fattorizzazione di Neymann e della partizione (indotta dalla funzione) di verosimiglianza, se tali statistiche siano sufficienti, non lo siano o siano sufficienti minimali.

Appare evidente che  $V_n(\mathbf{X})$  non sia una statistica *sufficiente* e che  $T_n(\mathbf{X})$  sia una statistica sufficiente e anche *minimale* in quanto induce sullo spazio campionario  $\mathfrak{X}$  una partizione che coincide con quella di verosimiglianza. Diversamente,  $W_n(\mathbf{X})$  è certamente statistica sufficiente ma *non minimale*, come è facile verificare con un controsenso: dati i due campioni  $(0, 1, 1, 0)$  e  $(1, 0, 1, 0)$ , aventi medesima verosimiglianza, in loro corrispondenza la statistica  $W_n$  assume due valori differenti, rispettivamente  $(2, 0)$  e  $(2, 1)$ .

Ciò si verifica facilmente anche confrontando le partizioni dello spazio campionario indotte da  $T_n$  e da  $W_n$  rispettivamente. In particolare, notiamo che  $T_n$  divide lo spazio campionario in 5 sottospazi, ciascuno caratterizzato da un diverso numero di successi. Invece,  $W_n$  divide lo spazio campionario in 10 sottospazi, in quanto, fissato il numero di successi, tale statistica distingue se il primo elemento del campione sia associato o meno a un successo.

Notiamo inoltre che  $T_n$  e  $U_n = T_n/n$  inducono partizioni equivalenti (e, in questo caso coincidenti con quella di verosimiglianza) essendo l'una trasformata biunivoca dell'altra.

Individuare una statistica sufficiente minimale sfruttando la definizione che di essa abbiamo poc'anzi dato, non è cosa agevole (non è "costruttiva" come abbiamo avuto modo di vedere). Ci può allora venire in aiuto il seguente teorema:

**Teorema 8.2.3** (Teorema di Lehmann-Sheffé). Sia  $(X_1, \dots, X_n)$  un campione casuale da una distribuzione avente funzione di massa (o di densità)  $f_X(x; \theta)$ , con  $\theta \in \Theta \subseteq \mathbb{R}^k$ . Supponiamo esista una funzione  $T_n(\mathbf{x}) = T_n(x_1, x_2, \dots, x_n)$  tale che per due punti  $\mathbf{x}$  e  $\mathbf{y}$  di  $\mathfrak{X}$  il rapporto (di verosimiglianza)

$$\frac{L(\theta|\mathbf{x})}{L(\theta|\mathbf{y})} = \frac{f_X(\mathbf{x}; \theta)}{f_X(\mathbf{y}; \theta)} \quad (8.50)$$

non dipenda da  $\theta$  se e solo se  $T(\mathbf{x}) = T(\mathbf{y})$ . Allora  $T(\mathbf{X})$  è statistica sufficiente *minimale* per  $\theta$ .

*Dimostrazione.* Per semplicità supponiamo che  $f_X(x; \theta) > 0$  per ogni  $\mathbf{x} \in \mathfrak{X}$  e per ogni  $\theta \in \Theta$ . Dimostreremo prima che la statistica  $T_n(\mathbf{X})$  è una statistica sufficiente per  $\theta$  e poi che essa è anche minimale.

Per ogni valore  $t$  assunto da  $T_n$  sia  $A_t = \{\mathbf{x} \in \mathfrak{X} : T(\mathbf{x}) = t\}$  l'elemento della partizione indotta da  $T(\mathbf{X})$  su  $\mathfrak{X}$ . In particolare per ogni  $\mathbf{x}, \mathbf{y} \in A_t$  si ha  $T_n(\mathbf{x}) = T_n(\mathbf{y})$  e per ipotesi il rapporto

$$\frac{f_X(\mathbf{x}; \theta)}{f_X(\mathbf{y}; \theta)} = \frac{L(\theta|\mathbf{x})}{L(\theta|\mathbf{y})}$$

non dipende da  $\theta$ . Ora per  $\mathbf{x} \in \mathfrak{X}$  si consideri  $\mathbf{y} \in A_{T_n(\mathbf{x})}$  e la funzione

$$h(\mathbf{x}) = \frac{f_X(\mathbf{x}; \theta)}{f_X(\mathbf{y}(\mathbf{x}); \theta)}$$

che non dipende da  $\theta$  per costruzione; per la precedente uguaglianza si può anche scrivere

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}; \theta) &= h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{y}(\mathbf{x}); \theta) = \frac{f_{\mathbf{X}}(\mathbf{x}; \theta)}{f_{\mathbf{X}}(\mathbf{y}(\mathbf{x}); \theta)} f_{\mathbf{X}}(\mathbf{y}(\mathbf{x}); \theta) \\ &= \frac{f_{\mathbf{X}}(\mathbf{x}; \theta)}{f_{\mathbf{X}}(\mathbf{y}; \theta)} f_{\mathbf{X}}(\mathbf{y}; \theta) \\ &= h(\mathbf{x}) g(T_n(\mathbf{x}); \theta) \end{aligned} \quad (8.51)$$

dove l'ultima uguaglianza segue dall'arbitrarietà di  $\mathbf{y}$  in  $A_{t_n}$ . Ma allora

$$L(\theta | \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \theta) = h(\mathbf{x}) g(T_n(\mathbf{x}); \theta)$$

e in virtù del teorema di fattorizzazione di Neyman la statistica  $T_n(\mathbf{X})$  è sufficiente per  $\theta$ .

Dimostriamo ora la minimalità della statistica sufficiente  $T(\mathbf{X})$ . A questo scopo, sia  $T^*(\mathbf{X})$  un'altra statistica sufficiente per  $\theta$  per cui, in virtù del teorema di fattorizzazione di Neyman, dovranno esistere due funzioni  $h^*$  e  $g^*$ , la prima funzione della sola  $\mathbf{x}$  e la seconda di  $\theta$  e  $T^*(\mathbf{x})$ , tali che

$$L(\theta | \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \theta) = h^*(\mathbf{x}) g^*(T^*(\mathbf{x}); \theta).$$

Per  $\mathbf{x}, \mathbf{y} \in \mathfrak{X}$  e poichè  $T^*$  è sufficiente per  $\theta$  si ha

$$\frac{f_{\mathbf{X}}(\mathbf{x}; \theta)}{f_{\mathbf{X}}(\mathbf{y}; \theta)} = \frac{h^*(\mathbf{x}) g^*(T^*(\mathbf{x}); \theta)}{h^*(\mathbf{y}) g^*(T^*(\mathbf{y}); \theta)} = \frac{h^*(\mathbf{x})}{h^*(\mathbf{y})}$$

per ogni  $\mathbf{x}$  e  $\mathbf{y}$  tali che  $T^*(\mathbf{x}) = T^*(\mathbf{y})$ . Quest'ultimo rapporto evidentemente non dipende da  $\theta$  e per ipotesi questo avviene se e solo se  $T(\mathbf{x}) = T(\mathbf{y})$  da cui  $T$  è funzione di  $T^*$ .  $\square$

Vediamo il seguente esempio.

**Esempio 8.2.8.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione Gamma  $\mathcal{G}(\alpha, \beta)$ ,  $\alpha > 0, \beta > 0$  e calcoliamo la corrispondente funzione di verosimiglianza:

$$L(\alpha, \beta; \mathbf{x}) = \left( \frac{1}{\Gamma(\alpha)\beta^\alpha} \right)^n \left[ \prod_{i=1}^n x_i \right]^{\alpha-1} \exp \left\{ -\frac{1}{\beta} \sum_{i=1}^n x_i \right\} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \quad (8.52)$$

Consideriamo la statistica sufficiente  $T_n = (\prod_{i=1}^n X_i, \sum_{i=1}^n X_i)$  e determiniamo se essa è minimale verificando che essa soddisfi le ipotesi del teorema 8.2.3. Ora,

$$\frac{L(\alpha, \beta | \mathbf{x})}{L(\alpha, \beta | \mathbf{y})} = \exp \left\{ -\frac{1}{\beta} \left( \sum_{i=1}^n x_i - \sum_{i=1}^n y_i \right) \right\} \left( \prod_{i=1}^n \frac{x_i}{y_i} \right)^{\alpha-1} \quad (8.53)$$

Tale rapporto non dipende da  $(\alpha, \beta)$  se (e solo se) sono soddisfatte le condizioni

$$\begin{cases} \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ \prod_{i=1}^n x_i = \prod_{i=1}^n y_i \end{cases} \quad (8.54)$$

Pertanto

$$\mathbf{T}_n(X_1, X_2, \dots, X_n) = \left( \sum_{i=1}^n X_i, \prod_{i=1}^n X_i \right) \quad (8.55)$$

è statistica congiuntamente sufficiente *minimale*.

Notiamo subito che, se utilizziamo il *metodo dei momenti* (vedi paragrafo 2.18), la stima dei parametri  $(\alpha, \beta)$  della distribuzione Gamma si otterrà quale soluzione del seguente sistema di equazioni ottenuto eguagliando rispettivamente il primo e il secondo momento non centrato della popolazione con gli equivalenti campionari

$$\begin{cases} \alpha\beta = \frac{1}{n} \sum_{i=1}^n X_i \\ \alpha\beta^2 + \alpha^2\beta^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases} \quad (8.56)$$

Ma così facendo perdiamo parte delle informazioni su  $(\alpha, \beta)$  contenute nel campione poiché nelle equazioni di stima (8.56) non viene coinvolta la statistica sufficiente minimale. Di questo fatto si pagherà prezzo in termini di *affidabilità* delle stime.

**Esempio 8.2.9.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente dalla distribuzione  $N(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$ ,  $\sigma^2 > 0$ . Ricordando che la distribuzione in questione appartiene a famiglia esponenziale a  $k = 2$ -parametri, dall'espressione analitica (8.40) dell'associata funzione di verosimiglianza, si ha immediatamente che

$$\mathbf{T}_n = \left( \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right) \quad (8.57)$$

è statistica sufficiente e anche minimale.

Il Teorema 8.2.2 stabilisce un'importante risultato che lega famiglie esponenziali a  $k$ -parametri e statistiche (congiuntamente) sufficienti per  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ , fornendone la seguente espressione analitica

$$\mathbf{T}_n(\mathbf{X}) = \left( \sum_{i=1}^n B_1(X_i), \sum_{i=1}^n B_2(X_i), \dots, \sum_{i=1}^n B_k(X_i) \right) \quad (8.58)$$

Grazie al Teorema 8.2.3 di Lehmann-Scheffé è immediato dimostrare che la statistica (congiuntamente) sufficiente  $\mathbf{T}_n(\mathbf{X})$  è anche *minimale*. La dimostrazione è lasciata per esercizio.

Una domanda: l'esistenza di statistica sufficiente non banale (si ricorda che lo stesso campione casuale è una statistica sufficiente sebbene lunghi, in molti casi, dall'essere minimale) *implica* che la distribuzione generante il campione appartenga a famiglia esponenziale? **La risposta è no!**

Basti pensare alla distribuzione Uniforme su  $(0, \theta)$ ,  $\theta > 0$  di cui sappiamo che

$$T_n = X_{(n)} = \max(X_1, X_2, \dots, X_n) \quad (8.59)$$

è statistica sufficiente minimale ma ciò non implica in alcun modo che la famiglia generante  $(X_1, X_2, \dots, X_n)$  appartenga a famiglia esponenziale.

# 9 Methods of Evaluating Estimators

I concetti di verosimiglianza e di sufficienza (minimale) che abbiamo precedentemente introdotto possono essere spesi nell'individuare metodi di costruzione di stimatori ottimali.

## 9.1 Viaggio alla ricerca dello stimatore migliore

Abbiamo già avuto modo di definire l'ottimalità in termini di minimo richio e abbiamo formalizzato l'argomento introducendo il concetto di errore quadratico medio che qui brevemente richiamiamo.

**Definizione 9.1.1** (Errore Quadratico Medio). Se  $T_n(X_1, X_2, \dots, X_n)$  è uno stimatore del parametro  $\theta$ , diremo *Errore Quadratico Medio* di  $T_n$  la seguente funzione di  $\theta \in \Theta$

$$MSE_\theta(T_n) = \mathbb{E}_\theta [(T_n - \theta)^2] = \text{Var}_\theta(T_n) + \mathbb{B}_\theta^2(T_n) \quad (9.1)$$

dove  $T_n(X_1, X_2, \dots, X_n)$  rappresenta la regola di decisione mentre  $\theta$  rappresenta l'incognito *stato della natura*.

Per quanto già visto,  $MSE_\theta(T_n)$ , noto anche come funzione di rischio (o di perdita attesa) fornisce una misura della *qualità* della funzione di decisione  $T_n$  laddove si è deciso di adottare una funzione di perdita quadratica  $(T_n - \theta)^2$ .

Alla luce di tutto ciò, potremmo pensare di scegliere, nell'insieme dei molti (potenzialmente infiniti) competitori, lo stimatore  $T_n$  con rischio *uniformemente* più piccolo; vale a dire, sia  $V_n$  un qualsiasi competitor di  $T_n$ ,

$$MSE_\theta(T_n) \leq MSE_\theta(V_n), \quad \forall \theta \in \Theta. \quad (9.2)$$

Ora, trovare uno stimatore  $T_n$  di  $\theta \in \Theta$  con errore quadratico medio *uniformemente minimo* nella classe di *tutti* gli stimatori di  $\theta$  non è cosa possibile.

Infatti, sia  $U_n$  uno stimatore di  $\theta \in \Theta$  così definito

$$U_n(X_1, X_2, \dots, X_n) = \theta_0, \quad \text{con } \theta_0 \in \Theta, \text{ valore fissato.} \quad (9.3)$$

E' evidente che non ci si può aspettare gran che da un siffatto stimatore che ignora completamente l'informazione contenuta nel campione; ma poiché

$$MSE_\theta(U_n) = \mathbb{E} [(U_n - \theta)^2] = 0 \quad (9.4)$$

quando  $\theta = \theta_0$ , nessun altro stimatore  $T_n(X_1, X_2, \dots, X_n)$  potrà avere errore quadratico medio uniformemente più piccolo di quello di  $U_n$  a meno che non abbia

errore quadratico medio uguale a zero ovunque (ossia, uniformemente nullo). Ma ciò è, in geneale, chiaramente impossibile.

La ragione del *fallimento* del criterio basato sull'errore quadratico medio uniformemente minimo poggia sulla *ampiezza* ed *eterogeneità* della classe di *tutti i* possibili stimatori di  $\theta$  che comprende, per esempio, sia stimatori non distorti che distorti; inoltre, l'errore quadratico medio è un criterio complesso perché coinvolge contemporaneamente aspetti relativi all'*accuratezza* (non distorsione) e alla *precisione* (varianza) degli stimatori e tutto ciò comporta che le curve dell'errore quadratico medio che si intersechino (non vi è uno stimatore che *domina* gli altri in termini di rischio).

L'idea più semplice e naturale è di provare a *restringere* la classe degli stimatori di  $\theta$  ai soli *stimatori non distorti* di  $\theta$  (o di funzioni di  $\theta$ ,  $a(\theta)$ ); in questo caso,

$$MSE_\theta(T_n) = \mathbb{E}_\theta [(T_n - \theta)^2] = \text{Var}_\theta(T_n) \quad (9.5)$$

e

$$\min_{\theta \in \Theta} MSE_\theta(T_n) = \min_{\theta \in \Theta} \text{Var}_\theta(T_n) \quad (9.6)$$

Possiamo subito osservare che quest'ultima *restrizione* mette definitivamente fuori gioco lo stimatore  $U_n = \theta_0$  poiché esso non soddisfa a

$$\mathbb{E}_\theta(U_n) = \theta, \quad \forall \theta \in \Theta \quad (9.7)$$

essendo  $U_n$  stimatore *distorto* di  $\theta$ .

Affinchè un criterio di scelta dello stimatore, pur nella classe degli stimatori non distorti per  $\theta$  produca effetti conclusivi, dovremo ora rispondere a due domande di cruciale importanza quali quelle che seguono.

- 1) Esiste un *limite inferiore* per la varianza di uno stimatore non distorto  $T_n$  di  $\theta$ ?
- 2) Se sì, la varianza dello stimatore non distorto  $T_n$  lo raggiunge? E in caso contrario, di quanto se ne discosta?

Prima di fornire una risposta formale alle due domande, consideriamo il seguente esempio.

**Esempio 9.1.1.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da un distribuzione di Poisson di parametro  $\theta > 0$  e siano

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{e} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (9.8)$$

rispettivamente media e varianza campionarie. Come già sappiamo, la distribuzione di Poisson gode della seguente proprietà caratterizzante

$$\mathbb{E}_\theta(X) = \text{Var}_\theta(X) = \theta \quad (9.9)$$

e perciò, essendo  $\mathbb{E}_\theta(\bar{X}_n) = \theta$  e  $\mathbb{E}_\theta(S_n^2) = \theta$ , sia  $\bar{X}_n$  che  $S_n^2$  sono stimatori *non distorti* del parametro  $\theta$ .c

La scelta tra  $\bar{X}_n$  e  $S_n^2$  può essere fatta in base al *confronto* delle loro varianze e si può dimostrare che

$$\text{Var}_\theta(\bar{X}_n) = \frac{\theta}{n} < \frac{\theta}{n} \left(1 + \frac{2n\theta}{n-1}\right) = \text{Var}_\theta(S_n^2) \quad (9.10)$$

essendo  $\frac{2n\theta}{n-1} > 0$  (Teorema 2.1.3 per ricavare la varianza di  $S_n^2$ )

Constatato ciò, consideriamo ora la classe  $\mathcal{W}_{n,a}$  di stimatori di  $\theta$  dati da

$$W_{n,a} = a \bar{X}_n + (1-a) S_n^2, \quad a \in [0, 1] \quad (9.11)$$

e possiamo subito osservare che

$$\mathbb{E}_\theta(W_{n,a}) = \theta, \quad \forall \theta > 0 \quad (9.12)$$

quali che siano  $n$  e la costante  $a \in [0, 1]$ , sicché avremo un'infinità (non numerabile) di nuovi stimatori *non distorti* di  $\theta$ , costruiti a partire da combinazioni lineari di  $\bar{X}_n$  e  $S_n^2$ .

È dunque lecito chiedersi:

- a) ammettendo che  $\bar{X}_n$  sia *migliore* di  $S_n^2$  perché *meno rischioso*, esso è anche *migliore* di ogni altro stimatore elemento della classe  $\mathcal{W}_{n,a}$ ?
- b) come possiamo essere sicuri che non esistano altri stimatore non distorti di  $\theta$  *migliori* di quelli fin qui considerati?

Rispondere alle precedenti due domande richiede di trovare una risposta alle domande 1) e 2) poste a pagina precedente. Dobbiamo dunque spendere un po' di tempo (e fatica), e lo faremo nel paragrafo seguente, per trovare un *limite inferiore* per la *varianza* di un qualsiasi stimatore *non distorto* di  $\theta$ .

## 9.2 Efficienza e stimatori efficienti

Prima di procedere diamo la seguente definizione.

**Definizione 9.2.1** (Condizioni di regolarità). Una famiglia di distribuzioni

$$\mathcal{F}_\theta = \{F_X(x; \theta), \theta \in \Theta \subseteq \mathbb{R}^k\} \quad (9.13)$$

è detta *regolare* se:

- a) a diversi valori del parametro  $\theta$  corrispondono distribuzioni *distinte*

$$\theta \neq \theta' \Rightarrow F_X(x; \theta) \neq F_X(x; \theta') \quad (9.14)$$

ovvero vale la *condizione di identificabilità*: i (valori dei) parametri *identificano* le distribuzioni.

- b) Le distribuzioni della medesima famiglia hanno *supporto comune* per ogni  $\theta \in \Theta \subseteq \mathbb{R}^k$ ; ovvero, il supporto  $\text{supp } X$  della distribuzione  $F_X(x; \theta)$  non dipende da  $\theta$ .
- c) laddove la distribuzione sia continua la densità corrispondente a  $F_X(x; \theta)$ ,  $f_X(x; \theta)$  è *almeno due volte derivabile* rispetto a  $\theta$ . Ciò permette di espandere in serie di Taylor la funzione di verosimiglianza.

d) L'integrale

$$\int_{S_X} f_X(x; \theta) dx \quad (9.15)$$

può essere derivato almeno due volte sotto il segno di integrale rispetto a  $\theta$ . Ciò permette di scambiare nell'ordine integrazione e derivazione della funzione di densità rispetto a  $\theta$ .

Il teorema che segue, noto come *disuguaglianza di Rao-Cramér*, gioca un ruolo chiave nella risposta alle due domande con cui abbiamo chiuso il precedente paragrafo poiché fornisce quel limite inferiore della varianza di un qualsiasi stimatore non distorto di  $\theta$  che stiamo cercando e sul quale poggiano le risposte alle due menzionate domande.

Ne daremo enunciato e la dimostrazione assumendo, per il momento, che  $\theta$  sia uno scalare; proveremo poi a generalizzare il risultato a  $\theta$  vettore di dimensione qualsiasi o a funzioni di  $\theta$ ,  $\tau(\theta)$ .

**Teorema 9.2.1** (Disuguaglianza di Rao-Cramér). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione avente funzione di massa (o di densità)  $f(x; \theta)$  che soddisfa alle condizioni di regolarità poc'anzi elencate. Allora, per un qualsiasi stimatore non distorto  $T_n$  di  $\theta$  vale

$$\text{Var}_\theta(T_n) \geq \frac{1}{I_n(\theta)} \quad (9.16)$$

dove

$$I_n(\theta) = \mathbb{E}_\theta \left[ \frac{d}{d\theta} \ln L(\theta | \mathbf{x}) \right]^2 = \mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{x}) \right]^2 \quad (9.17)$$

è nota come *informazione di Fisher* dell'intero campione (ovvero contenuta nella realizzazione  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ ).

Prima di dare la dimostrazione del teorema, osserviamo che:

- a) la quantità  $1/I_n(\theta)$  è il limite inferiore della varianza di un qualsiasi stimatore non distorto di  $\theta$
- b) si può dimostrare che

$$\mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{x}) \right]^2 = -\mathbb{E}_\theta \left[ \frac{d^2}{d\theta^2} \ell(\theta | \mathbf{x}) \right] \quad (9.18)$$

e, spesso, il secondo membro della identità (9.18) è più semplice da calcolare di quanto non sia il primo; la dimostrazione della precedente identità poggia in maniera cruciale sulle proprietà di regolarità e in particolare sul punto d) delle condizioni di regolarità.

- c) più la varianza di  $T_n$  si avvicina a  $1/I_n(\theta)$  è più grande e significativa (leggi, informativa) risulta la sintesi dell'informazione contenuta nel campione casuale in merito a  $\theta$  operata dallo stimatore non distorto  $T_n$ .

E' bene osservare che

$$I_n(\theta) = -\mathbb{E}_\theta \left[ \frac{d^2}{d\theta^2} \ell(\theta | \mathbf{x}) \right] \quad (9.19)$$

fornisce una *misura della intensità della curvatura* della funzione di log-verosimiglianza  $\ell(\theta|\mathbf{x})$  in un intorno di  $\theta$ ; in particolare, più accentuata è la curvatura di  $\ell(\theta|\mathbf{x})$  e più concentrata sarà quest'ultima intorno a  $\theta$  e, di conseguenza, più preciso sarà lo stimatore  $T_n$ .

Potremmo anche osservare che  $I_n(\theta)$  fornisce una *misura dell'efficienza* di  $T_n$  in termini di *capacità di far sintesi informativa* riguardo a  $\theta$ .

e) spesso si scrive

$$I_n(\theta) = n I_1(\theta) \quad (9.20)$$

a sottolineare il fatto che l'informazione contenuta nell'*intero campione* riguardo  $\theta$  che abbiamo detto essere  $I_n(\theta)$  e usualmente sintetizzata da una statistica sufficiente, altro non è che la *somma* dei *contributi informativi* riconducibili a ciascuna delle componenti  $X_i$  del campione casuale. Infatti si può facilmente dimostrare che se  $(X_1, X_2, \dots, X_n)$  è una  $n$ -pla di v.c. *indipendenti* allora

$$I_n(\theta) = \sum_{i=1}^n I_{X_i}(\theta_i) \quad (9.21)$$

che si riduce a

$$I_n(\theta) = \sum_{i=1}^n I_{X_1}(\theta) = n I_1(\theta) \quad (9.22)$$

se oltre che essere *indipendenti* le  $n$  v.c. sono anche *identicamente distribuite*. Inoltre vale anche il seguente risultato.

**Teorema 9.2.2.** L'informazione di Fisher fornita da uno stimatore  $T_n$  basato su una statistica sufficiente  $W_n$  coincide con quella fornita dall'*intero campione*.

*Dimostrazione.* Per il teorema di fattorizzazione di Neyman, l'informazione di Fisher sarà funzione del campione casuale *solo* tramite

$$\ln g(\theta; W_n(\mathbf{x})). \quad (9.23)$$

Infatti

$$\ln L(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x}) = \ln[h(\mathbf{x}) \cdot g(\theta; W_n(\mathbf{x}))] = \ln h(\mathbf{x}) + \ln g(\theta; W_n(\mathbf{x})) \quad (9.24)$$

ma

$$\frac{d}{d\theta} \left[ \frac{d}{d\theta} \ell(\theta|\mathbf{x}) \right] = \frac{d}{d\theta} \left[ \frac{d}{d\theta} \ln g(\theta; W_n(\mathbf{x})) \right] \quad (9.25)$$

sicché

$$I_n(\theta) = I_{\mathbf{X}}(\theta) = -E \left[ \frac{d^2}{d\theta^2} \ell(\theta|\mathbf{x}) \right] \equiv -E \left[ \frac{d^2}{d\theta^2} \ln g(\theta; W_n(\mathbf{X})) \right] = I_{W_n(\mathbf{X})}(\theta)$$

□

**Esempio 9.2.1.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione Normale  $N(\mu, \sigma^2)$  con  $\sigma^2$  noto, per cui

$$I_n(\mu) = -\mathbb{E}_\mu \left[ \frac{d^2}{d\mu^2} \ell(\mu | \mathbf{X}) \right] = -\mathbb{E}_\mu \left[ -\frac{n}{\sigma^2} \right] = \frac{n}{\sigma^2} \quad (9.26)$$

Consideriamo ora la statistica *sufficiente* per  $\mu$  data da  $W_n = \sum_{i=1}^n X_i$  e sia

$$T_n = \frac{W_n}{n} = \bar{X}_n \sim N(\mu, \sigma^2/n) \quad (9.27)$$

lo stimatore di  $\mu$ . Si mostra facilmente che

$$-\mathbb{E}_\mu \left[ \frac{d^2}{d\mu^2} \ell(\mu | \bar{X}_n) \right] = -\mathbb{E}_\mu \left[ -\frac{n}{\sigma^2} \right] = \frac{n}{\sigma^2} = I_n(\mu) \quad (9.28)$$

Veniamo ora alla **dimostrazione del Teorema 9.2.1** (*Disuguaglianza di Rao-Cramér*) che si rivelerà essere una semplice ed elegante applicazione della disuguaglianza di Cauchy-Schwarz.

*Dimostrazione.* Supponiamo che  $T_n$  sia una variabile casuale continua (la dimostrazione per il caso discreto è analoga). Essendo  $T_n$  uno stimatore non distorto di  $\theta$ , vale che

$$\theta = E_\theta(T_n) = \int_{\mathbb{R}^n} T_n(\mathbf{x}) \cdot f_{\mathbf{X}}(\mathbf{x}; \theta) d\mathbf{x} \quad (9.29)$$

Derivando entrambi i membri rispetto a  $\theta$  e scambiando nell'ordine derivate e integrali, si ottiene

$$1 = \int_{\mathbb{R}^n} T_n(\mathbf{x}) \left( \frac{d}{d\theta} f_{\mathbf{X}}(\mathbf{x}; \theta) \right) d\mathbf{x} \quad (9.30)$$

Ora

$$\frac{d}{d\theta} L(\theta | b\mathbf{x}) = L(\theta | b\mathbf{x}) \frac{d}{d\theta} \ln L(\theta | b\mathbf{x}) = L(\theta | b\mathbf{x}) \frac{d}{d\theta} \ell(\theta | b\mathbf{x}) \quad (9.31)$$

sicché sostituendo in (9.30) si ha

$$\begin{aligned} 1 &= \int_{\mathbb{R}^n} \left( T_n(\mathbf{x}) \cdot \frac{d}{d\theta} \ell(\theta | b\mathbf{x}) \right) L(\theta | b\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}_\theta \left[ T_n(\mathbf{x}) \cdot \frac{d}{d\theta} \ell(\theta | b\mathbf{x}) \right] \\ &= \text{Cov}_\theta \left[ T_n(\mathbf{x}), \frac{d}{d\theta} \ell(\theta | b\mathbf{x}) \right] \end{aligned} \quad (9.32)$$

Infatti se poniamo  $U = T_n(\mathbf{X})$  e  $V = \frac{d}{d\theta} \ell(\theta | b\mathbf{X})$ , tenuto conto del fatto che, in virtù delle proprietà di regolarità,

$$\mathbb{E}_\theta(V) = \int_{\mathbb{R}^n} \frac{d}{d\theta} \ell(\theta | b\mathbf{x}) L(\theta | b\mathbf{x}) d\mathbf{x} = \frac{d}{d\theta} \int_{\mathbb{R}^n} L(\theta | b\mathbf{x}) d\mathbf{x} = 0 \quad (9.33)$$

si ha

$$\begin{aligned}
 \mathbb{C}ov_{\theta}(U, V) &= \mathbb{E}_{\theta}((U - \mathbb{E}_{\theta}(U)) \cdot (V - \mathbb{E}_{\theta}(V))) \\
 &= E_{\theta}(U \cdot V) - E_{\theta}(U)E_{\theta}(V) \\
 &= E_{\theta}(U \cdot V) \quad \text{poichè } E_{\theta}(V) = 0 \\
 &= \int_{\mathbb{R}^n} \left[ T_n(\boldsymbol{x}), \frac{d}{d\theta} \ell(\theta | b\boldsymbol{x}) \right] L(\theta | \boldsymbol{x}) = 1
 \end{aligned} \tag{9.34}$$

Per la diseguaglianza di Cauchy-Schwarz, possiamo scrivere

$$\mathbb{C}ov_{\theta}^2(U, V) \leq \mathbb{V}ar_{\theta}(U) \cdot \mathbb{V}ar_{\theta}(V) \tag{9.35}$$

dove

$$\begin{aligned}
 \mathbb{V}ar_{\theta}(V) &= \mathbb{V}ar_{\theta}\left(\frac{d}{d\theta} \ell(\theta | \boldsymbol{x})\right) \\
 &= E_{\theta}(V^2) - [E_{\theta}(V)]^2 \\
 &= E_{\theta}(V^2) = E_{\theta}\left[\left(\frac{d}{d\theta} \ell(\theta | b\boldsymbol{x})\right)^2\right] = I_n(\theta)
 \end{aligned} \tag{9.36}$$

sicché si ha

$$1^2 \leq \mathbb{V}ar_{\theta}(T_n) \cdot I_n(\theta) \tag{9.37}$$

da cui

$$\mathbb{V}ar_{\theta}(T_n) \geq \frac{1}{I_n(\theta)} \tag{9.38}$$

e questo conclude la dimostrazione.  $\square$

**Esempio 9.2.2** (Continuazione). Nella prima parte dell'Esempio 9.1.1 abbiamo verificato che sia  $\bar{X}_n$  che  $S_n^2$  siano stimatori non distorti di  $\theta$  e confrontando le loro varianze

$$\mathbb{V}ar_{\theta}(\bar{X}_n) = \frac{\theta}{n} < \frac{\theta}{n} \left(1 + \frac{2n\theta}{n-1}\right) = \mathbb{V}ar_{\theta}(S_n^2) \tag{9.39}$$

abbiamo concluso che  $\bar{X}_n$  è da preferire a  $S_n^2$  in quanto ha *rischio inferiore*.

Osserviamo che, pur ammettendo che  $\bar{X}_n$  sia uno stimatore migliore rispetto a  $S_n^2$ , non abbiamo ancora la garanzia che esso sia anche migliore di qualsiasi altro stimatore  $W_{n,a} = a\bar{X}_n + (1-a)S_n^2$ . Inoltre come possiamo essere sicuri che non esista un altro stimatore di  $\theta$  che sia migliore di tutti quelli fin qui considerati?

Se applichiamo la diseguaglianza di Rao-Cramér, troviamo che

$$\begin{aligned}
 I_n(\theta) &= -E_{\theta}\left(\frac{d^2}{d\theta^2} \ell(\theta; \boldsymbol{x})\right) \\
 &= nI_1(\theta) \\
 &= n \left[ -E\left(\frac{d^2}{d\theta^2} \ln\left(\frac{e^{-\theta}\theta^{X_1}}{X_1!}\right)\right) \right] = \\
 &= n \left[ -E\left(-\frac{X_1}{\theta^2}\right) \right] \\
 &= \frac{n\theta}{\theta^2} = \frac{n}{\theta}
 \end{aligned} \tag{9.40}$$

da cui

$$\text{Var}_\theta(\bar{X}_n) = \frac{1}{I_n(\theta)} = \frac{\theta}{n} \quad (9.41)$$

sicché la media campionaria  $\bar{x}_n$  è lo stimatore la cui varianza raggiunge il limite inferiore di Rao-Cramér e pertanto lo stimatore a minima varianza nella classe degli stimatori non distorti (UMVU). Nessun altro stimatore non distorto di  $\theta$  può avere varianza più piccola di quella di  $\bar{X}_n$ . E questo risolve la duplice domanda con cui avevamo chiuso il paragrafo 9.1.

### 9.2.0.1 Efficienza (assoluta e relativa)

L'aver individuato l'espressione analitica del limite inferiore della varianza di un qualsiasi stimatore di  $\theta$  permette di costruire i concetti e le definizioni che seguono.

**Definizione 9.2.2** (Bahadur Efficiency). Sia  $T_n$  uno stimatore non distorto di  $\theta$ . Diremo *efficienza* di  $T_n$  la quantità

$$\text{eff}(T_n) = \frac{1/I_n(\theta)}{\text{Var}_\theta(T_n)} = \frac{1}{I_n(\theta)\text{Var}_\theta(T_n)} \in [0, 1] \quad (9.42)$$

Di conseguenza,

**Definizione 9.2.3** (Stimatore efficiente). Sia  $T_n$  stimatore non distorto di  $\theta$ . Diremo che  $T_n$  è *stimatore efficiente* per  $\theta$  se e solo se

$$\text{Var}_\theta(T_n) = \frac{1}{I_n(\theta)} \quad (9.43)$$

**Definizione 9.2.4** (Efficienza relativa). Dati due stimatori  $T_n$  e  $V_n$  non distorti di  $\theta$  definiamo *efficienza relativa* la quantità

$$\text{eff}(T_n, V_n) = \frac{\text{Var}_\theta(T_n)}{\text{Var}_\theta(V_n)} \quad (9.44)$$

E dunque,  $T_n$  è preferito a  $V_n$  se  $\text{eff}(T_n, V_n) < 1$ ; viceversa,  $V_n$  è preferito a  $T_n$  se  $\text{eff}(T_n, V_n) > 1$ .

**Definizione 9.2.5** (Stimatore asintoticamente efficiente). Sia  $T_n$  uno stimatore non distorto di  $\theta$ . Diremo che  $T_n$  è uno stimatore *asintoticamente efficiente* per  $\theta$  se

$$\lim_{n \rightarrow \infty} \text{Var}_\theta(T_n) = \frac{1}{I_n(\theta)} \quad (9.45)$$

ovvero, parleremo di *efficienza asintotica* di  $T_n$  se  $\lim_{n \rightarrow \infty} \text{eff}(T_n) = 1$ .

**Esempio 9.2.3** (Continuazione Esempio 9.1.1). Riprendiamo l'Esempio 9.1.1. Già sappiamo che le varianza dei due stimatori di  $\theta$ , sono date da

$$\text{Var}_\theta(\bar{X}_n) = \frac{\theta}{n} \quad \text{e} \quad \text{Var}_\theta(S_n^2) = \frac{\theta}{n} \left(1 + \frac{2n\theta}{n-1}\right) \quad (9.46)$$

sicché la loro *efficienza relativa* è data da

$$eff(S_n^2, \bar{X}_n) = \frac{\mathbb{V}ar_\theta(S_n^2)}{\mathbb{V}ar_\theta(\bar{X}_n)} = 1 + \frac{2n\theta}{n-1} > 1 \quad (9.47)$$

e ciò vuol dire che stimando  $\theta$  tramite  $S_n^2$ , per raggiungere la stessa precisione ottenuta con  $\bar{X}_n$  dovremmo disporre di un campione di ampiezza

$$n^* = \left(1 + \frac{2n\theta}{n-1}\right) \cdot n > n \quad (9.48)$$

Per esempio, fissato  $n = 21$  e  $\theta = 1$ , dalla precedente si ha

$$n^* = \left(1 + \frac{2 \cdot 21 \theta}{n-1}\right) \cdot 21 = 1 + 2.1 \cdot 21 = 65.1 \simeq 66 \quad (9.49)$$

e ciò vuol dire che per raggiungere la stessa precisione di  $\bar{X}_n$  usando  $S_n^2$  dobbiamo disporre di un campione di numerosità più che tripla.

L'efficienza è un concetto che può anche essere collegato alla capacità di sfruttare, intensamente e in maniera ottimale, l'informazione contenuta della realizzazione del campione casuale; più ciò è marcato e più piccola è la numerosità campionaria richiesta per costruire buone procedure inferenziali.

### 9.2.0.2 Estensioni della disuguaglianza di Rao-Cramér

a) *Limite inferiore di Rao-Cramér* per la varianza di stimatori non distorti di funzioni di  $\theta$

**Teorema 9.2.3.** Sotto le medesime assunzioni fatte per il Teorema 9.2.1, sia  $V_n$  uno stimatore non distorto per una funzione di  $\theta$ ,  $\tau(\theta)$ , ossia

$$\mathbb{E}_\theta(V_n) = \tau(\theta), \quad \forall \theta \in \Theta \quad (9.50)$$

Allora,

$$\mathbb{V}ar_\theta(V_n) \geq \frac{\left[\frac{d}{d\theta} \tau(\theta)\right]^2}{I_n(\theta)} \quad (9.51)$$

*Dimostrazione.* La dimostrazione è identica a quella svolta precedentemente a cui si rimanda.  $\square$

b) *Matrice di informazione di Fisher*

**Definizione 9.2.6.** La matrice il cui  $(i, j)$ -mo elemento è dato da

$$h_{ij} = -\mathbb{E}_\theta \left[ \frac{d^2}{d\theta_i d\theta_j} \ell(\theta | \mathbf{x}) \right] \quad (9.52)$$

è detta *matrice di informazione di Fisher* e la indicheremo con  $\mathbf{H}$ .

Sia  $h^{ij}$  l' $(i, j)$ -mo elemento di  $\mathbf{H}^{-1}$ , inversa della matrice di informazione  $\mathbf{H}$ ; sotto le usuali condizioni di regolarità si ha che

$$\text{Var}_\theta(T_i) \geq h^{ii}, \quad i = 1, 2, \dots, k \quad (9.53)$$

dove  $T_i$  è un qualsiasi stimatore non distorto di  $\theta_i$  e  $h^{ii}$  il *limite inferiore* della sua varianza resituo dalla disuguaglianza di Rao-Cramér. Laddove nella precedente valesse l'*uguaglianza*, diremo lo stimatore  $T_i$  *efficiente* per  $\theta_i$ .

L'inversa della matrice di informazione  $\mathbf{H}$ ,  $\mathbf{H}^{-1}$ , altro non è che la **matrice di covarianza asintotica** dello stimatore  $\mathbf{T} = (T_1, T_2, \dots, T_k)$  del vettore di parametri  $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ .

**Esempio 9.2.4.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Normale  $N(\mu, \sigma^2)$  e sia  $\mathbf{T}_n = (\bar{X}_n, S_n^2)$  uno stimatore di  $(\mu, \sigma^2)$ . Cosa possiamo dire in merito all'*efficienza* di questo stimatore?

La funzione di log-verosimiglianza è data da

$$\ell(\mu, \sigma^2 | \mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 + \ln \prod_{i=1}^n \mathbb{1}_{\mathbb{R}}(x_i)$$

e

$$\begin{aligned} h_{11} &= -\mathbb{E}_{\mu, \sigma^2} \left[ \frac{d^2}{d\mu^2} \ell(\mu, \sigma^2 | \mathbf{x}) \right] = -E_{\mu, \sigma^2} \left[ -\frac{n}{\sigma^2} \right] = \frac{n}{\sigma^2} \\ h_{12} &= h_{21} = -\mathbb{E}_{\mu, \sigma^2} \left[ \frac{d^2}{d\mu d\sigma^2} \ell(\mu, \sigma^2 | \mathbf{x}) \right] = \frac{n}{\sigma^4} E_{\mu, \sigma^2}(\bar{X}_n - \mu) = 0 \\ h_{22} &= -\mathbb{E}_{\mu, \sigma^2} \left[ \frac{d^2}{d\sigma^2} \ell(\mu, \sigma^2 | \mathbf{x}) \right] = -E_{\mu, \sigma^2} \left[ -\frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (X_i - \mu)^2 \right] = \frac{n}{2\sigma^4} \end{aligned}$$

per cui

$$\mathbf{H} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \quad (9.54)$$

Ora, l'inversa di  $\mathbf{H}$  data da

$$\mathbf{H}^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \quad (9.55)$$

restituisce la *matrice di covarianza asintotica* dello stimatore  $\mathbf{T}_n = (\bar{X}_n, S_n^2)$  di  $(\mu, \sigma^2)$ .

Da notare che in questo caso  $\text{Cov}(\bar{X}_n, S_n^2) = 0$  perchè in ambito di campionamento da distribuzione Normale, *media campionaria*  $\bar{X}_n$  e *varianza campionaria*  $S_n^2$  sono *indipendenti*, *ergo incorrelate* (vedere Teorema 5.2.1).

Inoltre, dall'analisi degli elementi di  $\mathbf{H}^{-1}$  risulta che la media campionaria  $\bar{X}_n$  è uno stimatore *efficiente* di  $\mu$  poichè

$$\text{Var}_{\mu, \sigma^2}(\bar{X}_n) = h^{11} = \frac{\sigma^2}{n} \quad (9.56)$$

mentre la varianza campionaria  $S_n^2$  è uno stimatore solo *asintoticamente efficiente* per  $\sigma^2$  (a meno che la media della popolazione  $\mu$  sia nota, in quel caso è efficiente), poiché

$$\mathbb{V}ar_{\mu, \sigma^2}(S_n^2) = \frac{2\sigma^4}{n-1} \neq h^{22} = \frac{2\sigma^4}{n} \quad (9.57)$$

e

$$eff(T_n) = \frac{h^{22}}{\mathbb{V}ar(S_n^2)} = \frac{2\sigma^4/n}{2\sigma^4/(n-1)} = \frac{n-1}{n} \neq 1 \quad (9.58)$$

ma

$$\lim_{n \rightarrow \infty} eff(T_n) = \lim_{n \rightarrow \infty} \frac{n-1}{n} = 1. \quad (9.59)$$

Di conseguenza  $(\bar{X}_n, S_n^2)$  non è uno stimatore efficiente di  $(\mu, \sigma^2)$  sebbene lo sia asintoticamente.

In generale, laddove uno stimatore non distorto e consistente di  $\theta$ , scalare o vettore che sia, sia asintoticamente efficiente (come vedremo essere, per esempio, gli stimatori di massima verosimmiglianza) potremmo usare il limite inferiore di Rao-Cramér come *approssimazione della varianza* dello stimatore

$$\mathbb{V}ar_\theta(T_n) \simeq I_n^{-1}(\theta) \quad (9.60)$$

e stimare  $I_n(\theta)$ , informazione di Fisher (teorica) con la quantità

$$i_n(T_n) = -\frac{d^2}{d\theta^2} \ell(T_n | \boldsymbol{x}) = I_n(T_n) \quad (9.61)$$

detta *informazione di Fisher osservata*; fatto questo, potremmo stimare la varianza di  $T_n$  nella seguente maniera

$$\hat{\mathbb{V}ar}_\theta(T_n) \simeq \hat{\mathbb{V}ar}_\theta(T_n) = i_n^{-1}(T_n) \quad (9.62)$$

Possiamo trovare una giustificazione teorica della conclusione raggiunta in (9.60) osservando che

- a) per definizione di consistenza (semplice)  $T_n \xrightarrow{P} \theta$
- b)  $i_1(\theta)$  è una funzione continua di  $\theta$  e allora, per il teorema del *continuous mapping* (Teorema 3.1.6),

$$n \cdot i_1(T_n) \xrightarrow{P} n \cdot I_1(\theta) = I_n(\theta) \quad (9.63)$$

per cui la conclusione in 9.60 è giustificata.

### 9.3 Score function ed efficienza

La definizione che segue introduce la nozione di *score function*; essa gode di alcune utili proprietà e ha una relazione interessante con l'*informazione di Fisher* e l'*efficienza*.

**Definizione 9.3.1** (Score function). La derivata prima rispetto  $\theta$  della funzione di log-verosimiglianza  $\ell(\theta|\mathbf{x})$

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta|\mathbf{x}) \quad (9.64)$$

è detta *score function*.

E' immediato osservare che l'*informazione di Fisher*  $I_n(\theta)$  altro non è che la varianza della *score function* poichè

$$I_n(\theta) = \mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta|\mathbf{x}) \right]^2 = \mathbb{E}_\theta [S(\theta; \mathbf{X})]^2 = \text{Var}_\theta [S(\theta; \mathbf{X})] \quad (9.65)$$

dal momento che  $\mathbb{E}_\theta [S(\theta; \mathbf{X})] = 0$ .

Possiamo subito notare che, in presenza di un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da una famiglia di distribuzioni *regolari* (ovvero che soddisfano le *condizioni di regolarità*), eguagliando a zero la *score function*

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta|\mathbf{x}) = 0 \quad (9.66)$$

si ottiene quella che è nota come *score equation*, la cui soluzione rispetto a  $\theta$  restituisce stimatore di massima verosimiglianza di  $\theta$ .

E' importante ricordare sempre che il *limite inferiore* per la varianza di un qualsiasi stimatore non distorto  $T_n$  di  $\theta$  fornito dalla *disuguaglianza di Rao-Cramér* è pur sempre un **limite inferiore** e pertanto, in generale, niente garantisce che esso sia raggiunto da  $\text{Var}(T_n)$  per  $n$  finito e neppure asintoticamente.

Sempre, per il momento, assumendo  $\theta$  scalare, vi è tuttavia un caso in cui una *forma particolare della score function* garantisce l'esistenza di uno stimatore *efficiente* di  $\theta$ .

**Teorema 9.3.1** (Score function e efficienza). Sotto le usuali *condizioni di regolarità*, esiste uno stimatore non distorto la cui varianza raggiunge il limite inferiore di Rao-Cramér se e solo se

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta|\mathbf{x}) = (T_n - \theta) \cdot I_n(\theta) \quad (9.67)$$

*Dimostrazione.* Nel dimostrare la disuguaglianza di Rao-Cramér, a un certo punto posto  $U = T_n$  e  $V = \frac{d}{d\theta} \ell(\theta|\mathbf{X})$ , avevamo ottenuto la seguente disuguaglianza

$$\text{Cov}_\theta^2(U, V) \leq \text{Var}_\theta(U) \cdot \text{Var}_\theta(V)$$

dove, per quanto già visto in merito di covarianza, vale l'uguaglianza se e solo se sussiste una relazione lineare tra  $U$  e  $V$  ossia

$$\frac{d}{d\theta} \ell(\theta|\mathbf{X}) = c_1 + c_2 T_n$$

con  $c_1$  e  $c_2$  costanti. Ora,

$$\begin{aligned}\mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \right] &= \mathbb{E}_\theta [c_1 + c_2 T_n] \\ &= c_1 + c_2 \mathbb{E}_\theta (T_n) \\ &= c_1 + c_2 \theta\end{aligned}$$

sicché  $c_1 = -c_2 \theta$  essendo  $\mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \right] = 0$ . Ma allora,

$$\frac{d}{d\theta} \ell(\theta | \mathbf{X}) = c_2 (T_n - \theta). \quad (9.68)$$

Ora, moltiplicando ambo i membri di (9.68) per  $\frac{d}{d\theta} \ell(\theta | \mathbf{X})$  si ha

$$\left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \right]^2 = c_2 \frac{d}{d\theta} \ell(\theta | \mathbf{X}) T_n - c_2 \theta \frac{d}{d\theta} \ell(\theta | \mathbf{X})$$

e prendendo i valori attesi su ambo i membri della precedente, si ha

$$\mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \right]^2 = c_2 \mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \cdot T_n \right] - c_2 \theta \mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \right]$$

sicché  $c_2 = I_n(\theta)$  poiché  $\mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \cdot T_n \right] = 1$  e  $\mathbb{E}_\theta \left[ \frac{d}{d\theta} \ell(\theta | \mathbf{X}) \right] = 0$ . Infine, sostituendo  $c_2 = I_n(\theta)$  in (9.68), si ha

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta | \mathbf{X}) = (T_n - \theta) \cdot I_n(\theta).$$

□

Questi risultati assicurano l'*efficienza* o almeno *efficienza asintotica* degli *stimatori di massima verosimiglianza* per i parametri di distribuzioni che appartengono a famiglie *regolari*, in particolare, a *famiglia esponenziale* a  $k$ -parametri. In questo caso sappiamo esistere una statistica *sufficiente minimale*  $W_n(\mathbf{X})$  e, in virtù del teorema di fattorizzazione di Neyman

$$L(\theta | \mathbf{X}) = g(W_n(\mathbf{x}); \theta) \cdot h(\mathbf{x}) \quad (9.69)$$

e

$$\ell(\theta | \mathbf{X}) = \ln [g(W_n(\mathbf{x}); \theta)] \cdot \ln [h(\mathbf{x})] \quad (9.70)$$

Ora, essendo la famiglia da cui proviene il campione *regolare*, lo stimatore di massima verosimiglianza  $\hat{\theta}_n$  di  $\theta$  si otterrà risolvendo l'equazione (o il sistema di  $k$  equazioni, se  $\theta$  è vettore  $k$ -dimensionale)

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta | \mathbf{X}) = \frac{d}{d\theta} \ln [g(W_n(\mathbf{x}); \theta)] = 0 \quad (9.71)$$

sicché

$$\hat{\theta}_n = q(W_n(\mathbf{X})) \quad (9.72)$$

assodato che  $\frac{d}{d\theta} S(\theta; \mathbf{x}) = \frac{d^2}{d\theta^2} \ell(\theta | \mathbf{X}) < 0$ .

Il ragionamento svolto qui sopra a proposito degli stimatori di massima verosimiglianza trova formalizzazione nel seguente teorema; esso afferma che se la varianza di uno stimatore non disorto raggiunge il limite inferiore di Rao-Cramér (per  $n$  finito o asintoticamente) questo stimatore sarà lo stimatore di massima verosimiglianza di  $\theta$ .

**Teorema 9.3.2** (Stimatori di massima verosimiglianza ed efficienza). Sotto le usuali *condizioni di regolarità*, sia  $T_n$  uno stimatore non distorto di  $\theta$  la cui varianza raggiunge il limite inferiore di Rao-Cramèr (ergo, uno stimatore *efficiente* di  $\theta$ ) e sia  $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$  lo stimatore di massima verosimiglianza di  $\theta$  restituito dalla soluzione dell'equazione

$$S(\theta; \mathbf{X}) = \frac{d}{d\theta} \ell(\theta | \mathbf{X}) = 0.$$

Allora,  $T_n = \hat{\theta}_n$ .

*Dimostrazione.* Per quanto visto nel Teorema 9.3.1,  $T_n$  deve soddisfare alla seguente identità

$$S(\theta; \mathbf{X}) = (T_n - \theta) I_n(\theta).$$

Lo stimatore di massima verosimiglianza di  $\theta$  è radice dell'equazione ottenuta eguagliando a zero la precedente funzione di  $\theta$ ; sicchè, essendo  $I_n(\theta) > 0$

$$S(\theta; \mathbf{X}) = (T_n - \theta) I_n(\theta) = 0 \quad (9.73)$$

solo se  $(T_n - \hat{\theta}_n) = 0$  ovvero se  $T_n = \hat{\theta}_n$ .  $\square$

**Esempio 9.3.1** (Continuazione Esempio 9.1.1). Nell'ambito dell'esempio in questione, la funzione di log-verosimiglianza è uguale a

$$\ell(\theta | \mathbf{x}) = -n\theta + \ln(\theta) \sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!) \quad (9.74)$$

e la score function sarà dunque

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta | \mathbf{x}) = -n + \frac{1}{\theta} \sum_{i=1}^n x_i \quad (9.75)$$

sicché

$$I_n(\theta) = -\mathbb{E}_\theta \left[ \frac{d}{d\theta} S(\theta; \mathbf{X}) \right] = \frac{n}{\theta} \quad (9.76)$$

Ora,

$$S(\theta; \mathbf{x}) = -n + \frac{1}{\theta} \sum_{i=1}^n x_i = \left( \frac{1}{n} \sum_{i=1}^n x_i - \theta \right) \frac{n}{\theta} \quad (9.77)$$

sicché, per il Teorema 9.3.1,  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  è stimatore *efficiente* di  $\theta$ . Inoltre, in base al Teorema 9.3.2, lo stimatore efficiente  $\bar{X}_n$  coincide con lo stimatore di massima verosimiglianza di  $\theta$  soluzione della *score equation*

$$S(\theta; \mathbf{x}) = -n + \frac{1}{\theta} \sum_{i=1}^n x_i = 0 \quad (9.78)$$

**Esempio 9.3.2.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Gamma di parametri  $\alpha > 0$  e  $\beta > 0$  con  $\alpha$  noto. Vogliamo trovare uno stimatore efficiente per  $\beta$ .

La funzione di verosimiglianza è

$$\begin{aligned} L(\beta|\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\Gamma(\alpha) \beta^\alpha} x_i^{\alpha_i-1} e^{-\frac{1}{\beta}x_i} \mathbb{1}_{\mathbb{R}^+}(x_i) \\ &= C \cdot \frac{1}{\beta^{n\alpha}} e^{-\frac{1}{\beta} \sum_{i=1}^n x_i} \mathbb{1}_{\mathbb{R}^+}(x_i) \end{aligned} \quad (9.79)$$

e

$$\ell(\beta|\mathbf{x}) = \ln L(\beta|\mathbf{x}) = \ln(C) - n\alpha \ln(\beta) + \frac{1}{\beta} \sum_{i=1}^n x_i \quad (9.80)$$

da cui

$$S(\beta; \mathbf{x}) = \frac{d}{d\beta} \ell(\beta|\mathbf{x}) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i \quad (9.81)$$

e

$$\begin{aligned} I_n(\beta) &= -E_\beta \left[ \frac{d}{d\beta} S(\beta; \mathbf{x}) \right] = -E_\beta \left[ \frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^n X_i \right] \\ &= -\frac{n\alpha}{\beta^2} + \frac{2}{\beta^3} \sum_{i=1}^n E_\beta(X_i) \\ &= -\frac{n\alpha}{\beta^2} + \frac{2}{\beta^3} n \alpha \beta \\ &= \frac{n\alpha}{\beta^2} \end{aligned} \quad (9.82)$$

Ora, possiamo subito osservare che

$$\begin{aligned} S(\beta; \mathbf{x}) &= -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i \\ &= \left[ \frac{\bar{x}_n}{\alpha} - \beta \right] \cdot \frac{n\alpha}{\beta^2} \end{aligned} \quad (9.83)$$

e tenuto conto del fatto che

$$\mathbb{E}_\beta(T_n) = \frac{1}{\alpha} \mathbb{E}_\beta(\bar{X}_n) = \frac{1}{\alpha} \mathbb{E}_\beta(X) = \frac{1}{\alpha} \alpha \beta = \beta, \quad \forall \beta > 0 \quad (9.84)$$

concludere che  $T_n$  è stimatore non distorto di  $\beta$  e, per il Teorema 9.3.1, è stimatore efficiente per  $\beta$ .

## 9.4 Stimatori UMVU

Abbiamo avuto modo di sottolineare che il limite inferiore di Rao-Cramèr è solo un limite inferiore e nulla, salvo i casi di cui si occupa il Teorema 9.3.1, garantisce

che esista uno stimatore non distorto la cui varianza lo raggiunga. Tanto più se vengono meno le *condizioni di regolarità* sulla famiglia di distribuzioni generante il campione e su cui si regge l'impianto che abbiamo finora trattato, situazione in cui il suddetto limite non ha più alcun senso.

**Esempio 9.4.1.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Uniforme su  $(0, \theta)$  sicché già sappiamo che

$$T_n = \frac{n+1}{n} X_{(n)} \quad (9.85)$$

con  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$ , è uno stimatore non distorto di  $\theta$  con

$$\text{Var}_\theta(T_n) = \left(\frac{n+1}{n}\right)^2 \text{Var}_\theta(X_{(n)}) = \frac{\theta^2}{n(n+2)} \quad (9.86)$$

Ma cosa possiamo dire in merito alla varianza di  $T_n$ ? Si può ipotizzare l'esistenza di uno stimatore  $U_n$  non distorto di  $\theta$  per il quale  $\text{Var}_\theta(U_n) < \text{Var}_\theta(T_n)$ ? E anche ciò fosse, la scelta di  $U_n$  è la migliore in assoluto?

E' del tutto evidente che la mancanza di un termine di paragone per la varianza di un qualsiasi stimatore non distorto di  $\theta$  quale quello fornito dalla disugualanza di Rao-Cramér complica enormemente arrivare a decisioni definitive in merito all'azione da intraprendere.

*Come ci dobbiamo comportare in questi casi?*

La *sufficienza*, unita alla *non distorsione*, può giocare un ruolo fondamentale nel trovare stimatori a uniformemente a minimo rischio, vale a dire, stimatori a varianza uniformemente minima nella classe degli stimatori non distorti per  $\theta$ , anche senza raggiungere il limite inferiore stabilito dalla disugualanza di Rao-Cramér.

Passiamo ora a enunciare uno dei più importanti teoremi della statistica matematica, il *teorema di Rao-Blackwell*, che mette per l'appunto in relazione statistiche sufficienti e stimatori non distorti di  $\theta$ .

**Teorema 9.4.1** (Teorema di Rao-Blackwell). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione avente funzione di densità (o di massa)  $f_X(x; \theta)$  e sia  $L(\theta|x)$  la corrispondente funzione di verosimiglianza. Sia inoltre  $T_n$  una statistica sufficiente per  $\theta$  e  $V_n$  un qualsiasi stimatore non distorto di  $\theta$ . Allora

$$V_{n;T_n} = \mathbb{E}_\theta(V_n|T_n) \quad (9.87)$$

è uno stimatore

- a) non distorto di  $\theta$ :  $\mathbb{E}_\theta(V_{n;T_n}) = \theta \quad \forall \theta \in \Theta$ ;
- b)  $\text{Var}_\theta(V_{n;T_n}) \leq \text{Var}_\theta(V_n)$
- c) è funzione di statistica sufficiente:  $V_{n;T_n} = \varphi(T_n)$ .

Possiamo subito osservare che il *condizionamento* attraverso statistica sufficiente: *preserva la non distorsione, migliora lo stimatore in termini di rischio* e restituisce lo stimatore in *funzione di statistica sufficiente*. In altre parole,  $V_{n;T_n}$  è lo stimatore a *minima varianza* nella classe degli stimatori non distorti di  $\theta$ .

*Dimostrazione.* Prima di procedere con la dimostrazione, richiamiamo le seguenti proprietà del valor medio e della varianza (Teorema 1.3.3 e Teorem 1.3.4)

$$\begin{aligned}\mathbb{E}_\theta(X) &= \mathbb{E}_\theta[(\mathbb{E}_\theta(X|Y))] \\ \text{Var}_\theta(X) &= \text{Var}_\theta[(\mathbb{E}_\theta(X|Y))] + \mathbb{E}_\theta[\text{Var}_\theta(X|Y)]\end{aligned}\quad (9.88)$$

Siano ora  $X = V_n$  e  $Y = T_n$ . Per le proprietà appena ricordate, vale che

$$\mathbb{E}_\theta(V_n) = \mathbb{E}_\theta[\mathbb{E}_\theta(V_n|T_n)] = \theta \quad \forall \theta \in \Theta \quad (9.89)$$

pertanto  $V_{n;T_n}$  è stimatore *non distorto* di  $\theta$ . Inoltre

$$\text{Var}_\theta(V_n) = \text{Var}_\theta(V_{n;T_n}) + \mathbb{E}_\theta[\text{Var}_\theta(V_n|T_n)] \geq \text{Var}_\theta(V_{n;T_n}) \quad (9.90)$$

in quanto il valore atteso di una quantità non negativa qual è la varianza condizionata  $\text{Var}_\theta(V_n|T_n)$  è sempre una quantità non negativa. Possiamo infine osservare che

$$V_{n;T_n} = \mathbb{E}_\theta(V_n|T_n) = \int_{S_V} v_n f_{V_n|T_n}(v_n|t_n) dv_n = \varphi(T_n). \quad (9.91)$$

□

Ora, essendo  $V_n$  un *qualsiasi* stimatore *non distorto* e valendo la disegualanza (9.90),  $V_{n;T_n}$  ha *varianza più piccola uniformemente in  $\theta$*  nella classe degli stimatori *non distorti* di  $\theta$ . Pertanto,  $V_{n;T_n}$  è a minima varianza. Ma, attenzione, varianza più piccola *non vuole necessariamente dire uguale al limite inferiore di Rao-Cramér*.

In buona sostanza, il cuore del teorema di Rao-Blackwell è costituito da una *procedura di miglioramento* di uno stimatore non distorto (e *non su di un principio di ottimalità*). Ciò implica che, data una *regola di decisione* (vale a dire, uno stimatore non distorto di  $\theta$ ) e una statistica *sufficiente*  $T_n$ , possiamo invocare il teorema di Rao-Blackwell per costruire una *nuova regola di decisione* data da  $V_{n;T_n} = E_\theta(V_n|T_n)$  che garantisce una *riduzione del rischio* per una qualsiasi funzione di perdita convessa poiché

$$\mathbb{E}_\theta[(V_{n;T_n} - \theta)^2] = \text{Var}_\theta(V_{n;T_n}) \leq \text{Var}_\theta(V_n) \quad (9.92)$$

Rimane comunque la possibilità che questa regola di decisione, pur essendo la migliore possibile, sia comunque *lontana dall'ottimalità* (limite inferiore di Rao-Cramér, laddove esso esista e abbia senso).

**Esempio 9.4.2.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione Esponenziale di parametro  $\beta$  con  $\beta > 0$  avente funzione di densità

$$f_X(x; \beta) = \frac{1}{\beta} e^{-\frac{1}{\beta}x} \mathbb{1}_{\mathbb{R}^+}(x) \quad (9.93)$$

Vogliamo trovare lo stimatore UMVU di  $\beta$ .

Ricordiamo che  $\text{Exp}(\beta) = \mathcal{G}(1, \beta)$ ; di conseguenza,  $V_n = X_1 \sim \mathcal{G}(1, \beta)$  è stimatore non distorto di  $\beta$  e  $T_n = \sum X_i \sim \mathcal{G}(n, \beta)$  statistica sufficiente. Allora

$$V_{n;T_n} = \mathbb{E}_\beta(X_1|T_n) = \int_0^{t_n} x_1 f_{X_1|T_n}(x_1|t_n) dx_1 \quad (9.94)$$

dove

$$\begin{aligned}
 f_{X_1|T_n}(x_1|t_n) &= \frac{f_{X_1, T_n}(x_1, t_n)}{f_{T_n}(t_n)} \\
 &= \frac{f_{X_1}(x_1)f_{T_n-X_1}(t_n - x_1)}{f_{T_n}(t_n)} \\
 &= \frac{\frac{1}{\beta} e^{-\frac{x_1}{\beta}} \frac{1}{\Gamma(n-1)\beta^{n-1}} (t_n - x_1)^{(n-1)-1} e^{-\frac{t_n-x_1}{\beta}}}{\frac{1}{\Gamma(n)\beta^n} t_n^{n-1} e^{-\frac{t_n}{\beta}}} \\
 &= \frac{n-1}{t_n} \left(1 - \frac{x_1}{t_n}\right)^{n-2}
 \end{aligned} \tag{9.95}$$

Al secondo passaggio, abbiamo scritto  $t_n$  come  $t_n = x_1 + \sum_{i=1}^n x_i$ ; sapendo che  $T_n = \sum_{i=1}^n X_i \sim \mathcal{G}(n, \beta)$ , allora  $T_n - X_1 = \sum_{i=2}^n X_i \sim \mathcal{G}(n-1, \beta)$ . Pertanto

$$\mathbb{E}_\beta \left( X_1 \mid \sum_{i=1}^n X_i \right) = \int_0^{t_n} x_1 \left( \frac{n-1}{t_n} \right) \left( 1 - \frac{x_1}{t_n} \right)^{n-2} dx_1 = \frac{t_n}{n} \tag{9.96}$$

Lo stimatore UMVU di  $\beta$  è dato dunque da

$$V_{n;T_n} = \bar{X}_n = \varphi \left( \sum_{i=1}^n X_i \right). \tag{9.97}$$

Notiamo che  $V_{n;T_n} = \bar{X}_n$  è anche stimatore efficiente per  $\beta$ . Infatti

$$\begin{aligned}
 \mathbb{V}ar_{n;T_n}(\bar{X}_n) &= \frac{\mathbb{V}ar_\beta(X)}{n} = \frac{\beta^2}{n} \\
 I_n(\beta) &= -E_\beta \left( \frac{d^2}{d\beta^2} l(\beta; \mathbf{x}) \right) = -E_\beta \left( \frac{n}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^n X_i \right) = -\frac{n}{\beta^2} + \frac{2n}{\beta^2} = \frac{n}{\beta^2}
 \end{aligned} \tag{9.98}$$

Pertanto la varianza di  $\bar{X}_n$  raggiunge il limite inferiore di Rao-Cramèr.

**Esempio 9.4.3.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione di Poisson  $\mathcal{P}(\theta)$ , con  $\theta > 0$  e sia

$$\eta(\theta) = P_\theta(X_1 > 0) = 1 - e^{-\theta} = 1 - P_\theta(X_1 = 0). \tag{9.99}$$

Vogliamo determinare lo stimatore UMVU di  $\eta(\theta) = P_\theta(X_1 > 0)$ .

A questo scopo definiamo la statistica

$$V_n = \mathbb{1}_{\{X_1>0\}}(x_1) = \begin{cases} 1 & \text{se } x_1 > 0 \\ 0 & \text{se } x_1 = 0 \end{cases} \tag{9.100}$$

e ricordando che

$$\mathbb{E}_\theta(\mathbb{1}_A(X)) = P(A) \tag{9.101}$$

si ha subito che  $V_n$  è stimatore non distorto di  $\eta(\theta)$ . Già sappiamo che

$$T_n = \sum_{i=1}^n X_i \sim \mathcal{P}(n\theta) \quad (9.102)$$

è statistica sufficiente minimale per  $\theta$ , la cui distribuzione si può facilmente determinare via *proprietà di riproducibilità*. Allora lo stimatore UMVU può essere ottenuto applicando il teorema di Rao-Blackwell, ovvero

$$\begin{aligned} V_{n;T_n} &= E_\theta [\mathbb{1}_{\{X_1>0\}} | T_n] \\ &= 1 - P_\theta(X_1 = 0 | T_n = t_n) \\ &= 1 - \frac{P_\theta(X_1 = 0, T_n = t_n)}{P(T_n = t_n)} \\ &= 1 - \frac{P_\theta(X_1 = 0) P_\theta(T_n = t_n | X_1 = 0)}{P_\theta(T_n = t_n)} \\ &= 1 - \frac{P_\theta(X_1 = 0) P_\theta(\sum_{i=2}^n X_i = t_n)}{P_\theta(T_n = t_n)} \\ &= 1 - \frac{e^{-\theta} ((n-1)\theta)^{t_n} e^{-(n-1)\theta} / t_n!}{(n\theta)^{t_n} e^{-n\theta} / t_n!} \\ &= 1 - \left( \frac{n-1}{n} \right)^{t_n} \end{aligned} \quad (9.103)$$

poiché

$$\sum_{i=2}^n X_i = T_n - X_1 \sim \mathcal{P}((n-1)\theta) \quad (9.104)$$

sempre per la *proprietà di riproducibilità* di cui gode la distribuzione di Poisson.

In generale, una volta individuato lo stimatore UMVU di  $\theta$  rimane da affrontare il problema della sua *unicità* per poter poi chiudere il cerchio. A questo fine intordurremo un concetto, piuttosto tecnico: la *completezza*.

**Definizione 9.4.1** (Completezza). Sia  $X$  una v.c. continua o discreta la cui funzione di densità/massa appartenga alla famiglia  $\mathcal{F}_\theta = \{f_X(x; \theta), \theta \in \Theta\}$ . Se la condizione  $E_\theta[\varphi(X)] = 0$ , per ogni  $\theta \in \Theta$ , richiede che  $\varphi(X)$  sia nulla eccetto che su un insieme di punti che ha probabilità zero per ogni  $f_X(x; \theta), \theta \in \Theta$  (vale a dire,  $\varphi(X)$  q.o. nulla), allora la famiglia di distribuzioni  $\{f_X(x; \theta), \theta \in \Theta\}$  è detta *famiglia di funzioni di densità/massa completa*.

Come si evince dalla definizione, la nozione di *completezza* riguarda la *famiglia delle distribuzioni*; spesso per convenienza, si usa dire che la *statistica*  $T_n$  è *completa* per  $\theta$ , sottointendendo che essa è *completa* per la famiglia di distribuzioni indicizzate dal parametro  $\theta$ . Potremmo allora formulare la seguente definizione di statistca completa.

**Definizione 9.4.2** (Statistica completa). Una statistica  $T_n$ , funzione misurabile di  $(X_1, X_2, \dots, X_n)$ , è detta *completa* per la famiglia di distribuzioni indicizzata dal parametro  $\theta$  e dalla quale proviene il campione casuale  $(X_1, X_2, \dots, X_n)$  se, per ogni funzione misurabile di  $T_n$ ,  $\varphi_{T_n}$ , vale la seguente implicazione

$$\mathbb{E}_\theta [\varphi(T_n)] = 0, \forall \theta \in \Theta \Rightarrow P_\theta [\varphi(T_n) = 0] = 1, \text{ quasi ovunque.} \quad (9.105)$$

**Esempio 9.4.4.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione  $b(1, \theta)$ . Abbiamo già dimostrato che  $T_n = \sum X_i$  è una statistica sufficiente minimale per  $\theta$  e che  $T_n \sim b(n, \theta)$ . Definiamo dunque  $\varphi(T_n)$  una qualunque funzione misurabile di  $T_n$  e calcoliamone il valor medio:

$$\begin{aligned} \mathbb{E}_\theta [\varphi(T_n)] &= \sum_{t_n=0}^n \varphi(t_n) \binom{n}{t_n} \theta^{t_n} (1-\theta)^{n-t_n} \\ &= (1-\theta)^n \sum_{t_n=0}^n \varphi(t_n) \binom{n}{t_n} \left(\frac{\theta}{1-\theta}\right)^{t_n} \end{aligned} \quad (9.106)$$

sicché

$$\mathbb{E}_\theta [\varphi(T_n)] = \mathbb{E}_\theta \left[ \varphi \left( \sum_{i=1}^n X_i \right) \right] = 0 \Rightarrow \varphi \left( \sum_{i=1}^n X_i \right) = 0 \quad (9.107)$$

da cui segue che  $T_n(\mathbf{X}) = \sum_{i=1}^n X_i$  è statistica (sufficiente minimale) *completa* per  $\theta$ .

**Esempio 9.4.5.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione  $N(\mu, \sigma^2)$  con  $\sigma^2$  noto. Consideriamo la statistica

$$\bar{X}_n = \frac{1}{n} \sum X_i \sim N(\mu, \sigma^2/n) \quad (9.108)$$

e calcoliamo il valor medio di una generica funzione  $\varphi(T_n)$ :

$$\begin{aligned} E_\mu(\varphi(\bar{X}_n)) &= \int_{-\infty}^{+\infty} \varphi(\bar{X}_n) \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{n}{2\sigma^2} (\bar{X}_n - \mu)^2 \right\} d\bar{X}_n \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{\frac{n\mu^2}{2\sigma^2}} \int_{-\infty}^{+\infty} \varphi(\bar{X}_n) \exp \left\{ -\frac{n\bar{X}_n^2}{2\sigma^2} + \frac{n\mu\bar{X}_n}{\sigma^2} \right\} d\bar{X}_n \end{aligned} \quad (9.109)$$

È evidente che, se  $E_\mu(\varphi(\bar{X}_n)) = 0$ , allora vale necessariamente che  $\varphi(\bar{X}_n) = 0 \ \forall \mu \in \mathbb{R}$ . Pertanto  $\bar{X}_n$  è statistica *completa* per  $\mu$ .

Una domanda a cui dovremmo rispondere è la seguente: la *completezza* è una proprietà *diffusa* ovvero, caratterizza molte famiglie di distribuzioni?

Il teorema che segue fornisce una risposta a questa domanda.

**Teorema 9.4.2** (Completezza e famiglie esponenziali). Se il campione casuale  $(X_1, X_2, \dots, X_n)$  proviene da una distribuzione appartenente a una famiglia esponenziale a  $k$ -parametri, allora la statistica che già sappiamo essere *sufficiente minimale*

$$\left( \sum_{i=1}^n B_1(X_i), \sum_{i=1}^n B_2(X_i), \dots, \sum_{i=1}^n B_k(X_i) \right) \quad (9.110)$$

è anche *completa* per  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ .

La *completezza* svolge un ruolo fondamentale nello stabilire

- a) l'*indipendenza* di  $V_{n;T_n}$ , stimatore UMVU di  $\theta$  da  $V_n$ , punto di partenza del processo di miglioramento sotteso dal teorema di Rao-Blackwell
- b) di conseguenza, l'*unicità* dello stimatore UMVU  $V_{n;T_n}$  di  $\theta$ .

**Teorema 9.4.3** (Unicità dello stimatore UMVU). Siano  $V_n^{(1)}$  e  $V_n^{(2)}$  due distinti estimatori non distorti di  $\theta$  e sia  $T_n$  una statistica sufficiente e completa per  $\theta$ . Allora

$$V_{n;T_n}^{(1)} = \mathbb{E}_\theta(V_n^{(1)}|T_n) = \mathbb{E}_\theta(V_n^{(2)}|T_n) = V_{n;T_n}^{(2)} \quad (9.111)$$

*Dimostrazione.* Posto  $\varphi(T_n) = V_{n;T_n}^{(1)} - V_{n;T_n}^{(2)}$ , si ha che

$$\mathbb{E}_\theta(\varphi(T_n)) = \mathbb{E}_\theta\left(V_{n;T_n}^{(1)} - V_{n;T_n}^{(2)}\right) = \mathbb{E}\left(V_{n;T_n}^{(1)}\right) - \mathbb{E}\left(V_{n;T_n}^{(2)}\right) = 0 \quad (9.112)$$

e, poiché  $T_n$  è statistica completa, ciò implica ha  $\varphi(T_n) = 0$  da cui  $V_{n;T_n}^{(1)} = V_{n;T_n}^{(2)}$ . Pertanto, lo stimatore restituito dal teorema di Rao-Blackwell è *unico*.  $\square$

Nel calcolare lo stimatore di Rao-Blackwell *per via diretta*, vale a dire svolgendo calcoli come abbiamo fatto negli esempi precedenti coinvolgendo valori attesi condizionati, spesso si possono incontrare non poche difficoltà. Ci viene allora in aiuto il seguente corollario.

**Corollario 9.4.1** (Corollario del Teorema di Rao-Blackwell). La ricerca di estimatori UMVU di  $\theta$  si può *restringere* alla classe delle *funzioni di statistica sufficiente (minimale) e completa* per  $\theta$  che siano estimatori *non distorti* del parametro  $\theta$ .

**Esempio 9.4.6.** Riprendiamo l'Esempio 9.4.2. Sappiamo che  $\beta = \mathbb{E}_\beta(X)$  e che  $\bar{X}_n$  è stimatore *plug-in* di  $\beta$  (in verità, è anche stimatore di massima verosimiglianza di  $\beta$  come si può facilmente verificare). Inoltre,

$$\begin{aligned} L(\beta|\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\beta} e^{-\frac{1}{\beta}x_i} \mathbb{1}_{\mathbb{R}^+}(x_i) \\ &= \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \left(\frac{1}{\beta}\right)^n e^{-\frac{1}{\beta} \sum_{i=1}^n x_i} \end{aligned} \quad (9.113)$$

sicché la distribuzione Esponenziale appartiene a famiglia esponenziale a  $k = 1$ -parametri e dunque

$$T_n = T_n(\mathbf{X}) = \sum_{i=1}^n X_i \quad (9.114)$$

è statistica sufficiente minimale per  $\beta$ . Ora,  $\bar{X}_n$  è stimatore non distorto di  $\beta$  e funzione di statistica sufficiente (minimale) per cui, per il Corollario 9.4.1,

$$\bar{X}_n = \varphi \left( \sum_{i=1}^n X_i \right) = V_{n; T_n} \quad (9.115)$$

è lo stimatore UMVU cercato per  $\beta$ .

**Esempio 9.4.7.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $U(0; \theta)$ . Sappiamo che in questo caso  $X_{(n)} = \max(X_1, X_2, \dots, X_n)$  è statistica sufficiente minimale completa per  $\theta$  e che  $E_\theta(X_{(n)}) = n\theta/(n+1)$ .

Abbiamo già verificato che la funzione  $\varphi(X_{(n)}) = (n+1)X_{(n)}/n$  è uno stimatore non distorto per  $\theta$  ed è, inoltre, funzione di statistica sufficiente minimale completa. Pertanto, per il Corollario 9.4.1,  $\varphi(X_{(n)}) = (n+1)X_{(n)}/n$  è uno stimatore UMVU di  $\theta$  ed è unico; semplicemente, è l'unico stimatore UMVU di  $\theta$ .

**Esempio 9.4.8.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $N(\mu, \sigma^2)$ , con  $\mu, \sigma^2$  non noti. Abbiamo già mostrato che

$$S_n^2 = \frac{n}{n-1} \hat{\sigma}_n^2 \quad \text{con } \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \quad (9.116)$$

è uno stimatore non distorto ed è funzione di

$$\left( \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right) \quad (9.117)$$

statistica congiuntamente sufficiente minimale per  $(\mu, \sigma^2)$ . Pertanto, sempre per il Corollario 9.4.1,  $S_n^2$  è lo stimatore UMVU di  $\sigma^2$  nonostante, come risulta da (9.58), non sia stimatore efficiente. In ogni caso, in ambito di stimatori non distorti della varianza della popolazione  $\sigma^2$ , nulla di meglio si può fare di quanto non faccia  $S_n^2$ .

**Esempio 9.4.9.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione di Bernoulli  $b(1, \theta)$ ,  $\theta \in [0, 1]$ .

Vogliamo trovare uno stimatore UMVU per la varianza della popolazione

$$\mathbb{V}ar_\theta(X) = \eta(\theta) = \theta(1-\theta) \quad (9.118)$$

La funzione di verosimiglianza è data da

$$\begin{aligned} L(\theta | \mathbf{x}) &= \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i) \\ &= \theta^{t_n} (1-\theta)^{n-t_n} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(x_i) \end{aligned} \quad (9.119)$$

e per il teorema di fattorizzazione e ricordando che la distribuzione in questione appartiene a famiglia esponenziale a  $k = 1$  parametri per cui vale il Teorema 9.4.2,  $T_n(\mathbf{X}) = \sum_{i=1}^n X_i$  è statistica sufficiente (minimale) e completa per  $\theta$ . E' anche immediato osservare che lo stimatore di massima verosimiglianza di  $\theta$  è dato da  $\hat{\theta}_n = \frac{1}{n} T_n$ .

Tra le molte interessanti e desiderabili proprietà di cui godono gli stimatori di massima verosimiglianza ve ne è una nota come *proprietà di invarianza* che presenteremo in maniera formale nel prossimo capitolo; per ora basti osservare che, se  $\eta(\theta)$  è una funzione di  $\theta$ , lo stimatore di massima verosimiglianza di  $\eta(\theta)$  è semplicemente dato da  $\eta(\hat{\theta}_n)$ .

Allora, nell'ambito dell'esempio che stiamo svolgendo,

$$\eta(\hat{\theta}_n) = \hat{\theta}_n(1 - \hat{\theta}_n) = \frac{1}{n^2} T_n(n - T_n) \quad (9.120)$$

è lo stimatore di massima verosimiglianza della varianza (9.118) della popolazione bernoulliana.

Essendo intenzionati ad applicare il Corollario 9.4.1 del teorema di Rao-Blackwell, dobbiamo verificare se lo stimatore  $\eta(\hat{\theta}_n)$  sia o meno stimatore non distorto di  $\text{Var}_\theta(X) = \eta(\theta) = \theta(1 - \theta)$ .

Sicché, ricordando che  $T_n \sim b(n, \theta)$ , dobbiamo calcolare

$$\begin{aligned} \mathbb{E}_\theta(\eta(\hat{\theta}_n)) &= \mathbb{E}_\theta \left[ \frac{1}{n^2} T_n(n - T_n) \right] \\ &= \frac{1}{n^2} \mathbb{E}_\theta [T_n(n - T_n)] \\ &= \frac{1}{n^2} [\mathbb{E}_\theta(nT_n) - \mathbb{E}_\theta(T_n^2)] \\ &= \frac{1}{n^2} [n\mathbb{E}_\theta(T_n) - (\text{Var}_\theta(T_n) + \mathbb{E}_\theta^2(T_n))] \\ &= \frac{1}{n^2} [n \cdot n\theta - (n\theta(1 - \theta) + (n\theta)^2)] \\ &= \frac{n-1}{n} \theta(1 - \theta) \neq \theta(1 - \theta) \end{aligned} \quad (9.121)$$

quindi  $\eta(\hat{\theta}_n)$  non è stimatore non distorto di  $\text{Var}_\theta(X) = \theta(1 - \theta)$  sebbene lo sia asintoticamente poiché  $\lim_{n \rightarrow +\infty} \frac{n-1}{n} = 1$ . E' però immediato verificare che

$$\mathbb{E}_\theta(W_n) = \mathbb{E}_\theta \left[ \frac{n-1}{n} \eta(\hat{\theta}_n) \right] = \theta(1 - \theta), \quad \forall \theta \in [0, 1] \quad (9.122)$$

ossia  $W_n$  è stimatore non distorto della varianza della popolazione  $\text{Var}_\theta(X) = \theta(1 - \theta)$ ; inoltre,  $W_n$  è funzione di  $T_n$ , statistica sufficiente (minimale) e completa per  $\theta$  e quindi, per il corollario 9.4.1 del teorema di Rao-Blackwell,  $W_n$  è lo stimatore UMVU di  $\text{Var}_\theta(X)$ .

**Esempio 9.4.10.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale proveniente da una distribuzione di Bernoulli  $b(1, \theta)$  con  $\theta \in (0, 1)$  e vogliamo trovare uno stimatore UMVU di  $\tau(\theta) = \theta^m$  dove  $m \leq n$  è un intero positivo.

Abbiamo già visto che la famiglia delle distribuzioni di Bernoulli appartiene a famiglia esponenziale a  $k = 1$ -parametri e che, pertanto,

$$T_n(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i \quad (9.123)$$

è statistica *sufficiente* (minimale) e *completa* per  $\theta$ .

I virtù del corollario 9.4.1 del teorema di Rao-Blackwell, lo stimatore UMVU di  $\tau(\theta) = \theta^m$  dovrà essere

a) funzione della statistica sufficiente e completa  $T_n$ ,  $\varphi_m(T_n)$

b) non distorto per  $\theta^m$  vale a dire  $\mathbb{E}_\theta[\varphi_m(T_n)] = \theta^m$ ,  $\forall \theta \in [0, 1]$ .

Cominciamo col trovare la funzione  $\varphi_m(T_n)$ . Ricordiamo che  $T_n \sim b(n, \theta)$  e perciò

$$\mathbb{E}_\theta[\varphi_m(T_n)] = \sum_{t_n=0}^n \binom{n}{t_n} \varphi_m(t_n) \theta^{t_n} (1 - \theta)^{n-t_n} \quad (9.124)$$

Ponendo  $\mathbb{E}_\theta[\varphi_m(T_n)] = \theta^m$  si ha

$$\sum_{t_n=0}^n \binom{n}{t_n} \varphi_m(t_n) \theta^{t_n} (1 - \theta)^{n-t_n} = \theta^m \quad (9.125)$$

e dividendo ambo i membri per  $\theta^m$  si ha

$$\sum_{t_n=0}^n \binom{n}{t_n} \varphi_m(t_n) \theta^{t_n-m} (1 - \theta)^{n-m-(t_n-m)} = 1 \quad (9.126)$$

per ogni  $\theta$ . Se  $m > t_n$ ,  $\theta^{t_n-m} \rightarrow \infty$  quando  $\theta \rightarrow 0$ ; perciò dobbiamo avere  $\varphi_m(t_n) = 0$  per  $t_n = 0, 1, 2, \dots, m-1$  e allora possiamo riscrivere la (9.126) come

$$\sum_{t_n=m}^n \binom{n}{t_n} \varphi_m(t_n) \theta^{t_n-m} (1 - \theta)^{n-m-(t_n-m)} = 1 \quad (9.127)$$

per ogni  $\theta$ . Per altro verso, in virtù di una proprietà della distribuzione binomiale

$$\sum_{t_n=m}^n \binom{n-m}{t_n-m} \theta^{t_n-m} (1 - \theta)^{(n-m)-(t_n-m)} = 1 \quad (9.128)$$

per ogni  $\theta$ . Di conseguenza,

$$\binom{n}{t_n} \varphi_m(t_n) = \binom{n-m}{t_n-m} \quad (9.129)$$

da cui

$$\varphi_m(t_n) = \frac{\binom{n-m}{t_n-m}}{\binom{n}{t_n}} \quad (9.130)$$

per  $t_n = m, m+1, \dots, n$ . Infine lo stimatore UMVU di  $\tau(\theta) = \theta^m$  risulta essere

$$\varphi_m(T_n) = \begin{cases} \frac{\binom{n-m}{T_n-m}}{\binom{n}{T_n}}, & T_n = m, (m+1), \dots, n \\ 0, & T_n = 0, 1, \dots, (m-1). \end{cases} \quad (9.131)$$

Un *commento conclusivo*: a differenza di quanto accade per gli stimatori di massima verosimiglianza, come avremo tra poco modo di vedere, per la classe degli stimatori UMVU non è disponibile una teoria asintotica altrettanto attraente, elegante e potente. Questo fatto limita o meglio, complica, la loro applicazione a fini inferenziali dove sappiamo essere cruciale disporre della distribuzione campionaria (esatta o approssimata che sia) delle statistiche coinvolte nel processo inferenziale stesso.

# 10 Stimatori di massima verosimiglianza

Trattando di verosimiglianza e sufficienza abbiamo introdotto, in modo del tutto consequenziale, il metodo di stima di massima verosimiglianza e il conseguente stimatore 8.1.2, introdotto in letteratura da R.A. Fisher a metà degli anni venti del secolo scorso.

Intuitivamente, lo stimatore di massima verosimiglianza di un parametro  $\theta$ , scalare o vettore che sia, è una scelta del tutto ragionevole alla luce di quanto detto in ambito di riduzione della dimensionalità dei dati. Nell'individuare lo stimatore di massima verosimiglianza dobbiamo comunque fare i conti con due aspetti problematici dell'analisi matematica:

- a) il primo è relativo al fatto che dobbiamo trovare il massimo *globale o assoluto* di una funzione (quella di verosimiglianza o di log-verosimiglianza) e verificare che proprio di un siffatto massimo si tratta: non sempre il problema si può risolvere usando il calcolo differenziale. Abbiamo già anticipando questo aspetto parlando di stima di massima verosimiglianza del parametro  $\theta$  della distribuzione Uniforme su  $(0, \theta)$
- b) il secondo aspetto problematico è relativo al fatto che spesso le equazioni di stima di massima verosimiglianza *non sono lineari nei parametri* per cui sovente non vi è soluzione analitica e si devono invocare *metodi numerici* di soluzione quali, per esempio, Newton-Raphson o altri metodi a esso collegati

Nonostante tutto questo, questo metodo di stima è ben presto divenuto la tecnica più popolare e utilizzata per individuare stimatori in miriadi di applicazioni, massimamente in virtù delle sue molte e mirabili proprietà alla cui analisi ora dedicheremo un po' di tempo e di dettaglio.

## 10.1 Proprietà degli stimatori di massima verosimiglianza

Supponiamo che una distribuzione  $F(x; \theta)$  sia indicizzata da un parametro (scalare o vettore che sia)  $\theta \in \Theta$  e di disporre di una realizzazione  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  di un campione casuale  $(X_1, X_2, \dots, X_n)$  di ampiezza  $n$  da essa proveniente; lo stimatore di massima verosimiglianza di  $\theta$  dato da

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L(\theta | \mathbf{x})$$

gode di una serie di interessanti proprietà che saranno illustrate nei seguenti teoremi.

**Teorema 10.1.1** (Relazione con statistica sufficiente minimale). Lo stimatore di massima verosimiglianza  $\hat{\theta}_n$  di un parametro  $\theta$  è funzione di statistica sufficiente minimale per  $\theta$ .

*Dimostrazione.* Qualora esista una statistica sufficiente minimale  $T_n$  per  $\theta$ , per il teorema di fattorizzazione di Neyman,

$$L(\theta|\mathbf{x}) = g(t_n(\mathbf{x}); \theta) \cdot h(\mathbf{x})$$

Dal momento che lo stimatore di massima verosimiglianza  $\hat{\theta}_n$ , per sua stessa definizione, massimizza  $L(\theta|\mathbf{x})$  quale funzione di  $\theta$  allora, quando esiste una statistica sufficiente (minimale), la procedura di massimizzazione di  $L(\theta|\mathbf{x})$  può essere "ristretta" a  $g(t_n(\mathbf{x}); \theta)$  dal momento che

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} L(\theta|\mathbf{x}) = \arg \max_{\theta \in \Theta} g(t_n(\mathbf{x}); \theta).$$

Di conseguenza,  $\hat{\theta}_n$  dovrà dipendere sul campione solo tramite il valore  $t_n$  della statistica sufficiente minimale  $T_n$  da cui segue la tesi del teorema.  $\square$

Supponiamo che una distribuzione  $F(x; \theta)$  sia indicizzata da un parametro  $\theta$  e di essere interessati a trovare una stima di una qualche sua funzione reale  $\tau(\theta)$ . Quello che il seguente teorema afferma è che se  $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$  è lo stimatore di massima verosimiglianza di  $\theta$  allora  $\tau(\hat{\theta}_n)$  è lo stimatore di massima verosimiglianza di  $\tau(\theta)$ . Giusto per fare un esempio, se  $\mu$  è la media di una distribuzione Normale, lo stimatore di massima verosimiglianza di  $\cos^{-1}(\mu)$  è  $\cos^{-1}(\bar{X}_n)$ . E questo non è poca cosa.

**Teorema 10.1.2** (Proprietà di invarianza). Sia  $\tau(\theta) : \Theta \rightarrow \mathcal{T}$  una funzione reale continua del parametro  $\theta$  che indica la famiglia di distribuzioni  $\mathcal{F}_\theta = \{F(x; \theta), \theta \in \Theta\}$  da cui proviene il campione casuale  $(X_1, \dots, X_n)$  e sia  $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$  il relativo stimatore di massima verosimiglianza. Allora lo stimatore di massima verosimiglianza  $\hat{\tau}(\theta)$  di  $\tau(\theta)$  è dato da

$$\hat{\tau}(\theta) = \tau(\hat{\theta}_n).$$

*Dimostrazione.* Spezziamo la dimostrazione del teorema in due passi assumendo prima che  $\tau(\theta)$  sia una funzione reale monotona di  $\theta$  e poi una funzione reale qualsiasi.

- a) Sia  $\tau(\theta)$  funzione monotona di  $\theta$  (vale a dire, per ogni valore di  $\theta$  si ha un solo valore di  $\tau(\theta)$  e viceversa). Se poniamo  $\eta = \tau(\theta)$  allora la funzione inversa  $\theta = \tau^{-1}(\eta)$  è ben definita e la funzione di verosimiglianza di  $\eta = \tau(\theta)$ , scritta dunque come funzione di  $\eta$ , è data da

$$L^*(\eta|\mathbf{x}) = \prod_{i=1}^n f_X(x_i; \tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|\mathbf{x})$$

e

$$\max_{\eta} L^*(\eta|\mathbf{x}) = \max_{\eta} L(\tau^{-1}(\eta)|\mathbf{x}) = \max_{\theta} L(\theta|\mathbf{x})$$

ovvero non fa differenza alcuna massimizzare la funzione di verosimiglianza come funzione di  $\theta$  o come funzione di  $\eta = \tau(\theta)$ . Perciò il massimo di  $L^*(\eta|\mathbf{x})$  si ottiene a  $\eta = \tau(\theta) = \tau(\hat{\theta}_n)$  dimostrando che lo stimatore di massima verosimiglianza  $\hat{\tau}(\theta)$  di  $\tau(\theta)$  è proprio  $\tau(\hat{\theta}_n)$  con  $\hat{\theta}_n$  stimatore di massima verosimiglianza di  $\theta$ .

- b) Se cade la monotonia della funzione  $\tau(\theta)$  allora per un dato valore di  $\eta$  ci può essere più di un valore di  $\theta$  che soddisfa  $\tau(\theta) = \eta$ . In questi casi la corrispondenza, cui si faceva prima riferimento, tra la massimizzazione della funzione di verosimiglianza rispetto a  $\eta$  e quella rispetto a  $\theta$  viene meno.

Per procedere nella dimostrazione del teorema in questa seconda ipotesi, introduciamo il concetto di *funzione di verosimiglianza indotta*  $L^*$  che definiamo nella seguente maniera

$$L^*(\eta|\mathbf{x}) = \max_{\{\theta: \tau(\theta)=\eta\}} L(\theta|\mathbf{x}). \quad (10.1)$$

Il valore  $\hat{\eta}_n$  che massimizza  $L^*(\eta|\mathbf{x})$  sarà detto stimatore di massima verosimiglianza di  $\eta = \tau(\theta)$  e si può vedere da (10.1) che i massimi di  $L^*$  e  $L$  coincidono.

Abbiamo ora tutti gli elementi per la dimostrazione del teorema nella sua forma più generale. Dobbiamo dimostrare che

$$L^*(\hat{\eta}_n|\mathbf{x}) = L^*(\tau(\hat{\theta}_n)|\mathbf{x}).$$

Ora, come stabilito sopra, i massimi di  $L^*$  e di  $L$  coincidono cosicchè abbiamo

$$\begin{aligned} L^*(\hat{\eta}_n|\mathbf{x}) &= \max_{\eta} L^*(\eta|\mathbf{x}) \\ &= \max_{\eta} \max_{\{\theta: \tau(\theta)=\eta\}} L(\theta|\mathbf{x}) \quad (\text{definizione di } L^*(\eta|\mathbf{x})) \\ &= \max_{\theta} L(\theta|\mathbf{x}) \\ &= L(\hat{\theta}_n|\mathbf{x}), \quad (\text{definizione di } \hat{\theta}_n) \end{aligned}$$

dove la seconda uguaglianza segue dal fatto che la massimizzazione iterata è uguale alla massimizzazione non condizionata rispetto a  $\theta$  che si ottiene in  $\theta = \hat{\theta}_n$ . Inoltre,

$$\begin{aligned} L(\hat{\theta}_n|\mathbf{x}) &= \max_{\{\theta: \tau(\theta)=\tau(\hat{\theta}_n)\}} L(\theta|\mathbf{x}) \quad (\hat{\theta}_n \text{ è stimatore di massima verosimiglianza di } \theta) \\ &= L^*(\tau(\hat{\theta}_n)|\mathbf{x}), \quad (\text{definizione di } L^*) \end{aligned}$$

Perciò, la sequenza delle uguaglianze dimostra che

$$L^*(\hat{\eta}_n|\mathbf{x}) = L^*(\tau(\hat{\theta}_n)|\mathbf{x})$$

e che, in conclusione,  $\tau(\hat{\theta}_n)$  è lo stimatore di massima verosimiglianza di  $\tau(\theta)$ .

*Dimostrazione alternativa*

Cominciamo con il considerare  $\tau(\theta)$  una funzione monotona di  $\theta$ . La funzione di verosimiglianza rispetto alla nuova parametrizzazione  $\eta = \tau(\theta)$  è  $L(\tau(\theta); \mathbf{x})$ , ma dal momento che  $\tau$  è monotona

$$\max L(\tau(\theta)|\mathbf{x}) = \max_{\eta=\tau(\theta)} L(\eta|\mathbf{x}) = \max_{\eta} L(\tau^{-1}(\eta)|\mathbf{x}).$$

Ma allora il massimo si ottiene quando  $\tau^{-1}(\eta) = \hat{\theta}_n$  ovvero  $\hat{\eta} = \tau(\hat{\theta}_n)$ .

Supponiamo ora che  $\tau$  non sia monotona. Per ogni  $\eta$  definiamo il seguente insieme (=pre-immagine)

$$\tau^{-1}(\eta) = \{\theta : \tau(\theta) = \eta\}.$$

Il massimo si ha in  $\hat{\theta}_n$  e il dominio di  $\tau$  è quell'insieme  $\mathcal{T}$  che contiene  $\hat{\theta}_n$ . Perciò,  $\hat{\theta}_n$  sarà contenuto in una di queste pre-immagini e, più precisamente, esso può essere contenuto in una sola di queste pre-immagini. Perciò per massimizzare  $L(\eta)$  scegliamo  $\hat{\eta}_n$  tale che  $\tau^{-1}(\hat{\eta}_n)$  sia quell'unica pre-immagine che contiene  $\hat{\theta}_n$ . Ma allora,  $\hat{\eta}_n = \tau(\hat{\theta}_n)$ .  $\square$

**Esempio 10.1.1.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale di ampiezza  $n$  proveniente da una distribuzione Esponenziale di parametro  $\beta > 0$ . Tutto quello che sappiamo in merito al campione è che  $k$  delle  $n$  osservazioni che lo compongono ( $0 < k < n$ ) sono *minori o uguali* a  $M > 0$ , quantità positiva.

Vogliamo trovare lo stimatore di massima verosimiglianza di  $\beta$

Ora, indicato con  $X_i$  l' $i$ -mo elemento del campione casuale, definiamo

$$Y_i = \begin{cases} 1 & \text{se } X_i \leq M \\ 0 & \text{se } X_i > M \end{cases}$$

e perciò  $Y_i \sim b(1, p)$  con  $p \in (0, 1)$  con

$$p = P_\beta(X_i \leq M) = F_X(M; \beta) = 1 - e^{-\frac{M}{\beta}} \mathbb{1}_{\mathbb{R}^+}(x_i) \quad (10.2)$$

sicché  $e^{-(M/\beta)} = 1 - p$  da cui segue

$$\beta = -\frac{M}{\ln(1-p)} = g(p) \quad (10.3)$$

con  $g(p)$  funzione di  $p$ . In fine, lo stimatore di massima verosimiglianza di  $p$  è dato da

$$\hat{p}_n = \bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{\#\{X_i \leq M\}}{n} = \frac{k}{n} \quad (10.4)$$

e dunque, invocando la proprietà di invarianza degli stimatori di massima verosimiglianza, lo stimatore di massima verosimiglianza di  $\beta$  sarà dato da

$$\hat{\beta}_n = g(\hat{p}_n) = -\frac{M}{\ln(1-\hat{p}_n)} = -\frac{M}{\ln(1-\frac{k}{n})}. \quad (10.5)$$

**Esempio 10.1.2.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione di Bernoulli,  $b(1, \theta)$  con  $\theta \in [0, 1]$ , probabilità di successo nella singola prova.

Vogliamo trovare lo stimatore di verosimiglianza della varianza della popolazione

$$\mathbb{V}ar_\theta(X) = \theta(1 - \theta) = g(\theta) \quad (10.6)$$

Tenuto conto del fatto che  $g(\theta)$  è una funzione continua di  $\theta$  e che lo stimatore di massima verosimiglianza di  $\theta$  è dato dalla frequenza relativa campionaria  $\hat{\theta}_n = \bar{X}_n$ , per la proprietà di invarianza degli stimatori di massima verosimiglianza si ha

$$\hat{\mathbb{V}ar}_\theta(X) = \bar{X}_n(1 - \bar{X}_n) \quad (10.7)$$

La migliore giustificazione per l'uso del metodo di stima di massima verosimiglianza sta nel **comportamento asintotico** degli stimatori da esso prodotti. In particolare, laddove la dimensione del campione sia abbastanza grande (per convenzione,  $n > 30$ ), lo stimatore di massima verosimiglianza restituisce - con probabilità elevata - un valore (=stima) molto vicino al vero valore del parametro che deve essere stimato (cioè lo stimatore di massima verosimiglianza è *consistente*). Nelle stesse condizioni, lo stimatore di massima verosimiglianza è *asintoticamente non distorto* e *asintoticamente efficiente* vale a dire, la sua varianza è asintoticamente uguale al limite inferiore fornito dalla disuguaglianza di Rao-Cramér per un qualsiasi stimatore non distorto del parametro; inoltre, la sua distribuzione è *approssimativamente normale*. Tutto ciò fa dello stimatore di massima verosimiglianza uno **stimatore BAN** (= Best Asymptotically Normal, dove l'aggettivo *best* va inteso a varianza più piccola).

**Teorema 10.1.3** (Proprietà asintotiche degli stimatori di massima verosimiglianza). Sotto condizioni di regolarità assai blande, usualmente soddisfatte da molte famiglie di distribuzioni, lo stimatore di massima verosimiglianza  $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$  di  $\theta$  è

1) *consistente*

$$P(|\hat{\theta}_n - \theta_0| > \epsilon) \rightarrow 0 \text{ quando } n \rightarrow \infty$$

dove  $\theta_0$  è il vero valore del parametro  $\theta$ ;

2) *asintoticamente non distorto, asintoticamente efficiente e asintoticamente normalmente distribuito*, ovvero

$$\hat{\theta}_n \underset{a}{\sim} N\left(\theta_0, \frac{1}{I_n(\theta_0)}\right)$$

dove  $I_n(\theta_0)$  è l'informazione di Fisher valutata in  $\theta = \theta_0$ .

*Dimostrazione.* Daremo la dimostrazione per il caso in cui  $\theta \in \Theta \subseteq \mathbb{R}$ ; questa poi può essere estesa caso multiparametrico.

1) Consideriamo la variabile casuale

$$Z(\theta) = \ln f(X; \theta)$$

che dipende su  $\theta$ ; la sua media rispetto alla vera distribuzione di  $X$  (che assumiamo essere continua) corrisponde a

$$\mu(\theta) = E_\theta(Z(\theta)) = \int_{\mathcal{S}_X} \ln f(x; \theta) f(X; \theta_0) dx$$

dove  $\mathcal{S}_X$  rappresenta il supporto della distribuzione di  $X$ . Quale funzione di  $\theta$ ,  $\mu(\theta)$  assume valore massimo in  $\theta = \theta_0$  (ricordando che  $\theta_0$  è il vero valore di  $\theta$ ) ed è conseguenza del fatto che, per ogni  $\theta \neq \theta_0$

$$\mu(\theta) - \mu(\theta_0) = \int_{\mathcal{S}_X} \ln \left( \frac{f(x; \theta)}{f(x; \theta_0)} \right) f(x; \theta_0) dx < \int_{\mathcal{S}_X} \left\{ \left( \frac{f(x; \theta)}{f(x; \theta_0)} \right) - 1 \right\} f(x; \theta_0) dx \quad (10.8)$$

dal momento che, per ogni  $u \neq 1$ , si ha  $\ln(u) < u - 1$ . Ma l'ultimo integrale in (10.8) è uguale a zero sicché si ha che  $\mu(\theta) < \mu(\theta_0)$  per ogni  $\theta \neq \theta_0$  o, in poche parole,  $\mu(\theta)$  raggiunge il massimo per  $\theta = \theta_0$ .

Ora, in virtù del fatto che  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  è un campione casuale, possiamo scrivere

$$\ln f_{\mathbf{X}}(\mathbf{X}; \theta) = \ln \left( \prod_{i=1}^n f(X_i; \theta) \right) = \sum_{i=1}^n \ln f(X_i; \theta) = \sum_{i=1}^n Z_i(\theta)$$

e le variabili casuali  $Z_i(\theta)$  sono indipendenti, ciascuna distribuita come la variabile casuale  $Z(\theta)$  poc'anzi definita. Ma per la Legge dei Grandi Numeri

$$\frac{1}{n} \sum_{i=1}^n Z_i(\theta) \rightarrow \mu(\theta) = E_\theta(Z(\theta)) \quad \text{quando } n \rightarrow \infty,$$

cioè, per  $n$  grande,  $\frac{1}{n} \ln f_{\mathbf{X}}(\mathbf{X}; \theta)$  è vicino a  $\mu(\theta)$  quale che sia  $\theta$ . Conseguentemente, il punto in cui  $\frac{1}{n} \ln f_{\mathbf{X}}(\mathbf{X}; \theta)$  assume valore massimo, vale a dire  $\hat{\theta}_n$ , deve essere vicino al punto in cui  $\mu(\theta)$  è massima ovvero deve essere vicino a  $\theta_0$  a patto che la convergenza di  $\frac{1}{n} \ln f_{\mathbf{X}}(\mathbf{X}; \theta)$  a  $\mu(\theta)$  sia uniforme in  $\theta$ . Ma allora  $\hat{\theta}_n \xrightarrow{P} \theta_0$  il che implica la consistenza dello stimatore di massima verosimiglianza  $\hat{\theta}_n$  di  $\theta$ .

2) Assumendo ora che  $\hat{\theta}_n$  sia un punto di svolta della funzione di log-verosimiglianza  $\ell(\theta; \mathbf{x}) = \ln L(\theta; \mathbf{x})$ , vale a dire

$$0 = \frac{d}{d\theta} \ell(\hat{\theta}_n; \mathbf{x}) \quad (10.9)$$

ed espandendo il membro di destra di (10.9) intorno a  $\theta_0$ , si ha

$$0 = \frac{d}{d\theta} \ell(\theta_0; \mathbf{x}) + (\hat{\theta}_n - \theta_0) \frac{d^2}{d\theta^2} \ell(\theta_0; \mathbf{x}) + R. \quad (10.10)$$

Dal momento che  $\hat{\theta}_n$  è stimatore consistente, per  $n$  grande il valore di  $\hat{\theta}_n$  sarà prossimo a  $\theta_0$  sicché, con elevata probabilità, la quantità  $R$  in (10.10) sarà di valore trascurabile e potrà essere ignorata. Perciò

$$\begin{aligned} \left( \hat{\theta}_n - \theta_0 \right) &= - \left[ \frac{d^2}{d\theta^2} \ell(\theta_0; \mathbf{x}) \right]^{-1} \frac{d}{d\theta} \ell(\theta_0; \mathbf{x}) \\ &= - \left[ \frac{d^2}{d\theta^2} \ell(\theta_0; \mathbf{x}) \right]^{-1} S(\theta_0; \mathbf{X}) \end{aligned}$$

o

$$\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) = - \left[ \frac{1}{n} \frac{d^2}{d\theta^2} \ell(\theta_0; \mathbf{x}) \right]^{-1} \frac{1}{\sqrt{n}} S(\theta_0; \mathbf{X}) \quad (10.11)$$

dove  $S(\theta_0; \mathbf{X})$  è la score function valutata in  $\theta = \theta_0$ .

Poichè le componenti  $X_1, X_2, \dots, X_n$  di un campione casuale sono indipendenti e identicamente distribuite, allora

$$f_{\mathbf{X}}(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

e

$$\ln f_{\mathbf{X}}(\mathbf{x}; \theta) = \sum_{i=1}^n \ln f(x_i; \theta) \quad (10.12)$$

sicchè

$$S(\theta; \mathbf{x}) = \sum_{i=1}^n \frac{d}{d\theta} \ln f(x_i; \theta).$$

Perciò

$$\frac{1}{\sqrt{n}} S(\theta_0; \mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i, \quad (10.13)$$

dove  $S_i = \frac{d}{d\theta} \ln f(X_i; \theta) |_{\theta=\theta_0}$ , per  $i = 1, 2, \dots, n$ , sono a loro volta v.c. indipendenti e identicamente distribuite. Ma allora, ricordando che  $\mathbb{E}(S_i) = 0$  e  $\mathbb{V}ar(S_i) = I_1(\theta_0)$  per  $i = 1, 2, \dots, n$ , per il Teorema Limite Centrale si ha

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n S_i \xrightarrow{d} N(0, \mathbb{V}ar(S_i)) = N(0, I_1(\theta_0)) \quad (10.14)$$

Vale la pena notare che la varianza di  $S_i$  valutata in  $\theta_0$  coincide con l'informazione di Fisher contenuta in una osservazione.

Inoltre, da (10.11) e (10.12), per la Legge dei Grandi Numeri, si ottiene

$$\frac{1}{n} \frac{d^2}{d\theta^2} \ell(\theta_0; \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{d^2}{d\theta^2} \ln f(x_i; \theta_0) \xrightarrow{P} \mathbb{E} \left( \frac{d^2}{d\theta^2} \ln f(x_i; \theta_0) \right) = -I_1(\theta_0). \quad (10.15)$$

e quindi, per il teorema del continuous mapping,

$$\left[ -\frac{1}{n} \frac{d^2}{d\theta^2} \ell(\theta_0; \mathbf{x}) \right]^{-1} \xrightarrow{P} [I_1(\theta_0)]^{-1}. \quad (10.16)$$

Infine, da (10.11), (10.14), (10.16), unitamente al teorema di Slutsky, si ha

$$\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{d} N \left( 0, \frac{1}{I_1(\theta_0)} \right)$$

da cui segue il risultato cercato

$$\hat{\theta}_n \xrightarrow{d} N \left( \theta_0, \frac{1}{I_n(\theta_0)} \right). \quad (10.17)$$

□

**Esempio 10.1.3.** La distribuzione di Pareto (Vilfredo Pareto, ingegnere, economista e sociologo italiano: Parigi, 15 luglio 1848 - Céligny, 19 agosto 1923) è stata frequentemente usata in ambito economico come modello per distribuzioni con *code* caratterizzate da un *decadimento lento*. In particolare, diremo che la v.c. continua  $X$  segue una distribuzione di Pareto di parametro  $\theta > 1$  se

$$f_X(x; \theta) = \theta x_0^\theta x^{-(\theta+1)} \mathbb{1}_{[x_0, +\infty)}(x) \quad (10.18)$$

con  $x_0$  quantità nota.

Supponiamo ora che  $(X_1, X_2, \dots, X_n)$  sia un campione casuale proveniente da una distribuzione di Pareto di parametro  $\theta > 1$ ; vogliamo trovare lo stimatore di massima verosimiglianza di  $\theta$  e la sua distribuzione asintotica.

La funzione di verosimiglianza, in questo caso, è data da

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n \theta x_0^\theta x_i^{-(\theta+1)} \mathbb{1}_{[x_0, +\infty)}(x_i) = \theta^n x_0^{n\theta} \left( \prod_{i=1}^n x_i \right)^{-(\theta+1)} \prod_{i=1}^n \mathbb{1}_{[x_0, +\infty)}(x_i)$$

da cui la log-verosimiglianza

$$\ell(\theta | \mathbf{x}) = n \ln(\theta) + n \ln(x_0) - (\theta + 1) \sum_{i=1}^n \ln(x_i) + \ln \prod_{i=1}^n \mathbb{1}_{[x_0, +\infty)}(x_i).$$

Lo stimatore di massima verosimiglianza si ottiene risolvendo l'equazione in  $\theta$  ottenuta equagliando a zero la score function  $S(\theta; \mathbf{x})$  ovvero

$$S(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta | \mathbf{x}) = \frac{n}{\theta} + n \ln(x_0) - \sum_{i=1}^n \ln(x_i) = 0$$

da cui

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^n \ln(x_i) - n \ln(x_0)} = \frac{n}{\sum_{i=1}^n \ln \left( \frac{x_i}{x_0} \right)} \quad (10.19)$$

L'informazione di Fisher è data da

$$I_n(\theta) = -\mathbb{E}_\theta \left[ \frac{d^2}{d\theta^2} \ell(\theta | \mathbf{X}) \right] = -\mathbb{E}_\theta \left[ \frac{d}{d\theta} S(\theta; \mathbf{X}) \right] = -\mathbb{E}_\theta \left[ -\frac{n}{\theta^2} \right] = \frac{n}{\theta^2}$$

sicchè in virtù della (10.17)

$$\hat{\theta}_n \underset{a}{\sim} N\left(\theta, \frac{1}{I_n(\theta)}\right) \equiv N\left(\theta, \frac{\theta^2}{n}\right). \quad (10.20)$$

Processi inferenziali (stime intervallari, test per la verifica di ipotesi,...) riguardo  $\theta$  possono essere costruiti a partire da  $\hat{\theta}_n$ , sfruttando approssimazione appena trovata della sua distribuzione campionaria.

Ribadiamo ancora una volta, se ce ne fosse bisogno, che il *limite inferiore di Rao-Cramèr* per la varianza di uno stimatore non distorto  $T_n$  del parametro  $\theta$  è pur sempre *solo un limite inferiore* e nulla garantisce che esso sia raggiunto dalla varianza di  $T_n$ , nè per  $n$  finito ma neanche asintoticamente.

Come stabilisce il Teorema 9.3.1, vi è tuttavia un caso in cui una particolare forma della score function garantisce l'esistenza di uno stimatore efficiente di  $\theta$ ; inoltre, se esiste uno stimatore efficiente di  $\theta$ , il Teorema 9.3.2 stabilisce che questo deve essere proprio lo stimatore di massima verosimiglianza  $\hat{\theta}_n$ .

**Esempio 10.1.4.** Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione  $N(\mu, \sigma^2)$ , con  $\mu \in \mathbb{R}$  e  $\sigma^2 > 0$ . Abbiamo già ricavato gli stimatori di massima verosimiglianza dei parametri della distribuzione Normale dati da

$$\hat{\mu}_n = \bar{X}_n \quad \hat{\sigma}_n^2 = \frac{n-1}{n} S_n^2 \quad (10.21)$$

Inoltre, il limite inferiore della varianza di uno stimatore non distorto di  $\sigma^2$  è dato da  $[I_n(\sigma^2)]^{-1} = 2\sigma^4/n$ . Calcoliamo ora la varianza dello stimatore di massima verosimiglianza di  $\sigma^2$ :

$$\begin{aligned} \text{Var}_{\sigma^2}(\hat{\sigma}_n^2) &= \text{Var}_{\sigma^2}\left(\frac{n-1}{n} S_n^2\right) = \frac{(n-1)^2}{n^2} \text{Var}_{\sigma^2}(S_n^2) \\ &= \frac{(n-1)^2}{n^2} \frac{2\sigma^4}{n-1} \\ &= \frac{2(n-1)}{n^2} \sigma^4 \end{aligned} \quad (10.22)$$

Pertanto lo stimatore di massima verosimiglianza di  $\sigma^2$  non è stimatore efficiente di  $\sigma^2$  ma comunque lo è *asintoticamente*. E questo vale in generale: spesso gli stimatori di massima verosimiglianza non sono stimatori efficienti del parametro  $\theta$  ma lo sono certamente asintoticamente.

**Corollario 10.1.1.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione  $F(x; \theta)$  con  $\theta$  scalare. Sotto condizioni assai blande di regolarità

$$\frac{1}{\sqrt{n}} S(\mathbf{X}; \theta) \rightarrow N(0, I_1(\theta)), \quad \text{quando } n \rightarrow \infty$$

dove  $S(\mathbf{X}; \theta)$  è la *statistica score*; in altre parole, per  $n$  grande,

$$S(\mathbf{X}; \theta) \xrightarrow{d} N(0, I_n(\theta)).$$

**OSS:** se la distribuzione fosse indicizzata da *due o più parametri*, ovvero  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  con  $k \geq 2$  e indicassimo con  $\hat{\theta}_n(X_1, X_2, \dots, X_n)$  lo stimatore di massima verosimiglianza di  $\theta \in \Theta \subseteq \mathbb{R}^k$ , avremo

$$\hat{\theta}_n \sim N_k(\theta_0, \mathbf{H}_n^{-1}(\theta_0)) \text{ quando } n \rightarrow \infty$$

dove, secondo quanto abbiamo visto a lezione,  $\mathbf{H}_n(\theta_0)$  è la matrice di informazione di Fisher valutata in  $\theta = \theta_0$ .

Possiamo estendere le conclusioni del Teorema 10.1.3 al caso di funzioni del parametro  $\theta$ .

**Teorema 10.1.4.** Sotto condizioni di regolarità assai blande, usualmente soddisfatte da molte famiglie di distribuzioni, lo stimatore di massima verosimiglianza  $\tau(\hat{\theta}_n)$  di una funzione continua reale  $\tau(\theta)$  di  $\theta$  è

1) *consistente*

$$P\left(|\tau(\hat{\theta}_n) - \tau(\theta_0)| > \epsilon\right) \rightarrow 0 \text{ quando } n \rightarrow \infty$$

dove  $\theta_0$  è il vero valore del parametro  $\theta$ ;

2) *asintoticamente efficiente, non distorto e normalmente distribuito*, ovvero

$$\tau(\hat{\theta}_n) \sim N\left(\tau(\theta_0), \frac{[\tau'(\theta_0)]^2}{I_n(\theta_0)}\right) \text{ quando } n \rightarrow \infty$$

dove  $I_n(\theta_0)$  è l'informazione di Fisher e  $\tau'(\theta_0)$  è la derivata prima rispetto a  $\theta$  di  $\tau(\theta)$  entrambe valutate in  $\theta = \theta_0$

*Dimostrazione.* La dimostrazione di questo teorema può essere facilmente ottenuta aggiustando la dimostrazione del teorema precedente oppure usando (a) la proprietà di invarianza degli stimatori di massima verosimiglianza, (b) il Teorema 10.1.3 e (c) il fatto che, data una funzione continua reale  $\tau(\theta)$  di  $\theta$ , l'informazione di Fisher riparametrizzata per  $\tau(\theta)$  è data da

$$I_n(\tau(\theta)) = \left[ \frac{d}{d\theta} \tau(\theta) \right]^{-2} I_n(\theta)$$

in cui reciproco, valutato in  $\theta = \theta_0$ , restituisce la varianza asintotica di  $\tau(\hat{\theta}_n)$ ,  $\frac{[\tau'(\theta_0)]^2}{I_n(\theta_0)}$ .  $\square$

**Esempio 10.1.5.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $b(1, \theta)$  e sia

$$g(\theta) = \ln\left(\frac{\theta}{1-\theta}\right) \tag{10.23}$$

Vogliamo stimare  $g(\theta)$  alla massima verosimiglianza. Per farlo, possiamo utilizzare lo stimatore di massima verosimiglianza  $\hat{\theta}_n$  di  $\theta$  e definire

$$g(\hat{\theta}_n) = \ln\left(\frac{\hat{\theta}_n}{1-\hat{\theta}_n}\right) \tag{10.24}$$

Abbiamo già dimostrato che per il teorema del limite centrale  $\hat{\theta}_n \sim \left( \theta, \frac{\theta(1-\theta)}{n} \right)$ . Pertanto, applicando il metodo delta, otteniamo che

$$g(\hat{\theta}_n) \stackrel{a}{\sim} N \left( \ln \left( \frac{\theta}{1-\theta} \right); \frac{1}{n\theta(1-\theta)} \right) \quad (10.25)$$

La funzione di massima verosimiglianza e l'informazione di Fisher sono rispettivamente

$$L(\theta; \mathbf{x}) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \quad I_n = -\frac{d^2}{d\theta^2} l(\theta; \mathbf{x}) = \frac{n}{\theta(1-\theta)} \quad (10.26)$$

Pertanto, applicando il Teorema 3.2.6, otteniamo che:

$$g(\hat{\theta}_n) \stackrel{a}{\sim} N \left( g(\theta); \frac{[g'(\theta)]^2}{I_n(\theta)} \right) \quad (10.27)$$

Concludiamo questo paragrafo con ultimo esempio in cui confrontiamo lo stimatore di massima verosimiglianza e lo stimatore UMVU per una funzione  $\eta(\theta)$  del parametro  $\theta$ . Svolgendo l'esempio avremo anche modo di ragionare sulle distribuzioni asintotiche dei due stimatori e in particolare, sfruttare quanto già sappiamo a proposito della distribuzione asintotica degli stimatori di massima verosimiglianza per dire qualcosa sulla distribuzione asintotica dello stimatore UMVU. Come abbiamo già avuto modo di dire, per gli stimatori UMVU non esiste una teoria asintotica così completa e strutturata come per gli stimatori di massima verosimiglianza, per cui ci appoggeremo a quest'ultima per dire qualcosa in merito al comportamento asintotico dei primi.

**Esempio 10.1.6.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $b(1, \theta)$ . Vogliamo stimare la funzione di  $\theta$

$$\eta(\theta) = \text{Var}_\theta(X) = \theta(1-\theta) \quad (10.28)$$

sia tramite lo stimatore di massima verosimiglianza che quello UMVU di  $\eta(\theta)$  e procedendo poi al loro confronto.

Per cominciare, cerchiamo una statistica sufficiente (minimale) e completa per il parametro  $\theta$  e, a questo scopo, calcoliamo la funzione di verosimiglianza

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i) \\ &= \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(x_i) \end{aligned} \quad (10.29)$$

da cui, per l'appartenenza della distribuzione generante il campione casuale a famiglia esponenziale a  $k = 1$ -parametri,  $T_n = \sum_{i=1}^n X_i$  sarà la statistica sufficiente (minimale) e completa cercata per  $\theta$ .

Già sappiamo che lo stimatore di massima verosimiglianza per  $\theta$  è  $\hat{\theta}_n = n^{-1} \sum X_i$ ; inoltre, per la proprietà di invarianza degli stimatori di massima verosimiglianza,

$$\hat{\eta}(\theta) = \eta(\hat{\theta}_n) = \hat{\theta}_n (1 - \hat{\theta}_n) = \frac{1}{n^2} T_n (n - T_n) \quad (10.30)$$

è lo stimatore di massima verosimiglianza per  $\eta(\theta)$ . Cerchiamo ora lo stimatore UMVU di  $\eta(\theta)$  sfruttando il corollario 9.4.1 del teorema di Rao-Blackwell; partiamo dunque dallo stimatore di massima verosimiglianza di  $\eta(\theta)$  che sappiamo essere funzione di statistica sufficiente (minimale) e completa  $T_n$  e verifichiamone la non distorsione:

$$\begin{aligned} E_\theta \left( \frac{1}{n^2} T_n (n - T_n) \right) &= \frac{1}{n^2} E_\theta(n T_n - T_n^2) = \frac{1}{n^2} [E_\theta(n T_n) - E_\theta(T_n^2)] \\ &= \frac{1}{n^2} (n^2\theta - n\theta(1-\theta) + n^2\theta^2) \\ &= \frac{1}{n^2} (n^2\theta - n\theta + n\theta^2 + n^2\theta^2) = \frac{n-1}{n} \theta(1-\theta) \end{aligned} \quad (10.31)$$

sicché

$$\tilde{\eta}(\theta) = \frac{n}{n-1} \eta(\hat{\theta}_n) \quad (10.32)$$

è uno stimatore non distorto di  $\eta(\theta)$  funzione di statistica sufficiente (minimale) completa, quindi è lo stimatore UMVU di  $\eta(\theta) = \theta(1-\theta)$ .

Sappiamo, per nota proprietà, che lo stimatore di massima verosimiglianza  $\hat{\eta}(\theta)$  è stimatore consistente di  $\eta(\theta)$ , ovvero

$$\hat{\eta}(\theta) \xrightarrow{P} \eta(\theta) \quad (10.33)$$

e poiché

$$\tilde{\eta}(\theta) - \hat{\eta}(\theta) = \frac{n}{n-1} \hat{\eta}(\theta) - \hat{\eta}(\theta) = \frac{1}{n-1} \hat{\eta}(\theta) \xrightarrow{P} 0 \quad (10.34)$$

anche lo stimatore UMVU  $\tilde{\eta}(\theta)$  sarà consistente per  $\eta(\theta)$ .

Notiamo inoltre che

$$\begin{aligned} \sqrt{n} [\tilde{\eta}(\theta) - \eta(\theta)] - \sqrt{n} [(\hat{\eta}(\theta) - \eta(\theta))] &= \sqrt{n} \left[ \frac{n}{n-1} \hat{\eta}(\theta) - \eta(\theta) \right] - \sqrt{n} [(\hat{\eta}(\theta) - \eta(\theta))] \\ &= \frac{\sqrt{n}}{n-1} \hat{\eta}(\theta) \xrightarrow{P} 0 \end{aligned}$$

perciò, tenuto conto che la convergenza in probabilità *implica* quella in distribuzione,  $\sqrt{n} [\tilde{\eta}(\theta) - \eta(\theta)]$  ha la stessa distribuzione asintotica di  $\sqrt{n} [\hat{\eta}(\theta) - \eta(\theta)]$ .

Valgono, infine, i seguenti limiti in distribuzione:

$$\sqrt{n} [(\hat{\eta}(\theta) - \eta(\theta))] \underset{a}{\sim} N(0, (1-2\theta)^2 \theta (1-\theta)) \quad (10.35)$$

quale prodotto del *Delta method*, tenuto conto del fatto che  $\eta(\theta) = \theta(1-\theta)$  è una funzione continua di  $\theta$  e che  $\frac{d}{d\theta}\eta(\theta) = 1-2\theta \neq 0$  purché  $\theta \neq \frac{1}{2}$ ; e conseguentemente,

$$\sqrt{n} [\tilde{\eta}(\theta) - \eta(\theta)] \underset{a}{\sim} N(0, (1-2\theta)^2 \theta (1-\theta)) \quad (10.36)$$

sempre che  $\theta \neq \frac{1}{2}$ .

## 10.2 Metodi numerici per la massima verosimiglianza

### 10.2.1 Tecnica di cattura-ricattura

Affrontiamo ora il problema di *stimare* l'incognita *numerosità*  $N$  di una popolazione di individui che non può essere censita direttamente e interamente. Questo tipo di problema assume una rilevanza notevole sia in ambito di bio-ecologico (stimare la numerosità di una specie animale o vegetale in un certo areale) che medico-epidemiologico (stimare la numerosità di una popolazione di soggetti affetti da una certa patologia) o ancora in ambito economico e di marketing (stimare la numerosità di una popolazione di consumatori di un certo bene o servizio).

La numerosità  $N$  della popolazione può essere stimata ricorrendo a una tecnica nota in letteratura come *tecnica di cattura-ricattura*. Essa era nota già a Laplace che la utilizzò nel 1802 per stimare la popolazione dell'intera Francia sulla base di una sua stima dei nati in Francia in un anno e dei dati di nati e residenti in alcune comunità francesi, particolarmente ordinate e accurate nella gestione amministrativa dei dati di censimento. La tecnica fu perfezionata e introdotta ufficialmente in letteratura dal biologo marino danese Carl G. J. Petersen che, grazie a essa, nel 1896 stimò la presenza numerica della popolazione marina delle platesse in un determinato tratto di mare.

In buona sostanza, questa tecnica sfrutta la possibilità di plottare la funzione di verosimiglianza in funzione del parametro  $N$ , incognita numerosità della popolazione e nell'analisi del grafico risultante proprio alla ricerca del valore che la rende massima.

Operativamente si tratta di applicare la seguente procedura:

- a) estrarre dalla popolazione un sottoinsieme di  $N_1$  elementi
- b) contrassegnare ciascun elemento estratto con una marcatura e poi re-inserirlo nella popolazione; fatto ciò, la popolazione orignaria sarà ora composta da due distinte sottopopolazioni: quella degli individui marcati, di dimensione  $N_1$  e quella degli individui non marcati, di dimensione  $(N - N_1)$
- c) successivamente, estrarre dalla popolazione quale quella risultante in b), un campione casuale di  $n$  elementi.
- d) indicato con  $n_1$  il numero di elementi marcati del campione casuale di ampiezza  $n$ , costruire la funzione di verosimiglianza associata ad  $n_1$

$$L(N|n_1) = \frac{\binom{N_1}{n_1} \binom{N-N_1}{n-n_1}}{\binom{N}{n}} = g(N) \quad (10.37)$$

che risulterà essere una funzione della sola quantità incognita  $N$ . Il ricorso alla *distribuzione ipergeometrica* è giustificato dal fatto che le estrazioni si eseguono *in blocco*

- e) per valori di  $N > N_1 + n - n_1$ , disegnare il grafico di  $L(N|n_1)$  e, procedendo con un'ispezione grafica, individuare quel valore di  $N$  in corrispondenza del quale

la verosimiglianza è massima, ottenendo dunque lo stimatore di massima verosimiglianza della numerosità  $N$  della popolazione

$$\hat{N}_n = \arg \max_{N \in \mathbb{N}} L(N|n_1), \text{ dove } N_1 \geq n_1 \quad (10.38)$$

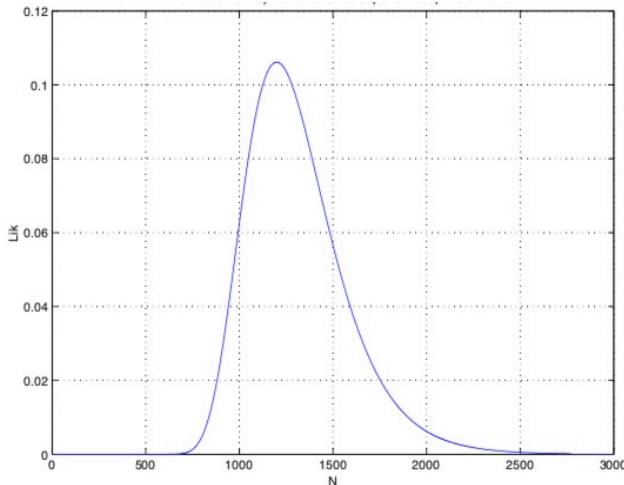
valore di  $N$  che è *massimamente verosimile* in concomitanza con le osservazioni, e dunque l'informazione, disponibili; o, in altre parole, il valore  $\hat{N}_n$  è quello che è più compatibile con il risultato osservato (30 soggetti marcatai si 80 estratti).

Con i dati a disposizione, l'intera procedura oggi richiede poche righe di codice, in un qualsiasi linguaggio di programmazione (R, Matlab, Fortran,...), per essere efficacemente implementata.

**Esempio 10.2.1.** Supponiamo di essere interessati a stimare la numerosità della popolazione di un certo tipo di pesci in uno stagno. Volendo applicare la tecnica di cattura e ricattura, fissiamo in  $N_1 = 300$  il numero di individui da prelevare e marcare; rilasciati gli individui marcatai in libertà e trascorso il tempo necessario per una omogeneizzazione delle due sottopololazioni (individui marcatai e non marcatai), estraiamo dallo stagno un campione casuale di  $n = 80$  individui,  $n_1 = 30$  dei quali risultano marcatai. Allora

$$L(N|n_1 = 30) = \frac{\binom{300}{30} \binom{N-300}{80-30}}{\binom{N}{80}} = g(N) \quad (10.39)$$

che possiamo plottare per valori di  $N > 300 + 80 - 30 = 350$ , ottenendo,



**Figura 10.1:** Stima della numerosità della popolazione via *tecnica di cattura e ricattura*

sicché, dall'ispezione del grafico della funzione di verosimiglianza, alla ricerca del massimante, possiamo concludere che  $\hat{N}_n = 1200$ .

## 10.2.2 Metodo di Newton-Raphson

Il *metodo di Newton*, anche conosciuto anche come *metodo delle tangenti* o ancora *algoritmo di Newton-Raphson*, è procedimento che produce *approssimazioni* via via

sempre migliori degli *zeri* di una funzione a valori reali a partire da una loro stima iniziale. Essenzialmente esso

- a) è un metodo *iterativo*
- b) quando esiste soluzione e si ha una *buona inizializzazione* della procedura ricorsiva, converge velocemente (usualmente, *quadraticamente* ovvero *raddoppiando il numero di cifre corrette a ogni iterazione*) alla soluzione.

In poche parole, l'algoritmo di Newton-Raphson cerca di risolvere un'equazione di forma  $m(x) = 0$  attraverso una procedura che poggia sullo *sviluppo in serie di Taylor* di  $m(x)$  attorno a  $x_0$  e che approssima la funzione  $m(x)$  con il suo sviluppo arrestato al primo ordine

$$m(x) \simeq m(x_0) + (x - x_0) m'(x_0) \quad (10.40)$$

da cui

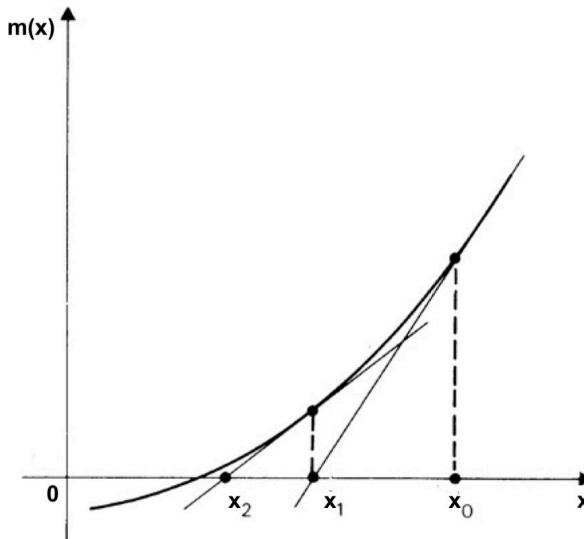
$$x = x_0 - \frac{m(x_0)}{m'(x_0)} \quad (10.41)$$

Da quest'ultima si può ricavare infine la seguente relazione ricorsiva

$$x_h = x_{h-1} - \frac{m(x_{h-1})}{m'(x_{h-1})}, \quad h = 1, 2, \dots \quad (10.42)$$

che produce una approssimazione della radice  $x$  di  $m(x)$  ottenuta a partire da una *stima iniziale*  $x_0$  di  $x$ , detta *valore di inizializzazione* della procedura.

Mediante questa procedura, costruiamo una successione convergente di zeri; possiamo quindi definire un *livello di precisione*  $\varepsilon > 0$  e interrompere l'esecuzione della procedura se  $x_h - x_{h-1} < \varepsilon$ .



**Figura 10.2:** Metodo delle tangenti (o *algoritmo di Newton-Raphson*): successione degli zeri  $x_0, x_1, x_2, \dots$  via via convergenti alla soluzione dell'equazione  $m(x) = 0$

Nell'ambito della stima di massima verosimiglianza l'algoritmo di Newton-Raphson trova applicazione quando le equazioni di stima (le cosiddette *score equation*) non sono *lineari nei parametri*, cosa abbastanza frequente in molti modelli comunemente impiegati nelle applicazioni.

Supponendo per il momento  $\theta$  scalare, l'equazione alla base del procedimento ricorsivo diventa

$$m(\theta; \mathbf{x}) = \frac{d}{d\theta} \ell(\theta | \mathbf{x}) = 0 \quad (10.43)$$

e

$$\theta^{(h)} = \theta^{(h-1)} - \frac{\ell'(\theta^{(h-1)} | \mathbf{x})}{\ell''(\theta^{(h-1)} | \mathbf{x})}, \quad h = 1, 2, \dots \quad (10.44)$$

sicché per  $|\theta^{(h)} - \theta^{(h-1)}| < \varepsilon$  interromperemo la procedura ponendo  $\hat{\theta}_n = \theta^{(h)}$ .

Dal momento che l'algoritmo è decisamente performante in presenza di funzioni regolari e di un buon punto di partenza  $\theta_0$  spesso si sceglie

$$\theta^{(0)} = \tilde{\theta}_n(X_1, X_2, \dots, X_n) \quad (10.45)$$

dove  $\tilde{\theta}_n$  è lo stimatore di  $\theta$  ottenuto con il *metodo dei momenti*.

Questa procedura può essere facilmente estesa al caso di un vettore di più parametri  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$ . In tal caso, l'equazione alla base del procedimento ricorsivo diventa

$$m(\boldsymbol{\theta}; \mathbf{x}) = \frac{d}{d\boldsymbol{\theta}} \ell(\boldsymbol{\theta} | \mathbf{x}) = 0 \quad (10.46)$$

e

$$\boldsymbol{\theta}^{(h)} = \boldsymbol{\theta}^{(h-1)} - \mathbf{H}_n^{-1}(\boldsymbol{\theta}^{(h-1)}) \mathbf{u}_n(\boldsymbol{\theta}^{(h-1)}), \quad h = 1, 2, \dots \quad (10.47)$$

dove  $\mathbf{H}_n(\boldsymbol{\theta}^{(h-1)})$  è la matrice delle derivate seconde di  $\ell(\boldsymbol{\theta} | \mathbf{x})$  rispetto  $\boldsymbol{\theta}$  (o *matrice hessiana* valutata in  $\boldsymbol{\theta}^{(h-1)}$  mentre  $\mathbf{u}_n(\boldsymbol{\theta}^{(h-1)})$  è il vettore delle derivate prime di  $\ell(\boldsymbol{\theta} | \mathbf{x})$  rispetto  $\boldsymbol{\theta}$  (o *vettore gradiente*) anch'esso valutato in  $\boldsymbol{\theta}^{(h-1)}$ .

Spesso al posto di  $-\mathbf{H}_n(\boldsymbol{\theta}^{(h-1)})$  si utilizza la matrice di informazione di Fisher  $\mathbf{I}_n(\boldsymbol{\theta}^{(h-1)})$ . Così facendo, si ottiene il metodo iterativo noto come *scoring di Fisher*.

**Esempio 10.2.2.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $\mathcal{G}(\alpha, \beta)$  con  $\alpha, \beta > 0$ . Vogliamo determinare lo stimatore di massima verosimiglianza del parametro  $\eta = \beta/\alpha$ .

La funzione di verosimiglianza è data da

$$\begin{aligned} L(\alpha, \beta | \mathbf{x}) &= \prod_{i=1}^n \frac{1}{\Gamma(\alpha) \beta^\alpha} x_i^{\alpha-1} e^{\frac{x_i}{\beta}} \mathbb{1}_{\mathbb{R}^+}(x_i) \\ &= \left[ \frac{1}{\Gamma(\alpha) \beta^\alpha} \right]^n \left[ \prod_{i=1}^n x_i \right]^{\alpha-1} \exp \left\{ \frac{1}{\beta} \sum_{i=1}^n x_i \right\} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \end{aligned}$$

sicché la funzione di log-verosimiglianza è data da

$$\ell(\alpha, \eta | \mathbf{x}) = -n \ln(\Gamma(\alpha)) - n\alpha \ln(\beta) + (\alpha - 1) \sum_{i=1}^n \ln(x_i) - \frac{1}{\beta} \sum_{i=1}^n x_i + \ln \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i)$$

La famiglia delle distribuzioni Gamma è regolare sicché gli stimatori di massima verosimiglianza di  $\alpha$  e  $\beta$ ,  $\hat{\alpha}_n$  e  $\hat{\beta}_n$ , si ottengono quali soluzioni del seguente sistema di *equazioni non lineari nei parametri*  $\alpha$  e  $\beta$

$$\begin{cases} \frac{d}{d\alpha} \ell(\alpha, \beta | \mathbf{x}) = -n \frac{\psi_0(\alpha)}{\Gamma(\alpha)} - n \ln(\beta) + \sum_{i=1}^n \ln(x_i) = 0 \\ \frac{d}{d\beta} \ell(\alpha, \beta | \mathbf{x}) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i = 0 \end{cases} \quad (10.48)$$

con  $\psi_0(\alpha) = \frac{d}{d\alpha} \Gamma(\alpha)$  funzione digamma, derivata prima rispetto  $\alpha$  della funzione Gamma  $\Gamma(\alpha)$ . Ottenuti gli stimatori di massima verosimiglianza di  $(\alpha, \beta)$ , lo stimatore di massima verosimiglianza di  $\eta = \alpha/\beta$ , in virtù della proprietà di invarianza 10.1, sarà  $\hat{\eta}_n = \hat{\alpha}_n/\hat{\beta}_n$ .

Per risolvere il sistema di equazioni (10.48) dobbiamo ricorrere a metodi di soluzione numerica quali di Newton-Rapshon o lo scoring di Fisher, per cui dobbiamo determinare *gradiente* e *matrice hessiana* che sono dati da

$$\mathbf{u}(\alpha, \beta) = \begin{pmatrix} \frac{d}{d\alpha} \ell(\alpha, \beta | \mathbf{x}) \\ \frac{d}{d\beta} \ell(\alpha, \beta | \mathbf{x}) \end{pmatrix} = \begin{pmatrix} -n \frac{\psi_0(\alpha)}{\Gamma(\alpha)} - n \ln(\beta) + \sum_{i=1}^n \ln(x_i) \\ -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i \end{pmatrix}$$

e

$$\mathbf{H}(\alpha, \beta) = \begin{pmatrix} \frac{d^2}{d\alpha^2} \ell(\alpha, \beta | \mathbf{x}) & \frac{d}{d\alpha d\beta} \ell(\alpha, \beta | \mathbf{x}) \\ \frac{d}{d\alpha d\beta} \ell(\alpha, \beta | \mathbf{x}) & \frac{d^2}{d\beta^2} \ell(\alpha, \beta | \mathbf{x}) \end{pmatrix} = \begin{pmatrix} -n \frac{\psi_1(\alpha) \Gamma(\alpha) - [\psi_0(\alpha)]^2}{[\Gamma(\alpha)]^2} & -\frac{n}{\beta} \\ -\frac{n}{\beta} & \frac{n\alpha}{\beta^2} - \frac{2}{\beta^3} \sum_{i=1}^n x_i \end{pmatrix}$$

con  $\psi_1(\alpha) = \frac{d^2}{d\alpha^2} \Gamma(\alpha)$  funzione trigamma, derivata seconda rispetto  $\alpha$  della funzione Gamma  $\Gamma(\alpha)$ .

Fissato piccolo a piacere il valore del livello di precisione  $\varepsilon > 0$ , siamo ora in grado di scrivere la relazione ricorsiva da *reiterare fino a convergenza* e che produrrà la soluzione numerica dei sistemi di equazioni (10.48); per semplicità di notazione poniamo  $\boldsymbol{\theta}^{(h)} = (\alpha^{(h)}, \beta^{(h)})$

$$\boldsymbol{\theta}^{(h)} = \boldsymbol{\theta}^{(h-1)} - \mathbf{H}^{-1}(\boldsymbol{\theta}^{(h-1)}) \mathbf{u}(\boldsymbol{\theta}^{(h-1)}) \quad (10.49)$$

dove  $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}_n$  ovvero, inizializziamo la relazione ricorsiva, cuore dell'algoritmo di Newton-Rapson, con lo stimatore  $\tilde{\boldsymbol{\theta}}_n$  di  $\boldsymbol{\theta} = (\alpha, \beta)$  ottenuto con il metodo dei momenti e riportato in 2.25

Si può subito osservare che la matrice di informazione di Fisher per  $\boldsymbol{\theta} = (\alpha, \beta)$  è data da

$$\mathbf{I}_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{H}_n) = \begin{pmatrix} n \frac{\psi_1(\alpha) \Gamma(\alpha) - [\psi_0(\alpha)]^2}{[\Gamma(\alpha)]^2} & \frac{n}{\beta} \\ \frac{n}{\beta} & \frac{n\alpha}{\beta^2} \end{pmatrix} \quad (10.50)$$

sicché, sempre per la proprietà di invarianza,

$$\mathbf{I}_n^{-1}(\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} n \frac{\psi_1(\hat{\alpha}_n) \Gamma(\hat{\alpha}_n) - [\psi_0(\hat{\alpha}_n)]^2}{[\Gamma(\hat{\alpha}_n)]^2} & \frac{n}{\hat{\beta}_n} \\ \frac{n}{\hat{\beta}_n} & \frac{n\hat{\alpha}_n}{\hat{\beta}_n^2} \end{pmatrix}^{-1} \quad (10.51)$$

restituisce una *stima* di massima verosimiglianza della *matrice di covarianza asintotica* di  $\hat{\boldsymbol{\theta}}_n = (\hat{\alpha}_n, \hat{\beta}_n)$ , stimatori di massima verosimiglianza di  $(\alpha, \beta)$ .

### 10.2.3 Modello di regressione logistica

Il modello di *regressione logistica* trova applicazione quando si ritiene che la *probabilità* di un certo evento sia *influenzata* dal valore assunto da un'altra (o da più di un'altra) variabile (modello di regressione logistica semplice e modello di regressione logistica multipla).

In termini generali, dunque, si vuole *spiegare* la relazione di *dipendenza* di una variabile risposta *dicotomica*  $Y$  da *una o più* variabili indipendenti o esplicative (o, regressori)  $X_1, X_2, \dots, X_p$  di natura qualsiasi (quantitativa o qualitativa).

Alcuni ambiti di applicazione della regressione logistica (semplice):

- a) *sopravvivenza* (o meno) ( $Y$ ), a distanza di un certo lasso di tempo, dalla somministrazione di un farmaco o trattamento alla luce del dosaggio somministrato ( $x$ )
- b) *restituzione* (o meno) ( $Y$ ) di un prestito, a scadenza, da parte di chi lo ha ricevuto alla luce del reddito (o altra caratteristica) del medesimo ( $x$ )

Le ragioni che conducono al modello di regressione logistica sono molteplici; tra queste,

- a) individuare, tra le variabili ritenute esplicative, quelle che meglio si prestano a *spiegare la presenza o l'assenza* dell'attributo studiato rappresentato dalla variabile risposta  $Y$  (a seconda che le prime siano correlate *positivamente* o *negativamente* con  $Y$  possono essere considerate *fattori di rischio* o *fattori di protezione*)
- b) ricercare una *combinazione lineare* delle variabili esplicative che meglio *discrimina* tra il gruppo degli individui che posseggono l'attributo e quello degli individui che non lo posseggono
- c) *stimare la probabilità di possesso* dell'attributo per una *nuova* unità statistica su cui è stato osservato il vettore delle variabili esplicative  $x_1, x_2, \dots, x_p$  e, fissato un *valore soglia* per tale probabilità, *classificare* la nuova unità statistica associandola al gruppo dei possessori o a quello dei non possessori dell'attributo.

Siano  $Y_1, Y_2, \dots, Y_n$  v.c. bernoulliane  $b(1, \theta_i)$ ,  $\theta_i \in (0, 1)$  per ogni  $i = 1, 2, \dots, n$  *indipendenti* (ma come si vede immediatamente, *non identicamente distribuite*) in quanto la *probabilità di successo*  $\theta_i$  varia in funzione delle caratteristiche dell'unità statistica a cui si riferisce, descritte in termini di una (o più) variabile esplicativa  $x$  che assume valori  $x_1, x_2, \dots, x_n$  sulle  $n$  unità statistiche. In particolare, assumiamo,

$$\theta_i = P_{\beta_0, \beta_1}(Y_i = 1 | x = x_i) = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}, \quad i = 1, 2, \dots, n \quad (10.52)$$

dove  $x_1, x_2, \dots, x_n$  sono valori osservati della variabile esplicativa (o regressore)  $x$  e  $\beta_0, \beta_1$  parametri incogniti che modellano la relazione tra  $Y$  e  $x$ .

Stando, a finora quanto assunto, la relazione che descrive la *dipendenza* della probabilità di successo (o accadimento)  $\theta_i$  dalla variabile esplicativa  $x$  segue una *distribuzione logistica*.

Definiamo la funzione *logit* (o *logOdds*) di  $w$  nella maniera seguente:

$$\text{logit}(w) = \ln\left(\frac{w}{1-w}\right), \quad 0 < w < 1 \quad (10.53)$$

sicché

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 x_i, \quad i = 1, 2, \dots, n \quad (10.54)$$

La funzione di verosimiglianza relativa al vettore dei valori osservati della variabile risposta  $Y$ ,  $\mathbf{y} = (y_1, y_2, \dots, y_n)$ , è data da

$$\begin{aligned} L(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) &= \prod_{i=1}^n \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \mathbb{1}_{\{0,1\}}(y_i) \\ &= \prod_{i=1}^n \left[ \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \right]^{y_i} \left[ 1 - \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \right]^{1-y_i} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i) \\ &= \left[ \prod_{i=1}^n (1 + \exp\{\beta_0 + \beta_1 x_i\}) \right]^{-1} \exp \left\{ \beta_0 \sum_{i=1}^n y_i + \beta_1 \sum_{i=1}^n x_i y_i \right\} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i) \end{aligned}$$

ed emerge da quest'ultima espressione della funzione di verosimiglianza la struttura di una *famiglia esponenziale* a  $k = 2$  parametri sicché possiamo subito concludere, sulla base del Teorema 8.2.2 e nel Teorema 9.4.2 che la statistica

$$\mathbf{T}_n = \left( \sum_{i=1}^n Y_i, \sum_{i=1}^n x_i Y_i \right) \quad (10.55)$$

è *congiuntamente sufficiente* (minimale) e *completa* per il parametro  $\beta = (\beta_0, \beta_1)$ . In conseguenza, è del tutto ragionevole pensare di appoggiare su di essa una buona inferenza riguardo a  $\beta = (\beta_0, \beta_1)$  e lo stimatore di massima verosimiglianza  $\hat{\beta}_n = (\hat{\beta}_0, \hat{\beta}_1)$  che ne è funzione (vedere Teorema 10.1.1), si candida naturalmente a essere il perno (o *pivot*) di procedure inferenziali ottimali.

La funzione di log-verosimiglianza è data da

$$\ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = \ln \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i) - \sum_{i=1}^n \ln(1 + \exp\{\beta_0 + \beta_1 x_i\}) + \sum_{i=1}^n (\beta_0 + \beta_1 x_i) y_i$$

sicché gli stimatori di massima verosimiglianza di  $\beta_0$  e  $\beta_1$  si otterranno quali soluzioni del seguente sistema di equazioni

$$\begin{cases} \frac{d}{d\beta_0} \ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = \sum_{i=1}^n \left( y_i - \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \right) = 0 \\ \frac{d}{d\beta_1} \ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = \sum_{i=1}^n \left( x_i y_i - \frac{\exp\{\beta_0 + \beta_1 x_i\} x_i}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \right) = 0 \end{cases}$$

ma, non essendo queste lineari nei parametri  $\beta_0$  e  $\beta_1$ , esso non può essere risolto per via analitica: dovremo dunque ricorrere a *metodi numerici* quali Newton-Raphson o lo scoring di Fisher.

A tale scopo calcoliamo gli elementi del vettore gradiente  $\mathbf{u}$

$$u_1 = \frac{d}{d\beta_0} \ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = \sum_{i=1}^n \left( y_i - \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \right)$$

$$u_2 = \frac{d}{d\beta_0} \ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = \sum_{i=1}^n \left( x_i y_i - \frac{\exp\{\beta_0 + \beta_1 x_i\} x_i}{1 + \exp\{\beta_0 + \beta_1 x_i\}} \right)$$

e gli elementi della *matrice hessiana*  $\mathbf{H}$

$$h_{11} = \frac{d^2}{d\beta_0^2} \ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = - \sum_{i=1}^n \frac{\exp\{\beta_0 + \beta_1 x_i\}}{(1 + \exp\{\beta_0 + \beta_1 x_i\})^2}$$

$$h_{12} = \frac{d^2}{d\beta_0 d\beta_1} \ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = - \sum_{i=1}^n \frac{\exp\{\beta_0 + \beta_1 x_i\} x_i}{(1 + \exp\{\beta_0 + \beta_1 x_i\})^2} = h_{21}$$

$$h_{22} = \frac{d^2}{d\beta_1^2} \ell(\beta_0, \beta_1 | \mathbf{y}, \mathbf{x}) = - \sum_{i=1}^n \frac{\exp\{\beta_0 + \beta_1 x_i\} x_i^2}{(1 + \exp\{\beta_0 + \beta_1 x_i\})^2}$$

che, nell'ambito della procedura di stima numerica basata sull'algoritmo di Newton-Raphson, conducono alle seguente equazione ricorsiva di stima

$$\boldsymbol{\beta}^{(h)} = \boldsymbol{\beta}^{(h-1)} - \mathbf{H}_n^{-1}(\boldsymbol{\beta}^{(h-1)}) \mathbf{u}_n(\boldsymbol{\beta}^{(h-1)}) \quad (10.56)$$

dove  $\mathbf{H}_n(\boldsymbol{\beta}^{(h-1)})$  è la matrice delle derivate seconde di  $\ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{x})$  rispetto agli elementi di  $\boldsymbol{\beta}$  (o *matrice hessiana* valutata in  $\boldsymbol{\beta}^{(h-1)}$ ) mentre  $\mathbf{u}_n(\boldsymbol{\beta}^{(h-1)})$  è il vettore delle derivate prime di  $\ell(\boldsymbol{\beta} | \mathbf{y}, \mathbf{x})$  rispetto agli elementi di  $\boldsymbol{\beta}$  (o *vettore gradiente*) anch'esso valutato in  $\boldsymbol{\beta}^{(h-1)}$ .

L'equazione (10.56), giunta a convergenza, restituisce lo stimatore di massima verosimiglianza  $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_0, \hat{\beta}_1)$  di  $\boldsymbol{\beta} = (\beta_0, \beta_1)$  da cui la possibilità di stimare sempre a massima verosimiglianza in virtù della proprietà di invarianza, la probabilità  $\theta_i$  di accadimento dell'evento di interesse per l' $i$ -ma unità statistica.

$$\hat{\theta}_i = \frac{\exp\{\hat{\alpha}_n + \hat{\beta}_n x_i\}}{1 + \exp\{\hat{\alpha}_n + \hat{\beta}_n x_i\}} \quad (10.57)$$

In relazione alle proprietà (asintotiche) egli stimatori di massima verosimiglianza, possiamo anche osservare che

$$\hat{\boldsymbol{\beta}}_n \underset{a}{\sim} N_2(\boldsymbol{\beta}, \mathbf{I}_n^{-1}(\boldsymbol{\beta})) \quad (10.58)$$

dove  $\mathbf{I}_n^{-1}(\boldsymbol{\beta})$  è l'inversa della *matrice di informazione* di Fisher

$$\mathbf{I}_n(\boldsymbol{\beta}) = -\mathbb{E}(\mathbf{H}_n(\boldsymbol{\beta})) \quad (10.59)$$

ovvero, la *matrice di covarianza asintotica* dello stimatore di massima verosimiglianza  $\hat{\boldsymbol{\beta}}_n$  di  $\boldsymbol{\beta}$ .

A questo punto abbiamo tutti gli elementi necessari alla costruzione di procedure inferenziali (asintotiche) per i parametri del modello di regressione logistica e, in ultima analisi, per il modello stesso.

**Esempio 10.2.3.** In uno studio clinico si è studiata la relazione che corre tra presenza di *attacchi di panico* e *livello di ansia* ipotizzando il seguente modello di regressione logistica a descrivere la probabilità  $\theta_i$  di incorrere in un attacco di panico ( $Y = 1$ ) in funzione del livello di ansia ( $x$ ) misurato su di una scala (*scala di Hamilton*) con un intervallo di punteggio totale di 0 – 56, dove  $x < 17$  indica lieve entità,  $18 \leq x \leq 24$  da lieve a moderata e  $25 \leq x \leq 30$  da moderata a grave:

$$\theta_i = P_{\beta_0, \beta_1}(Y_i = 1|x = x_i) = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}, \quad i = 1, 2, \dots, n \quad (10.60)$$

con  $n$  numero di unità statistiche coinvolte nello studio clinico.

Sulla base di un campione di 14 osservazioni relative alla coppia di variabili appaiate  $(x, Y)$ , vale a dire

$$(23, 0), (23, 0), (23, 0), (23, 0), (23, 0), (23, 1), (27, 0) \\ (27, 1), (28, 0), (28, 1), (28, 1), (28, 1), (28, 1), (28, 1)$$

si è ottenuta la stima di massima verosimiglianza, per via numerica usando l'algoritmo di Newton-Raphson, dei parametri del modello; riportiamo qui sotto il protocollo dell'esecuzione dell'algoritmo stesso fino a convergenza (sono richieste sono 6 iterazioni per raggiungere la precisione desiderata  $\varepsilon = 10^{-10}$ , a riprova della veloce convergenza (quadratica) dell'algoritmo di Newton-Raphson)

Iterazione	$\alpha$	$\beta$	$\ln \ell(\boldsymbol{\beta} \mathbf{y}, \mathbf{x})$
1	-13.043	0.507	-7.0142
2	-15.325	0.594	-6.9594
3	-15.518	0.601	-6.599
4	-15.520	0.601	-6.959
5	-15.520	0.601	-6.959
6	-15.520	0.601	-6.959

per cui, le stime di massima verosimiglianza dei parametri sono

$$\hat{\beta}_0 = -15.520 \quad \hat{\beta}_1 = 0.601 \quad (10.61)$$

mentre quella della matrice di covarianza asintotica di  $\hat{\boldsymbol{\beta}}$  data da  $\mathbf{I}_n(\hat{\boldsymbol{\beta}}_n)$  è

$$\mathbf{I}_n(\hat{\boldsymbol{\beta}}_n) = \begin{pmatrix} 59.234 & -2.258 \\ -2.258 & 0.0867 \end{pmatrix} \quad (10.62)$$

Possiamo allora riassumere i risultati rilevanti ai fini dell'inferenza sul modello di regressione logistica ipotizzato nella tabella che segue, ricordando la Normalità asintotica che caratterizza la distribuzione asintotica dello stimatore di massima verosimiglianza:

Parametro	MLE	$\beta$	z	p-value
$\beta_0$	-15.520	7.696	-2.016	0.0437
$\beta_1$	0.601	0.294	-2.041	0.0412

e concludere sulla base dei valori del *p*-value per la significatività di entrambi i parametri  $\beta_0$  e  $\beta_1$ , e quindi del modello di regressione logistica ipotizzato (almeno al 5%).

Possiamo ora analizzare brevemente i risultati ottenuti: la stima del parametro  $\beta_1$  è  $\hat{\beta}_1 = 0.601$  e quindi possiamo concludere che la relazione tra il livello di ansia e gli attacchi di panico è di tipo positivo: per un incremento unitario del punteggio di ansia, il valore del logOdds della variabile attacchi di panico

$$\text{logit}(\hat{\theta}_i) = \ln \left( \frac{P_{\hat{\beta}_0, \hat{\beta}_1}(Y_i = 1|x = x_i)}{P_{\hat{\beta}_0, \hat{\beta}_1}(Y_i = 0|x = x_i)} \right) = \hat{\beta}_0 + \hat{\beta}_1 x_i = -15.52 + 0.60 x_i \quad (10.63)$$

aumenta di 0.60 unità. In poche parole, all'aumentare del livello di ansia aumenta la probabilità di avere un attacco di panico.

E' dunque possibile conoscere la probabilità di osservare un attacco di panico in corrispondenza di ogni specifico livello di ansia:

$$\hat{\theta}_i = P_{\hat{\beta}_0, \hat{\beta}_1}(Y_i = 1|x = x_i) = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}} = \frac{\exp\{-15.52 + 0.601 x_i\}}{\exp\{-15.52 + 0.601 x_i\}}$$

per  $i = 1, 2, \dots, n$ ; per esempio, per  $x = 30$ , si ha la probabilità di osservare un attacco di panico è pari a 0.9227.

Ora, se stabiliamo che chi ha una probabilità stimata di avere un attacco di panico *superiore* a un *valore soglia*, per esempio 0.5, è da considerare individuo suscettibile ad attacco di panico, mediante le probabilità stimate  $\hat{\theta}_i$  possiamo *classificare* le unità osservate in *individui suscettibili* e *individui non suscettibili* ad attacchi di panico.

Possiamo anche calcolare una *misura di adeguatezza* del modello ai dati (o di *goodness-of-fit*) confrontando la tabella di classificazione ottenuta mediante il modello di regressione logistica stimato con quella costruita sulla base dei valori realmente osservati sulla variabile  $Y$  tramite una tabella a due vie nota come *tabella di classificazione*

	$Y = 0$	$Y = 1$	
$\hat{Y} = 0$	$n_{00}$	$n_{01}$	$n_{0\bullet}$
$\hat{Y} = 1$	$n_{10}$	$n_{11}$	$n_{1\bullet}$
	$n_{\bullet 0}$	$n_{\bullet 1}$	$n$

sulla base della quale calcolare la seguente *misura di adeguatezza* del modello proposto ai dati

$$A = \frac{n_{00} + n_{11}}{n} \quad (10.64)$$

dove il numeratore  $n_{00} + n_{11}$  rappresenta il *numero* di individui *correttamente classificati*; per differenza  $\bar{A} = 1 - A$  restituisce un *indice di misclassificazione*. Nell'esempio che stiamo trattando  $A = \frac{5+6}{14} = 0.786 \simeq 79\%$ .

Da ultimo osserviamo che il modello stimato può anche essere usato anche a *fini estrapolativi* vale a dire per assegnare probabilità all'evento che un nuovo individuo, diverso da quelli su cui si è basata la stima del modello, possa presentare attacchi di panico.

# 11 Likelihood Ratio Tests

Rivisitiamo l'argomento dei test per la verifica di ipotesi alla luce dei concetti di verosimiglianza e sufficienza introdotti nei capitoli precedenti. Questo darà modo di costruire una teoria organica e coerente con cui affrontare questa ampia classe di problemi inferenziali.

## 11.1 Test basati sul rapporto di verosimiglianza

Consideriamo un campione casuale  $(X_1, X_2, \dots, X_n)$  proveniente da una distribuzione con funzione di distribuzione cumulata  $F_X(x; \theta)$ ,  $\theta \in \Theta \subset \mathbb{R}^k$  che ammette funzione di massa/densità  $f_X(x; \theta)$ ,  $\theta \in \Theta$ .

Se siamo interessati a verificare il seguente sistema di ipotesi

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 < \theta_0 \end{cases}$$

possiamo, in virtù di quanto visto nei capitoli precedenti, sfruttare a tale scopo l'informazione contenuta nel seguente rapporto (di verosimiglianza)

$$\frac{L(\theta_1 | \mathbf{x})}{L(\theta_0 | \mathbf{x})} \tag{11.1}$$

ricordando che la funzione di verosimiglianza induce una *misura di preferenza* (relativa) associata ai diversi valori del parametro  $\theta$  alla luce dell'informazione riguardo  $\theta$  contenuta nella determinazione campionaria  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ . In conseguenza di ciò, è ragionevole rifiutare  $H_0$  a favore di  $H_1$  per *grandi valori* del rapporto (11.1), ovvero se

$$\frac{L(\theta_1 | \mathbf{x})}{L(\theta_0 | \mathbf{x})} > A \tag{11.2}$$

per  $A > 0$  essendo, in questo caso,  $\theta_0$  *meno verosimile* di  $\theta_1$  alla luce di quanto osservato.

Possiamo dunque rivisitare quanto visto nel paragrafo 5.4 alla luce delle precedenti considerazioni.

**Esempio 11.1.1.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Normale  $N(\mu, \sigma^2)$  con  $\sigma^2$  nota e supponiamo di essere interessati a verificare il seguente sistema di ipotesi (al momento semplici):

$$\begin{cases} H_0 : \mu = \mu_0 \\ vs. \\ H_1 : \mu = \mu_1 \quad (\mu_1 > \mu_0) \end{cases} \tag{11.3}$$

La funzione di verosimiglianza associata alla distribuzione Normale per il caso che stiamo trattando è

$$L(\mu; \mathbf{x}) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \quad (11.4)$$

sicché

$$\begin{aligned} \frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} &= \frac{\exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_1)^2 \right\}}{\exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right\}} \\ &= \frac{\exp \left\{ -\frac{1}{2\sigma^2} [\sum_{i=1}^n x_i^2 - 2\mu_1 \sum_{i=1}^n x_i + n\mu_1^2] \right\}}{\exp \left\{ -\frac{1}{2\sigma^2} [\sum_{i=1}^n x_i^2 - 2\mu_0 \sum_{i=1}^n x_i + n\mu_0^2] \right\}} \\ &= \dots \\ &= \exp \left\{ \frac{n}{\sigma^2} \bar{x}_n (\mu_1 - \mu_0) + \frac{n}{2\sigma^2} (\mu_0^2 - \mu_1^2) \right\} > A \end{aligned} \quad (11.5)$$

ossia

$$\frac{n}{\sigma^2} \bar{x}_n (\mu_1 - \mu_0) + \frac{2}{2\sigma^2} (\mu_0^2 - \mu_1^2) > \ln(A) \quad (11.6)$$

quindi

$$\bar{x}_n > \frac{\frac{n}{2\sigma^2} (\mu_0^2 - \mu_1^2) - \ln A}{\frac{n}{\sigma^2} (\mu_0 - \mu_1)} = B \quad (11.7)$$

e dunque la regione critica di livello  $\alpha$  sarà

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{L(\theta_1 | \mathbf{x})}{L(\theta_0 | \mathbf{x})} > A \right\} = \{ \mathbf{x} \in \mathfrak{X} : \bar{x}_n > B \}. \quad (11.8)$$

Ora, fissato  $\alpha$ , la soluzione dell'equazione che segue restituisce proprio il valore di  $B$ , necessario a rendere "operativo" il test:

$$\alpha = P_\mu(\bar{X}_n > B | \mu = \mu_0) = \int_B^{+\infty} \frac{1}{\sqrt{(2\pi)\sigma^2/n}} \exp^{-\frac{n}{2\sigma^2}(\bar{x}_n - \mu_0)^2} d\bar{x}_n. \quad (11.9)$$

Infatti, posto  $z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{(n)}} = z_{1-\alpha}$ , si ha

$$\frac{B - \mu_0}{\sigma/\sqrt{(n)}} = z_{1-\alpha} \quad (11.10)$$

da cui

$$B = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad (11.11)$$

Ma allora la regola di decisione per il sistema di ipoesti in questione si riduce a

a) **rifiutare**  $H_0 : \mu = \mu_0$  se  $\bar{x}_n > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$

ovvero in alternativa,

a') **rifiutare**  $H_0 : \mu = \mu_0$  se  $z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{(n)}} > z_{1-\alpha}$ .

Considerato che

$$C_\alpha = \left\{ \bar{x}_n \in \mathfrak{X} : \bar{x}_n > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\} \quad (11.12)$$

la potenza del test in questione sarà

$$\begin{aligned} \eta_{C_\alpha} &= 1 - \beta = 1 - P_\mu(\bar{X}_n \leq \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} | \mu = \mu_1) \\ &= 1 - P_\mu\left(\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq \frac{\mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}}\right) \\ &= 1 - P_\mu\left(Z \leq \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \\ &= 1 - \Phi_Z\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \end{aligned} \quad (11.13)$$

E cosa cambierebbe se fosse  $H_1 : \mu = \mu_1, (\mu_0 > \mu_1)$ ?

E' bene ricordare ancora una volta che la regione critica  $C_\alpha$  identifica il test e la corrispondente (funzione di) potenza. In altre parole, poggiare la costruzione della regola di decisione sul rapporto di verosimiglianza (11.1) evita di dover porci il problema dell'individuazione di una statistica pivot a partire dalla quale costruire la regola di decisione.

**Esempio 11.1.2.** Sia  $X_1, X_2, \dots, X_n$  un campione casuale da una distribuzione esponenziale di parametro  $\beta > 0$  e si voglia sottoporre a verifica il seguente sistema di ipotesi

$$\begin{cases} H_0 : \beta = \beta_0 \\ vs. \\ H_1 : \beta = \beta_1, \quad (\beta_1 > \beta_0) \end{cases} \quad (11.14)$$

Ora,

$$L(\beta|\mathbf{x}) = \prod_{i=1}^n \frac{1}{\beta} e^{-\frac{1}{\beta}x_i} \mathbb{1}_{\mathbb{R}^+}(x_i) = \frac{1}{\beta^n} e^{-\frac{1}{\beta} \sum_{i=1}^n x_i} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i) \quad (11.15)$$

per cui

$$\frac{L(\beta_1|\mathbf{x})}{L(\beta_0|\mathbf{x})} = \frac{\frac{1}{\beta_1^n} e^{-\frac{1}{\beta_1} \sum_{i=1}^n x_i} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i)}{\frac{1}{\beta_0^n} e^{-\frac{1}{\beta_0} \sum_{i=1}^n x_i} \prod_{i=1}^n \mathbb{1}_{\mathbb{R}^+}(x_i)} > A \quad (11.16)$$

sicché, con qualche semplice passaggio, si arriva a

$$\sum_{i=1}^n x_i > \frac{\ln(A) + \ln(\beta_1) - \ln(\beta_0)}{\frac{1}{\beta_0} - \frac{1}{\beta_1}} = B \quad (11.17)$$

con  $\sum_{i=1}^n X_i \sim \mathcal{G}(n, \beta)$ , e saremo portati a rifiutare  $H_0$  per valori di  $W = \sum_{i=1}^n X_i$  più grandi di  $B$ . Ora, la probabilità di commettere un errore di prima specie è data da

$$\alpha = P_\beta\left(\sum_{i=1}^n X_i > B | \beta = \beta_0\right) = \int_B^\infty \frac{1}{\Gamma(n) \beta_0^n} w^{n-1} e^{-\frac{1}{\beta_0} w} dw \quad (11.18)$$

sicché fissato il valore di  $\alpha$ , troveremo  $B = w_{n,\beta_0;1-\alpha}$ , quantile di ordine  $1 - \alpha$  della distribuzione di  $W$  sotto  $H_0$  ossia  $\mathcal{G}(n, \beta_0)$ .

Di conseguenza, la regione critica per il test in questione sarà

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \sum_{i=1}^n x_i > w_{n,\beta_0;1-\alpha} \right\} \quad (11.19)$$

da cui segue la *regola di decisione*

"... rifiutare  $H_0$  se  $\sum_{i=1}^n x_i > w_{n,\beta_0;1-\alpha}$ ".

Per fare qualche conto, supponiamo  $n = 45$  e  $\beta_0 = 1.7$ . Per  $\alpha = 0.05$  si ottiene  $B = w_{45,1.7;0.95} = 96.17$  (usando R per calcolare il quantile  $w$  di interesse). La potenza del test in questione sarà, per un fissato valore di  $\beta = \beta_1$  (sotto ipotesi alternativa), data da

$$\begin{aligned} \eta_{C_\alpha} &= \int_{96.17}^{\infty} \frac{1}{\Gamma(n) \beta_1^n} w^{n-1} e^{-\frac{1}{\beta_1} w} dw \\ &= 1 - \int_0^{96.17} \frac{1}{\Gamma(n) \beta_1^n} w^{n-1} e^{-\frac{1}{\beta_1} w} dw \\ &= 1 - P_\beta \left( \sum_{i=1}^n X_i \leq w_{n,\beta_0;1-\alpha} \mid \beta = \beta_1 \right) \\ &= 1 - P_\beta (W \leq w_{45,1.7;0.95} \mid \beta = \beta_1) \\ &= 1 - P_\beta (W \leq 96.17 \mid \beta = \beta_1) \\ &= 1 - F_W (96.17; \beta_1) \end{aligned} \quad (11.20)$$

Ora, posto  $\beta_1 = 2.3$  avremo  $\eta_{C_\alpha} = 1 - 0.33 = 0.67$  mentre per  $\beta_1 = 2.7$  avremo  $\beta_{C_\alpha} = 1 - 0.07 = 0.93$  nonché posto  $\beta_1 = 1.7 = \beta_0$  avremo  $\eta_{C_\alpha} = 1 - 0.95 = 0.05$ .

Rimane una domanda a cui dovremmo rispondere: i test che abbiamo individuato nei due esempi (e in generale, i test basati sul rapporto di verosimiglianza) sono i *migliori* a cui possiamo pensare tra tutti competitori di medesimo livello  $\alpha$  per il sistema di ipotesi considerato. In analogia con quanto visto in ambito di stima, si tratta di introdurre un *criterio di ottimalità* per la scelta del test (e come vedremo, questo criterio potrà essere individuato a partire dal rapporto di verosimiglianza).

Stabiliamo dunque che cosa intendiamo per test *migliore*.

**Definizione 11.1.1** (Test più potente di livello  $\alpha$ ). Fissato  $\alpha$ , un test che minimizza la probabilità di commettere un errore di secondo tipo (ovvero che massimizza la potenza  $\eta_{C_\alpha}$ ) è detto test *più potente* di livello  $\alpha$ .

Il teorema che segue formalizza il procedimento con cui individuare il test ottimo.

**Teorema 11.1.1** (Lemma di Neyman-Pearson). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione avente funzione di densità (o di massa)  $f_X(x; \theta)$  e sia

$$\begin{cases} H_0 : \theta = \theta_0 \\ vs. \\ H_1 : \theta = \theta_1, \quad \theta_0 < \theta_1 \text{ (oppure } \theta_0 > \theta_1) \end{cases} \quad (11.21)$$

il sistema di ipotesi semplici che si intende sottoporre a verifica alla luce dell'informazione contenuta nella determinazione del campione casuale. Se  $L(\theta|\mathbf{x})$  è la funzione di verosimiglianza per il problema allora il test più potente di livello  $\alpha$  ha regione critica

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq A \right\} \quad (11.22)$$

con  $A$  costante non negativa, determinata dal livello  $\alpha$  del test.

*Dimostrazione.* Svolgiamo la dimostrazione nel caso di una distribuzione continua; nel caso discreto il lemma si dimostra in modo analogo, sostituendo opportunamente integrali con somme.

Sia  $C_\alpha$  la regione critica di livello  $\alpha$  restituita dal lemma di Neyman-Pearson e sia  $C_\alpha^*$  la regione critica di un qualsiasi altro test di livello  $\alpha$  per il medesimo sistema di ipotesi. Allora per dimostrare il lemma bisognerà dimostrare che

$$\eta_{C_\alpha^*} \leq \eta_{C_\alpha} \quad (11.23)$$

ovvero indicata con  $\beta$  la probabilità di commettere un errore di secondo tipo ed essendo  $\eta_{C_\alpha^*} = 1 - \beta^*$  e  $\eta_{C_\alpha} = 1 - \beta$ , si tratterà dunque di dimostrare che

$$\beta^* \geq \beta \equiv \beta^* - \beta \geq 0 \quad (11.24)$$

Ora

$$\begin{aligned} \beta^* - \beta &= P(\mathbf{X} \in \overline{C}_\alpha^* | \theta = \theta_1) - P(\mathbf{X} \in \overline{C}_\alpha | \theta = \theta_1) \\ &= \int_{\mathbf{x} \in \overline{C}_\alpha^*} L(\theta_1; \mathbf{x}) d\mathbf{x} - \int_{\mathbf{x} \in \overline{C}_\alpha} L(\theta_1; \mathbf{x}) d\mathbf{x} \\ &= \int_{\overline{C}_\alpha^* \cap C_\alpha} L(\theta_1; \mathbf{x}) d\mathbf{x} + \int_{\overline{C}_\alpha^* \cap \overline{C}_\alpha} L(\theta_1; \mathbf{x}) d\mathbf{x} - \int_{\overline{C}_\alpha \cap C_\alpha^*} L(\theta_1; \mathbf{x}) d\mathbf{x} - \int_{\overline{C}_\alpha \cap \overline{C}_\alpha^*} L(\theta_1; \mathbf{x}) d\mathbf{x} \end{aligned} \quad (11.25)$$

dove per scrivere l'ultimo passaggio abbiamo osservato che

$$\begin{aligned} \overline{C}_\alpha^* &= \overline{C}_\alpha^* \cap (C_\alpha \cup \overline{C}_\alpha) = (\overline{C}_\alpha^* \cap C_\alpha) \cup (\overline{C}_\alpha^* \cap \overline{C}_\alpha) \\ \overline{C}_\alpha &= \overline{C}_\alpha \cap (C_\alpha^* \cup \overline{C}_\alpha^*) = (\overline{C}_\alpha \cap C_\alpha^*) \cup (\overline{C}_\alpha \cap \overline{C}_\alpha^*) \end{aligned} \quad (11.26)$$

Notiamo che il secondo e il quarto termine della somma (11.25) si elidono reciprocamente. Inoltre, in  $C_\alpha$  e dunque anche in  $C_\alpha^* \cap \overline{C}_\alpha$ , si ha che  $L(\theta_1; \mathbf{x}) \geq A L(\theta_1; \mathbf{x})$ , mentre in  $\overline{C}_\alpha$ , e dunque anche in  $\overline{C}_\alpha \cap C_\alpha^*$  si ha che  $L(\theta_1; \mathbf{x}) < A L(\theta_0; \mathbf{x})$ . Perciò

$$\beta^* - \beta \geq A \left[ \int_{C_\alpha \cap \overline{C}_\alpha^*} L(\theta_0; \mathbf{x}) d\mathbf{x} - \int_{\overline{C}_\alpha \cap C_\alpha^*} L(\theta_0; \mathbf{x}) d\mathbf{x} \right] \quad (11.27)$$

Sommando e sottraendo  $A \int_{C_\alpha \cap C_\alpha^*} L(\theta_0; \mathbf{x}) d\mathbf{x}$  al membro di destra della disegualanza si ottiene

$$\begin{aligned}\beta^* - \beta &\geq A \left[ \int_{(C_\alpha \cap \bar{C}_\alpha^*) \cup (C_\alpha \cap C_\alpha^*)} L(\theta_0; \mathbf{x}) d\mathbf{x} - \int_{(\bar{C}_\alpha \cap C_\alpha^*) \cup (C_\alpha \cap C_\alpha^*)} L(\theta_0; \mathbf{x}) d\mathbf{x} \right] \\ &\geq A \left[ \int_{C_\alpha} L(\theta_0; \mathbf{x}) d\mathbf{x} - \int_{C_\alpha^*} L(\theta_0; \mathbf{x}) d\mathbf{x} \right] \\ &= A(\alpha - \alpha) = 0\end{aligned}\tag{11.28}$$

da cui

$$\beta^* - \beta \geq 0 \equiv \eta_{C_\alpha^*} \leq \eta_{C_\alpha} \tag{11.29}$$

e quindi il lemma è dunque dimostrato.  $\square$

Possiamo subito osservare che, se esiste una statistica sufficiente  $T_n$  per il parametro  $\theta$ , per il teorema di fattirizzazione

$$L(\theta|\mathbf{x}) = g(\theta, t_n) h(x) \tag{11.30}$$

sicché la regione critica del test più potente (=ottimale) di livello  $\alpha$  per il sistema di ipotesi di interesse costituita da

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq A \right\} \tag{11.31}$$

sarà sempre esprimibile per mezzo dell' statistica test  $\phi(T_n)$  che sarà funzione di  $T_n$ , statistica sufficiente per  $\theta$ .

Tre considerazioni vanno fatte in merito al lemma di Neyman-Pearson:

- a) il teorema resta valido qualunque sia il numero di parametri (purché finito) che caratterizzano la funzione di distribuzione da cui proviene il campione
- b) il teorema non richiede esplicitamente l'indipendenza stocastica delle  $n$  osservazioni costituenti il campione (in questo caso, la funzione di verosimiglianza non coincide con il prodotto delle funzioni di massa/ densità marginali ma è pur sempre ben definita)
- c) nel teorema sono fissate le condizioni necessarie affinché il test sia il più potente tra quelli di livello  $\alpha$  fissato, ma vengono anche indicate le regole l'individuazione della regione critica (e conseguentemente, della regola di decisione).

**Esempio 11.1.3.** Consideriamo due variabili casuali indipendenti  $U$  e  $V$ , entrambe distribuite come  $N(0, \sigma^2)$ , e definiamo una nuova variabile casuale  $X = \sqrt{U^2 + V^2}$  che risulta avere una distribuzione di Rayleigh di parametro  $\sigma^2$  la cui funzione di densità è così definita

$$f_X(x; \sigma^2) = \frac{x}{\sigma^2} \exp \left\{ -\frac{x^2}{2\sigma^2} \right\} \mathbb{1}_{\mathbb{R}^+}(x), \quad \sigma^2 > 0 \tag{11.32}$$

- a) Trovare lo stimatore di massima verosimiglianza di  $\sigma^2$  e la sua distribuzione asintotica.

b) Considerato il seguente sistema di ipotesi,

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ vs. \\ H_1 : \sigma^2 = \sigma_1^2, \quad \sigma_1^2 > \sigma_0^2 \end{cases} \quad (11.33)$$

determinare la regione critica del test più potente di prefissato livello  $\alpha$ .

La funzione di log-verosimiglianza è data da

$$\ell(\sigma^2 | \mathbf{x}) = \sum_{i=1}^n \ln(x_i) - n \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \quad (11.34)$$

e

$$\hat{\sigma}_n^2 = \arg \max_{\sigma^2 > 0} \ell(\sigma^2 | \mathbf{x}) = \arg \max_{\sigma^2 > 0} \left\{ \sum_{i=1}^n \ln(x_i) - n \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right\}. \quad (11.35)$$

Sicché, lo stimatore di massima verosimiglianza sarà dato dalla soluzione dell'equazione

$$\frac{d}{d\sigma^2} \ell(\sigma^2 | \mathbf{x}) = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n x_i^2 = 0 \quad (11.36)$$

e dunque,

$$\hat{\sigma}_n^2 = \frac{1}{2n} \sum_{i=1}^n x_i^2 \quad (11.37)$$

è lo stimatore di massima verosimiglianza di  $\sigma^2$  tenuto conto del fatto che

$$\frac{d^2}{d\sigma^4} \ell(\sigma^2 | \mathbf{x}) \Big|_{\sigma^2=\hat{\sigma}_n^2} = -\frac{n}{\hat{\sigma}_n^2} < 0. \quad (11.38)$$

L'informazione di Fisher associata all'inter campione è

$$\begin{aligned} I_n(\sigma^2) &= -\mathbb{E} \left[ \frac{d^2}{d\sigma^4} \ell(\sigma^2 | \mathbf{x}) \right] \\ &= -\mathbb{E} \left[ \frac{n}{\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n X_i^2 \right] \\ &= -\frac{n}{\sigma^4} + \frac{1}{\sigma^4} \sum_{i=1}^n \mathbb{E}_{\sigma^2} \left[ \frac{X_i^2}{\sigma^2} \right]. \end{aligned} \quad (11.39)$$

Poiché

$$\left( \frac{X}{\sigma} \right)^2 = \left( \frac{U}{\sigma} \right)^2 + \left( \frac{V}{\sigma} \right)^2 \quad (11.40)$$

con  $\left( \frac{U}{\sigma} \right) \sim N(0, 1)$  e  $\left( \frac{V}{\sigma} \right) \sim N(0, 1)$ , indipendenti, si ha

$$\frac{X^2}{\sigma^2} = \sum_{i=1}^n Z_i^2 \sim \chi_2^2 \quad (11.41)$$

con

$$\mathbb{E}_{\sigma^2} \left( \frac{X^2}{\sigma^2} \right) = 2. \quad (11.42)$$

Ma allora,

$$I_n(\sigma^2) = -\frac{n}{\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^n \mathbb{E}_{\sigma^2} \left( \frac{X_i^2}{\sigma^2} \right) = \frac{n}{\sigma^4}. \quad (11.43)$$

Valendo per la famiglia distribuzioni da cui proviene il campione le usuali condizioni di regolarità, per i teoremi relativi alle proprietà (asintotiche) degli stimatori di massima verosimiglianza si ha

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma_*^2) \xrightarrow{D} N(0, \sigma^4) \quad (11.44)$$

con  $\sigma_*^2$  vero valore del parametro e dunque,

$$\hat{\sigma}_n^2 \xrightarrow{D} N\left(\sigma_*^2, \frac{\sigma^4}{n}\right) \quad (11.45)$$

In virtù del lemma di Neyman-Pearson, la regione di rifiuto per il sistema di ipotesi al punto b), fissato il livello di significatività  $\alpha$  del test, è data da

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{L(\sigma_1^2; \mathbf{x})}{L(\sigma_0^2; \mathbf{x})} \geq A \right\} \quad (11.46)$$

con  $A$  costante che è determinata dal livello  $\alpha$  del test. Sicché,

$$\ln L(\sigma_1^2 | \mathbf{x}) - \ln L(\sigma_0^2 | \mathbf{x}) \geq \ln(A) \quad (11.47)$$

ossia

$$\sum_{i=1}^n \ln(x_i) - n \ln(\sigma_1^2) - \frac{1}{2\sigma_1^2} \sum_{i=1}^n x_i^2 - \sum_{i=1}^n \ln(x_i) + n \ln(\sigma_0^2) + \frac{1}{2\sigma_0^2} \sum_{i=1}^n x_i^2 \geq A \quad (11.48)$$

per cui

$$-n(\ln(\sigma_1^2) - \ln(\sigma_0^2)) - \left[ \frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2} \right] \sum_{i=1}^n x_i^2 \geq \ln(A) \quad (11.49)$$

e dunque

$$\left( \frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2} \right) \sum_{i=1}^n x_i^2 \geq \ln(A) - n[\ln(\sigma_1^2) - \ln(\sigma_0^2)] \quad (11.50)$$

o, in maniera equivalente,

$$\frac{1}{2} \left[ \frac{\sigma_0^2 - \sigma_1^2}{\sigma_0^2 \sigma_1^2} \right] \sum_{i=1}^n x_i^2 < B \quad (11.51)$$

con  $B = \ln(A) - n[\ln(\sigma_1^2) - \ln(\sigma_0^2)]$ . Poiché  $\sigma_1^2 > \sigma_0^2$  avremo infine

$$\frac{1}{2} \sum_{i=1}^n x_i^2 \geq \frac{\sigma_0^2 \sigma_1^2}{\sigma_0^2 - \sigma_1^2} \cdot B \equiv \frac{1}{2n} \sum_{i=1}^n x_i^2 \geq \frac{\sigma_0^2 \sigma_1^2}{n(\sigma_0^2 - \sigma_1^2)} \cdot B \quad (11.52)$$

dove la costante  $B^* = \frac{\sigma_0^2 \sigma_1^2}{n(\sigma_0^2 - \sigma_1^2)} \cdot B$  è determinata dal livello di significatività  $\alpha$  del test.

In conseguenza di ciò, la regione critica del test più potente di livello  $\alpha$  è data da

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \hat{\sigma}_n^2 = \frac{1}{2n} \sum_{i=1}^n x_i^2 \geq B^* \right\} \quad (11.53)$$

con  $\hat{\sigma}_n^2$  stimatore di massima verosimiglianza di  $\sigma^2$ .

La costante  $\mathbb{B}_{\sigma^2*}$ , una volta che si è fissato il livello di significatività  $\alpha$  del test si ottiene dalla soluzione dell'equazione in  $B^*$

$$\alpha = P_{\sigma^2} (\hat{\sigma}_n^2 \geq B^* | H_0) \quad (11.54)$$

e sotto  $H_0$ , per  $n$  grande,

$$\hat{\sigma}_n^2 \underset{a}{\sim} N \left( \sigma_0^2, \frac{\sigma_0^4}{n} \right). \quad (11.55)$$

Allora, standardizzando,

$$1 - \alpha = P_{\sigma^2} \left( \frac{\hat{\sigma}_n^2 - \sigma_0^2}{\sigma_0^2 / \sqrt{n}} < \frac{B^* - \sigma_0^2}{\sigma_0^2 / \sqrt{n}} \right) \quad (11.56)$$

da cui

$$\frac{B^* - \sigma_0^2}{\sigma_0^2 / \sqrt{n}} = \Phi_Z^{-1}(1 - \alpha) = z_{1-\alpha} \quad (11.57)$$

e perciò

$$B^* = \sigma_0^2 + \frac{\sigma_0^2}{\sqrt{n}} z_{1-\alpha}. \quad (11.58)$$

Infine, possiamo riscrivere le regione di rifiuto di  $H_0$  di livello  $\alpha$  come

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \hat{\sigma}_n^2 > \sigma_0^2 + \frac{\sigma_0^2}{\sqrt{n}} z_{1-\alpha} \right\} \quad (11.59)$$

Per esempio, sia  $H_0 : \sigma^2 = 2$  vs  $H_1 : \sigma^2 = 3$  e sia  $n = 100$  e  $\sum_{i=1}^{100} x_i = 470$ . Allora, fissato  $\alpha = 0.1$  avremo

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \hat{\sigma}_n^2 > 2 + \frac{2}{\sqrt{100}} \Phi_Z^{-1}(0.9) \right\} = \left\{ \mathbf{x} \in \mathfrak{X} : \hat{\sigma}_n^2 > 2.256 \right\} \quad (11.60)$$

ed essendo

$$\hat{\sigma}_n^2 = \frac{1}{2 \cdot 100} \cdot 470 = 2.35 \quad (11.61)$$

a livello  $\alpha = 0.1$ , rifiuteremo  $H_0 : \sigma_0^2 = 2$  a favore di  $\sigma_1^2 = 3$ .

La potenza (asintotica) per il test in questione è data da

$$\begin{aligned} \eta_{C_\alpha} &= 1 - \beta = 1 - P_{\sigma^2} (\hat{\sigma}_n^2 < B^* | \sigma^2 = \sigma_1^2) \\ &= 1 - P_{\sigma^2} \left( \frac{\hat{\sigma}_n^2 - \sigma_1^2}{\sigma_1^2 / \sqrt{n}} < \frac{B^* - \sigma_1^2}{\sigma_1^2 / \sqrt{n}} \right) \\ &= 1 - P_{\sigma^2} \left( Z < \frac{\sigma_0^2 + \frac{\sigma_0^2}{\sqrt{n}} z_{1-\alpha} - \sigma_1^2}{\sigma_1^2 / \sqrt{n}} \right) \\ &= 1 - P_{\sigma^2} \left( Z < \frac{\sqrt{n}(\sigma_0^2 - \sigma_1^2) + \frac{\sigma_0^2}{\sqrt{n}} z_{1-\alpha}}{\sigma_1^2} \right) \end{aligned} \quad (11.62)$$

con  $P_{\sigma^2}(Z < \dots)$  sempre più trascurabile via via che  $\sigma_1^2$  si allontana da  $\sigma_0^2$  diventando sempre più grande (si ricordi che si è assunto  $\sigma_1^2 > \sigma_0^2$ ) e di conseguenza  $\eta_{C_\alpha} \rightarrow 1$  al crescere di  $n$ .

Generalmente non abbiamo quasi mai informazione sufficiente per fissare a un preciso valore  $\theta_1$  il valore del parametro  $\theta$  sotto ipotesi alternativa  $H_1$ .

E' perciò necessario cercare una procedura che permetta di sottoporre a verifica ipotesi composite individuando una regola di decisione che sia *più potente* quale sia il valore di  $\theta$  sotto  $H_1$  vale a dire, *uniformemente più potente* (*Uniformly Most Powerful*).

**Definizione 11.1.2** (Test UMP). Un test UMP è una regola di decisione la cui funzione di potenza,  $\eta_{C_\alpha}(\theta_1)$ ,  $\forall \theta_1 \in \Theta_1$  per  $\alpha$  fissato, domina le funzioni di potenza di un qualsiasi altro test competitore di medesimo livello  $\alpha$ .

In maniera alternativa, possiamo anche dire

**Definizione 11.1.3** (Test UMP). La regione critica  $C_\alpha$  è una regione *uniformemente più potente* di livello  $\alpha$  per la verifica di  $H_0$  ipotesi semplice contro  $H_1$  composita unilaterale se  $C_\alpha$  corrisponde alla regione critica ottimale di livello  $\alpha$  restituita dal lemma di Neyman-Pearson.

In buona sostanza, la precedente definizione sottolinea che la *forma* della regione critica di un test UMP rimane *invariata* per ogni valore di  $\theta$  sotto  $H_1$ .

La definizione che segue esplicita la condizione sotto la quale esiste un test UMP.

**Definizione 11.1.4** (Rapporto di verosimiglianza monotono). La famiglia di distribuzioni  $\mathcal{F}_\theta = \{F(x; \theta), \theta \in \Theta\}$  gode della proprietà di *monotonia del rapporto di verosimiglianza* se il rapporto di verosimiglianza

$$\frac{L(\theta_1 | \mathbf{x})}{L(\theta_0 | \mathbf{x})} = \frac{L(\theta_1 | T_n(\mathbf{x}))}{L(\theta_0 | T_n(\mathbf{x}))} = \frac{L(\theta_1 | t_n)}{L(\theta_0 | t_n)} \quad (11.63)$$

è funzione *non decrescente* della statistica sufficiente  $T_n(X_1, X_2, \dots, X_n)$  per  $\theta$  per ogni scelta di  $\theta_0$  e  $\theta_1$  con  $\theta_1 > \theta_0$ .

Conviene notare che

- a) se  $L(\theta_1 | \mathbf{x})$  soddisfa alla proprietà di *monotonia del rapporto di verosimiglianza* rispetto a  $L(\theta_0 | \mathbf{x})$ , più grande è il valore di  $T_n(\mathbf{x}) = t_n$  più verosimile è che il campione (di cui si osserva la realizzazione  $\mathbf{x}$ ) sia stato estratto dalla distribuzione indicizzata da  $\theta_1$  piuttosto che da  $\theta_0$
- b) in presenza di campionamento da distribuzione appartenente a *famiglia esponenziale*, la monotonia in  $\theta$  della componente  $A(\theta)$  garantisce la monotonia del rapporto di verosimiglianza in

$$T_n(\mathbf{X}) = \sum_{i=1}^n B(X_i) \quad (11.64)$$

**Esempio 11.1.4.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Normale  $N(0, \sigma^2)$  e sia dato il seguente sistema di ipotesi

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ vs. \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases} \quad (11.65)$$

da verificare fissato in  $\alpha$  illivello di significatività.

Il rapporto di verosimiglianza è dato da

$$\frac{L(\sigma^2; \mathbf{x})}{L(\sigma_0^2; \mathbf{x})} = \left( \frac{\sigma^2}{\sigma_0^2} \right)^{-\frac{n}{2}} \exp \left\{ \frac{1}{2} \left( \frac{1}{\sigma_0^2} - \frac{1}{\sigma^2} \right) \sum_{i=1}^n x_i^2 \right\} \quad (11.66)$$

che, in virtù della monotonia della componente  $A(\sigma) = (\sigma_0^{-2} - \sigma^{-2})$ , risulta essere funzione monotonamente crescente di  $T_n(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i^2$  attraverso la quale (e solo) dipende da  $(X_1, X_2, \dots, X_n)$ . Di conseguenza, il test individuato tramite il lemma di Neyman-Pearson (ipotizzando  $H_1 : \sigma^2 = \sigma_1^2, \sigma_1^2 > \sigma_0^2$ ) è non solo più potente come afferma il lemma ma anche *uniformemente più potente* (vale a dire, per ogni valore di  $\sigma^2$  sotto ipotesi alternativa).

Passando al logaritmo del rapporto di verosimiglianza, con qualche passaggio si ha

$$\frac{1}{\sigma_0^2} \sum_{i=1}^n x_i^2 > \frac{2 \ln(A) - n(\ln(\sigma_0^2) - \ln(\sigma^2))}{\sigma^2 - \sigma_0^2} \cdot \sigma^2 = B \quad (11.67)$$

Ora, sotto  $H_0$ :  $\frac{X_i}{\sigma_0} \sim N(0, 1)$  sicché  $\frac{X_i^2}{\sigma_0^2} \sim \chi_1^2$  e per la proprietà di riproducibilità,

$$\sum_{i=1}^n \frac{X_i^2}{\sigma_0^2} \sim \chi_n^2 \quad (11.68)$$

per cui, sfruttando il lemma di Neyman-Pearson e la monotonia del rapporto di verosimiglianza, la regione critica del test UMP per il sistema di ipotesi in questione è data da

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{1}{\sigma_0^2} \sum_{i=1}^n X_i^2 > \chi_{n;\alpha}^2 \right\} \quad (11.69)$$

Ora, generalizzando un po',

a) se  $\mu \neq 0$  e comunque nota, sotto  $H_0 : \sigma^2 = \sigma_0^2$

$$\frac{1}{\sigma_0^2} \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma_0} \right)^2 \sim \chi_n^2 \quad (11.70)$$

e la regione critica  $C_\alpha$  del test UMP sarà data da

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \mu)^2 > \chi_{n;\alpha}^2 \right\}, \quad \text{per } H_1 : \sigma^2 > \sigma_0^2 \quad (11.71)$$

mentre

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \mu)^2 < \chi_{n;1-\alpha}^2 \right\}, \quad \text{per } H_1 : \sigma^2 < \sigma_0^2 \quad (11.72)$$

b) se  $\mu$  non è nota

$$\frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{(n-1) S_n^2}{\sigma_0^2} \sim \chi_{n-1}^2 \quad (11.73)$$

e la regione critica  $C_\alpha$  del test UMP sarà data da

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{(n-1) s_n^2}{\sigma_0^2} > \chi_{n-1; \alpha}^2 \right\}, \quad \text{per } H_1 : \sigma^2 > \sigma_0^2 \quad (11.74)$$

mentre

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{(n-1) s_n^2}{\sigma_0^2} < \chi_{n-1; 1-\alpha}^2 \right\}, \quad \text{per } H_1 : \sigma^2 < \sigma_0^2 \quad (11.75)$$

**Esempio 11.1.5.** Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale dalla distribuzione  $N(\mu, \sigma^2)$  con  $\sigma^2$  noto e consideriamo il seguente sistema di ipotesi:

$$\begin{cases} H_0 : \mu = \mu_0 \\ vs. \\ H_1 : \mu > \mu_0 \end{cases} \quad (11.76)$$

In questo caso il rapporto di verosimiglianza è dato da

$$\frac{L(\mu | \mathbf{x})}{L(\mu_0 | \mathbf{x})} = \exp \left\{ \frac{n}{2\sigma^2} (\mu - \mu_0) \bar{x}_n - \frac{n}{2\sigma^2} (\mu^2 - \mu_0^2) \right\} \quad (11.77)$$

che, essendo  $(\mu - \mu_0) > 0$ , risulta essere funzione monotona crescente di  $\bar{x}_n$ . Sicché il più potente test individuato dal lemma di Neyman-Pearson per  $H_1 : \mu = \mu_1$  è anche uniformemente più potente per alternative unilaterali del tipo  $H_1 : \mu > \mu_0$  e la regione critica (anch'essa UMP) sarà data da

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \bar{x}_n > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\}. \quad (11.78)$$

Analogo discorso può essere fatto in presenza di alternative unilaterali del tipo  $H_1 : \mu < \mu_0$ . Lascio per esercizio individuare la regione critica del test UMP.

Ma in presenza di un sistema di ipotesi del tipo

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ vs. \\ H_1 : \theta > \theta_0 \end{cases} \quad (11.79)$$

con  $H_0$  e  $H_1$  entrambe unilaterali, esiste ancora un test UMP nella classe dei test di livello  $\alpha$ ? il seguente teorema fornisce la risposta a questa domanda.

**Teorema 11.1.2** (Teorema di Karlin e Rubin). Sia dato il sistema di ipotesi

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_1 : \theta > \theta_0 \quad (11.80)$$

e

a) sia  $T_n(X_1, X_2, \dots, X_n)$  una statistica sufficiente per  $\theta$

b) la famiglia di distribuzioni da cui proviene il campione  $(X_1, X_2, \dots, X_n)$  goda della proprietà di monotonia del rapporto di verosimiglianza in  $T_n$ .

Allora, il test con regione critica

$$C_\alpha = \{ \mathbf{x} \in \mathfrak{X} : T_n(\mathbf{x}) > t_\alpha \} \quad (11.81)$$

con  $t_\alpha$  valore critico del test, è un test UMP di livello  $\alpha = P(T_n > t_\alpha | H_0 : \theta = \theta_0)$ .

## 11.2 Test del rapporto di verosimiglianza generalizzato

Finora abbiamo considerato

- a) test *più potenti* per *ipotesi nulla e alternativa entrambe semplici* la cui regione critica era restituita dal lemma di Neyman-Pearson
- b) test *uniformemente più potenti*, estendendo il lemma a *ipotesi nulla semplice* contro *alternativa composita unilaterale* in presenza di famiglie di distribuzioni con rapporto di verosimiglianza monotono nella statistica (sufficiente)  $T_n$ .

E che cosa si può dire del caso generale, ovvero *ipotesi nulla e alternativa entrambe composite*? E in particolare, esiste ancora un test UMP?

**Definizione 11.2.1** (Test del rapporto di massima verosimiglianza). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione  $F_X(x; \theta), \theta \in \Theta \subset \mathbb{R}^k$  e si voglia testare il seguente sistema di ipotesi

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ vs. \\ H_1 : \theta \in \Theta_1 \end{cases} \quad (11.82)$$

con  $\Theta = \Theta_0 \cup \Theta_1$  e  $\Theta_0 \cap \Theta_1 = \emptyset$  e definiamo nella seguente maniera il *rapporto di massima verosimiglianza* (o *generalizzato*)

$$\lambda(\mathbf{x}) = \frac{\max_{\theta \in \Theta_0} L(\theta | \mathbf{x})}{\max_{\theta \in \Theta} L(\theta | \mathbf{x})} = \frac{L(\hat{\theta}_0 | \mathbf{x})}{L(\hat{\theta}_n | \mathbf{x})} \quad (11.83)$$

Allora la regione critica  $C_\alpha$  del test del rapporto di massima verosimiglianza è formata da tutti i punti  $\mathbf{x}$  dello spazio campionario  $\mathfrak{X}$  che soddisfano alla relazione

$$\lambda(\mathbf{x}) = \frac{\max_{\theta \in \Theta_0} L(\theta | \mathbf{x})}{\max_{\theta \in \Theta} L(\theta | \mathbf{x})} \leq A \quad (11.84)$$

ovvero

$$C_\alpha = \left\{ \mathbf{x} \in \mathfrak{X} : \frac{L(\hat{\theta}_0 | \mathbf{x})}{L(\hat{\theta}_n | \mathbf{x})} \leq A \right\} \quad (11.85)$$

dove  $A$  è scelto in modo che la probabilità di commettere un errore di primo tipo sia uguale a  $\alpha$ .

Conviene rimarcare che la costante  $A$  sarà sempre inferiore a uno e potrà essere determinata sulla base della distribuzione di probabilità di  $\lambda(\mathbf{x})$  in corrispondenza del livello di significatività del test fissato tramite  $P(\lambda(\mathbf{x}) < A | H_0) = \alpha$ . Inoltre, come suggerisce l'intuito,  $\lambda(\mathbf{x})$  sarà *prossimo a uno* se  $\theta \in \Theta_0$  mentre sarà *prossimo a zero* se  $\theta \notin \Theta_0$ .

La distribuzione (esatta) di probabilità di

$$\lambda(\mathbf{x}) = \frac{\max_{\theta \in \Theta_0} L(\theta | \mathbf{x})}{\max_{\theta \in \Theta} L(\theta | \mathbf{x})} \quad (11.86)$$

a volte è così complicata da rendere molto difficoltoso il calcolo dei suoi quantili che sono quantità indispensabili per formulare esplicitamente la regione critica del test di interesse di livello  $\alpha$  fissato.

Quando però la dimensione del campione è *sufficientemente grande* per verificare il sistema di ipotesi (11.82) possiamo utilizzare il seguente risultato asintotico.

**Teorema 11.2.1** (Teorema di Wilks). Sotto le usuali e blande condizioni di regolarità che devono essere soddisfatte dalla distribuzione da cui proviene il campione, per ogni  $\theta \in \Theta_0$

$$-2 \ln(\lambda(\mathbf{X})) \underset{a}{\sim} \chi_{\nu}^2 \quad (11.87)$$

dove  $\nu$  è dato dalla *differenza* tra il numero di parametri che si devono stimare sotto  $H_1$  e quello sotto  $H_0$ .

Qualche considerazione:

- a) l'applicazione del test di verosimiglianza generalizzato a un sistema di ipotesi semplici dà luogo a  
i risultati identici a quelli che si ottengono con il lemma di Neyman-Pearson essendo in quel caso

$$\lambda(\mathbf{x}) = \frac{\max_{\theta \in \Theta_0} L(\theta | \mathbf{x})}{\max_{\theta \in \Theta} L(\theta | \mathbf{x})} = \frac{L(\theta_0 | \mathbf{x})}{L(\theta_1 | \mathbf{x})} \quad (11.88)$$

- b) la statistica test  $\lambda(\mathbf{x})$  dipende dal campione casuale *solo* attraverso una *statistica sufficiente (minimale)* al pari di quanto accadeva alla statistica test restituita dal lemma di Neyman-Pearson
- c) il test del rapporto di verosimiglianza generalizzato dà luogo a test *più potenti* o a test *uniformemente più potenti* se questi esistono (alternative semplici o composite unilaterali)
- d) il test basato sul rapporto di verosimiglianza generalizzato è *asintoticamente UMP* nella classe dei *test non distori* ovvero per i quali

$$\eta_{C_\alpha}(\theta) \geq \alpha, \quad \forall \theta \in \Theta. \quad (11.89)$$

**Esempio 11.2.1** (Test t-Student a un campione unilaterale). Sia  $(X_1, X_2, \dots, X_n)$  un campione casuale da una distribuzione Normale  $N(\mu, \sigma^2)$  con entrambi i parametri non noti e si vuole verificare

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ vs. \\ H_1 : \mu > \mu_0 \end{cases} \quad (11.90)$$

Ora,  $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$  sicché le stime di massima verosimiglianza di  $(\mu, \sigma^2)$  saranno

- a) su tutto  $\Theta$  (*non vincolate*):  $\hat{\mu}_n = \bar{x}_n$  e  $\hat{\sigma}_n^2 = \frac{n-1}{n} s_n^2$

b) su  $\Theta_0 = \{(\mu, \sigma^2) : \mu \leq \mu_0, \sigma^2 > 0\}$ :

$$\hat{\mu}_0 = \min(\mu_0, \hat{\mu}_n)$$

e

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_0)^2 \rightarrow \begin{cases} = \hat{\sigma}_n^2, & \text{se } \hat{\mu}_0 = \bar{x}_n \\ > \hat{\sigma}_n^2, & \text{se } \hat{\mu}_0 = \mu_0 \neq \bar{x}_n \end{cases}$$

Allora,

$$\begin{aligned} \lambda(\mathbf{x}) &= \frac{\max_{(\mu, \sigma^2) \in \Theta_0} L(\mu, \sigma^2 | \mathbf{x})}{\max_{(\mu, \sigma^2) \in \Theta} L(\mu, \sigma^2 | \mathbf{x})} = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_n^2} = \left[ \frac{\sum_{i=1}^n (x_i - \hat{\mu}_0)^2}{\sum_{i=1}^n (x_i - \hat{\mu}_n)^2} \right]^{-\frac{n}{2}} \\ &= \left[ \frac{\sum_{i=1}^n (x_i - \hat{\mu}_n)^2 + n(\hat{\mu}_n^2 - \hat{\mu}_0^2)}{\sum_{i=1}^n (x_i - \hat{\mu}_n)^2} \right]^{-\frac{n}{2}} \\ &= \left[ 1 + \frac{t^2}{n-1} \right]^{-\frac{n}{2}} \end{aligned} \quad (11.91)$$

dove  $T = (\bar{X}_n - \hat{\mu}_0) / (S_n / \sqrt{n}) \underset{H_0}{\sim} t_{n-1}$ .

La regione critica  $C_\alpha$  del test associato al sistema di ipotesi (11.90) è allora data dai valori di  $T > T_{n-1; \alpha}$  (vale a dire, da *piccoli valori* di  $\lambda(\mathbf{x})$ ) ossia,

$$C_\alpha = \{\mathbf{x} \in \mathfrak{X} : t > t_{n-1; \alpha}\} \quad (11.92)$$

Se fosse stato da testare  $H_0 : \mu \geq \mu_0$  vs.  $H_1 : \mu < \mu_0$  la regione critica associata al test in questione sarebbe stata

$$C_\alpha = \{\mathbf{x} \in \mathfrak{X} : t < t_{n-1; 1-\alpha}\} \quad (11.93)$$

ricordando che, per la simmetria della distribuzione *t-Student*,  $t_{n-1; 1-\alpha} = -t_{n-1; \alpha}$ .

**Esempio 11.2.2** (Test t-Student a un campione bilaterale). Consideriamo il campione casuale  $(X_1, X_2, \dots, X_n)$  dalla distribuzione  $N(\mu, \sigma^2)$ , con  $\mu$  e  $\sigma^2$  non noti. Consideriamo il seguente sistema di ipotesi composite:

$$\begin{cases} H_0 : \mu = \mu_0 \\ \text{vs.} \\ H_1 : \mu \neq \mu_0 \end{cases} \quad (11.94)$$

Abbiamo già mostrato che

$$\lambda(\mathbf{x}) = \left( 1 + \frac{t^2}{n-1} \right)^{-\frac{n}{2}} \quad (11.95)$$

con

$$T = \frac{\sqrt{n}(\bar{X}_n - \hat{\mu}_0)}{S_n} \underset{H_0}{\sim} t_{n-1} \quad (11.96)$$

Pertanto, tenuto conto della forma dell'ipotesi alternativa  $H_1 : \mu \neq \mu_0$ , la regione critica di livello  $\alpha$  per il test in questione è data da

$$C_\alpha = \{\mathbf{x} \in \mathfrak{X} : |t| > t_{n-1, \alpha/2}\} \quad (11.97)$$

Torniamo sul risultato asintotico contenuto nel teorema di Wilks, ossia sul fatto che

$$-2 \ln(\lambda(\mathbf{X})) \underset{a}{\sim} \chi_{\nu}^2 \quad (11.98)$$

Senza alcuna pretesa di dimostrare questo risultato, a puro scopo illustrativo, è interessante (e istruttivo) notare che nel caso del test t-Student unilaterale

$$-2 \ln(\lambda(\mathbf{X})) = n \ln \left( 1 + \frac{T^2}{n-1} \right) \underset{a}{\sim} \chi_{1-\alpha}^2 \quad (11.99)$$

Infatti, ricorrendo all'espansione in serie di Taylor della funzione  $g(t) = n \ln \left( 1 + \frac{t^2}{n-1} \right)$  intorno a  $t = 0$  si ha che

$$\begin{aligned} g(t) &= n \ln \left( 1 + \frac{t^2}{n-1} \right) \Big|_{t=0} + \frac{2nt}{(n-1)+t^2} \Big|_{t=0} \cdot t + 2n \frac{(n-1)-t^2}{[(n-1)+t^2]^2} \Big|_{t=0} \cdot \frac{t^2}{2} \\ &= n \frac{(n-1)}{(n-1)^2} \cdot t^2 \\ &= \frac{n}{n-1} \cdot t^2. \end{aligned} \quad (11.100)$$

Già sappiamo che  $T \xrightarrow{D} N(0, 1)$  sicché

$$\lim_{n \rightarrow \infty} \frac{n}{n-1} \cdot T^2 = T^2 \xrightarrow{D} \chi_1^2 \quad (11.101)$$

essendo  $T^2$  il quadrato di una variabile casuale asintoticamente distribuita come una Normale standard.

### 11.2.1 Test $t$ per campioni appaiati

Nell'ambito di campioni (=dati) appaiati, ciascuno degli  $n$  soggetti (o unità sperimentali) appartenenti a un gruppo viene osservato in due distinti momenti (per esempio, *prima* e *dopo* di un certo trattamento) e si procede a rilevare il valore della variabile di interesse assunto nei due momenti (siamo quindi nell'ambito di *misure ripetute* su di uno stesso soggetto). La ragione di questo modo di procedere consiste nel valutare l'*efficacia* del trattamento.

Se indichiamo con  $x_i$  il valore della variabile  $X$  osservata sull'i-mo soggetto *prima* del trattamento e con  $v_i$  il valore della medesima variabile sempre osservata sul medesimo soggetto ma *dopo* il trattamento, otterremo  $n$  coppie  $(x_i, v_i)$  di valori di  $X$  pre e post trattamento.

Possiamo supporre che  $x_i$  sia stata tratta da distribuzione  $N(\eta, \sigma^2)$  e che che  $v_i$  sia stata invece tratta da distribuzione  $N(\mu, \sigma^2)$  sotto ipotesi di *omoschedasticità* delle due distribuzioni generanti. Di conseguenza, il problema relativo alla verifica dell'efficacia del trattamento si può ridurre a testare il seguente sistema di ipotesi

$$\begin{cases} H_0 : \mu - \eta = 0 & \equiv \delta_0 = \mu - \eta = 0 \\ vs. \\ H_1 : \mu - \eta \neq 0 \end{cases} \quad (11.102)$$

Definiamo pertanto  $y_i = v_i - x_i$ , con  $i = 1, \dots, n$ . Ci aspettiamo che  $Y_i \sim N(\mu - \eta, \sigma_*^2)$  dove la varianza  $\sigma_*^2$  dipenderà da  $\sigma^2$  e dall'incognita correlazione tra  $x_i$  e  $v_i$ :

$$\sigma_*^2 = \text{Var}(X - V) = \text{Var}(X) + \text{Var}(V) - 2 \text{Cov}(X, Y). \quad (11.103)$$

A questo punto è sufficiente applicare un *t-test* a un campione al precedente sistema di ipotesi coinvolgendo le nuove variabili  $Y_i = V_i - X_i$  che possono essere trattate come determinazioni di variabili casuali indipendenti in quanto riferite a soggetti distinti. Allora,

$$\lambda(\mathbf{y}) = \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} \quad (11.104)$$

con

$$T = \frac{\sqrt{n}(\bar{Y}_n - \delta_0)}{S_n} \stackrel{H_0}{=} \frac{\sqrt{n}\bar{Y}_n}{S_n} \stackrel{H_0}{\sim} t_{n-1} \quad (11.105)$$

e siccome  $\lambda(\mathbf{y})$  è una funzione monotona descrescente di  $|t|$ , allora rifiutare  $H_0$  per piccoli valori di  $\lambda(\mathbf{y})$  è equivalente rifiutare  $H_0$  per grandi valori di  $|t|$  sicché la *regione critica* di livello  $\alpha$  fissato sarà data da

$$C_\alpha = \{y \in \mathcal{Y} : |t| > t_{n-1; \alpha/2}\}. \quad (11.106)$$

### 11.2.2 Test sulle medie di m popolazioni (o One-way ANOVA)

L'analisi della varianza (ANalysis Of VAriance) è un tecnica di fondamentale importanza in statistica in particolare in ambito di disegno degli esperimenti ed è stata introdotta in letteratura da R. Fisher negli anni venti dello scorso secolo (*Statistical Methods for Research Workers* (1925)); essa consente di stabilire se le medie di due o più gruppi di unità statistiche possono essere ritenute uguali o meno; in altre parole, l'analisi della varianza permette di sottoporre a verifica il seguente sistema di ipotesi

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_m \\ vs. \\ H_1 : \text{almeno una delle } m \text{ medie è diversa dalle altre} \end{cases} \quad (11.107)$$

Alla base dell'analisi della varianza vi è un set standard di assunzioni che brevemente richiamiamo (sono peraltro già note per quanto visto in precedenza); queste in sostanza affermano che

- a) le unità campionarie coinvolte sono scelte attraverso uno schema di campionamento casuale semplice e assegnate casualmente ai  $k$  gruppi
- b) la variabile risposta  $Y$  è assunta essere normalmente distribuita all'interno di ciascuno dei  $k$  gruppi ovvero  $Y_{ij} \sim N(\mu_i, \sigma^2)$  con  $i = 1, 2, \dots, m$  e  $j = 1, 2, \dots, n_i$  con  $n$  numerosità dell' $i$ -mo gruppo (se gli  $m$  gruppi hanno tutti lo stesso numero di unità si dice che il disegno sperimentale è *bilanciato*)
- c) le medie possono differire da gruppo a gruppo mentre le deviazioni standard si assumono essere uguali per tutti i gruppi (ipotesi di omoschedasticità).

Come si può facilmente intuire, nel caso in cui il numero  $m$  dei gruppi sia uguale a due, nulla cambia rispetto al tradizionale test  $t$  di Student per la verifica dell'uguaglianza delle medie (basti ricordare l'esempio sulle due popolazioni di ragni, *Deinopis* e *Menneus*); è però vero che tale test diviene impraticabile se  $m > 2$  sicché, in quel caso, l'analisi della varianza mette in gioco tutta la sua utilità.

Ricorrendo all'approccio basato sul rapporto di verosimiglianza generalizzato abbiamo che sotto  $H_0$  le stime di  $\mu_i$  e di  $\sigma^2$  sono

$$\hat{\mu}_0 = \frac{1}{n \cdot m} \sum_{i=1}^m \sum_{j=1}^n y_{ij} = \bar{y}_n \quad \text{e} \quad \hat{\sigma}_0^2 = \frac{1}{n \cdot m} \sum_{j=1}^n (y_{ij} - \bar{y}_n)^2 \quad (11.108)$$

dal momento che potremmo ritenere le unità degli  $m$  gruppi formare un solo gruppo di numerodità  $(n \cdot m)$ ; sotto  $H_1$  le stime di  $\mu_i$  sono date da

$$\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^n y_{ij} \quad i = 1, 2, \dots, m \quad (11.109)$$

mentre la stima di  $\sigma^2$  è data da

$$\hat{\sigma}^2 = \sum_{i=1}^m \hat{\sigma}_i^2 \frac{1}{m} = \frac{1}{m} \sum_{i=1}^m \left( \frac{1}{n} \sum_{j=1}^n (y_{ij} - \hat{\mu}_i)^2 \right) \quad (11.110)$$

e quindi

$$\begin{aligned} \lambda(\mathbf{y}) &= \frac{L(\hat{\mu}_0, \dots, \hat{\mu}_0, \hat{\sigma}_0^2 | \mathbf{y})}{L(\hat{\mu}_1, \dots, \hat{\mu}_m, \hat{\sigma}^2 | \mathbf{y})} \\ &= \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-\frac{(n \cdot m)}{2}} \\ &= \left[ 1 + \frac{n \sum_{i=1}^m (\hat{\mu}_i - \hat{\mu}_0)^2}{\sum_{i=1}^m \sum_{j=1}^n (y_{ij} - \hat{\mu}_i)^2} \right]^{-\frac{(n \cdot m)}{2}} \\ &= \left[ 1 + \frac{D_0}{D} \right]^{-\frac{(n \cdot m)}{2}} \end{aligned} \quad (11.111)$$

dove  $D_0$  e  $D$  rappresentano rispettivamente la *devianza tra i gruppi* e la *devianza nei gruppi*. Ed è naturale rifiutare  $H_0$  in favore di  $H_1$  se  $D_0$  è grande rispetto  $D$ .

La regione critica del test è allora data da quei valori per cui

$$\frac{n \sum_{i=1}^m (\hat{\mu}_i - \hat{\mu}_0)^2}{\sum_{i=1}^m \sum_{j=1}^n (y_{ij} - \hat{\mu}_i)^2} = \frac{D_0}{D} \quad (11.112)$$

è sufficientemente elevato (e ciò comporta un valore di  $\lambda(\mathbf{y})$  piccolo). In altri termini, si rifiuta  $H_0$  per grandi valori di  $\frac{D_0}{D}$

Sotto  $H_0$  il termine  $\frac{D_0}{\sigma^2}$  è una somma di  $m$  termini *indipendenti* ognuno distribuito come  $\chi^2_{n-1}$  sicché, per la proprietà di riproducibilità

$$\frac{D}{\sigma^2} \sim \chi^2_{m(n-1)} \quad (11.113)$$

mentre, sempre sotto  $H_0$ ,

$$\frac{D_0}{\sigma^2} = \sum_{i=1}^m \left( \frac{\hat{\mu}_i - \hat{\mu}_0}{\sigma/\sqrt{n}} \right)^2 \sim \chi_{m-1}^2 \quad (11.114)$$

essendo quest'ultima la varianza campionaria delle quantità

$$\frac{\hat{\mu}_1}{\sigma/\sqrt{n}}, \dots, \frac{\hat{\mu}_m}{\sigma/\sqrt{n}} \quad (11.115)$$

Inoltre,  $D_0$  e  $D$  sono determinazioni di variabili casuali indipendenti essendo funzione rispettivamente delle sole medie campionarie  $\hat{\mu}_i$  e delle sole varianze campionarie  $\hat{\sigma}_i^2$ ,  $i = 1, 2, \dots, m$ , che sono tra loro indipendenti in virtù di un noto teorema. Ma allora, sotto  $H_0$ ,

$$F = \frac{D_0/(m-1)\sigma^2}{D/(m(n-1)\sigma^2)} = \frac{(D_0/\sigma^2)/(m-1)}{(D/\sigma^2)/(m(n-1))} \underset{H_0}{\sim} F_{(m-1),m(n-1)} \quad (11.116)$$

ossia è una determinazione di una variabile casuale  $F$  di Fisher-Snedecor con  $(m - m(n - 1))$  gradi di libertà e la regione critica del test sarà data dai valori della statistica  $F$  maggiori del valore critico  $F_{(m-1),m(n-1);\alpha}$  che può essere ricavato dalle tavole della distribuzione  $F$ , ossia

$$C_\alpha = \{ \mathbf{y} \in \mathcal{Y} : F > F_{(m-1),m(n-1);\alpha} \} \quad (11.117)$$

dove  $F_{(m-1),m(n-1);\alpha}$  è il quantile di ordine  $(1 - \alpha)$  della distribuzione  $F$  con  $(m - 1)$  e  $m(n - 1)$  gradi di libertà.

### Esempio (One-way ANOVA)

Un produttore di borse di carta per la spesa è interessato a studiare come la resistenza alla trazione della borsa sia in relazione con la percentuale di pasta di legno duro presente nell'impasto di cui le borse sono fatte.

I risultati che seguono sono relativi alla resistenza alla trazione di un campione 24 borse costruite con diverse percentuali di pasta di legno duro.

<b>Unità</b>	<b>Resistenza alla trazione (psi)</b>	<b>Concentrazione di legno duro (%)</b>	
1	7	5	
2	8	5	
3	15	5	
4	11	5	
5	9	5	
6	10	5	
7	12	10	
8	17	10	
9	13	10	
10	18	10	
11	19	10	
12	15	10	
13	14	15	
14	18	15	
15	19	15	
16	17	15	
17	16	15	
18	18	15	
19	19	20	
20	25	20	
21	22	20	
22	23	20	
23	18	20	
24	20	20	
			<b>10,000</b>
			<b>8,000</b>
			<b>2,828</b>
			<b>15,667</b>
			<b>7,867</b>
			<b>2,805</b>
			<b>17,000</b>
			<b>3,200</b>
			<b>1,789</b>
			<b>21,167</b>
			<b>6,967</b>
			<b>2,639</b>

<b>Media</b>	<b>15,958</b>
Varianza	22,303
DevStd	4,723

<b>Dev_TOT</b>	<b>512,958</b>
----------------	----------------

Possiamo riassumere l'informazione nella tabella che segue:

<i>Tr_1</i>	<i>Tr_2</i>	<i>Tr_3</i>	<i>Tr_4</i>
<b>Numerosità</b>	6	6	6
<b>Media</b>	10,000	15,667	17,000
<b>Devianza</b>	40,000	39,333	16,000

sicché

<b>( D<sub>0</sub> : DEV<sub>TRA</sub> )</b>	
	<b>382,792</b>
<b>DEV<sub>TOT</sub></b>	
	<b>512,958</b>
<b>( D : DEV<sub>IN</sub> )</b>	
	<b>130,167</b>

Ora,

<b>Variable</b>	<b>Observations</b>	<b>Mean</b>	<b>Std. deviation</b>
Resistenza alla trazione (psi)	24	15,958	4,723

### ANalysis Of VAriance:

<b>Source</b>	<b>DF</b>	<b>Sum of squares</b>	<b>Mean squares</b>	<b>F</b>	<b>Pr &gt; F</b>
Model	3	382,792	127,597	<b>19,605</b>	< 0,0001
Error	20	130,167	6,508		
Corrected Total	23	512,958			

Computed against model Y=Mean(Y)

Model parameters:

<b>Source</b>	<b>Value</b>	<b>Std error</b>	<b>t</b>	<b>Pr &gt;  t </b>	<b>LB (95%)</b>	<b>UB (95%)</b>
Intercept	21,167	1,041	20,323	< 0,0001	18,994	23,339
Conc(5 %)	-11,167	1,473	-7,581	< 0,0001	-14,239	-8,094
Conc(10 %)	-5,500	1,473	-3,734	0,001	-8,572	-2,428
Conc(15 %)	-4,167	1,473	-2,829	0,010	-7,239	-1,094
Conc(20 %)	0,000	0,000				

Equation of the model:

$$\text{Resistenza alla trazione} = 21,167 - 11,167 * \text{Conc}(5\%) - 5,500 * \text{Conc}(10\%) - 4,167 * \text{Conc}(15\%)$$