Alma Mater Studiorum · Università di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea in Matematica

TEOREMA DI COCHRAN E APPLICAZIONI

Tesi di Laurea in Probabilità e Statistica Matematica

Relatore:
Chiar.mo Prof.
ANDREA PASCUCCI

Presentata da:
MARTINA MANCINI

II Sessione
Anno Accademico 2014-2015

...Lentamente muore chi abbandona un progetto prima di iniziarlo,
chi non fa domande sugli argomenti che non conosce,
chi non risponde quando gli si chiede qualcosa che conosce.
Evitiamo la morte a piccole dosi,
ricordando sempre che essere vivo richiede uno sforzo di
gran lunga maggiore
del semplice fatto di respirare!
Soltanto l'ardente pazienza porterà al raggiungimento di
una splendida
felicità.

Pablo Neruda

Indice

Introduzione

La statistica è un ramo della matematica che studia i metodi per raccogliere, organizzare e analizzare un insieme di dati numerici, la cui variazione è influenzata da cause diverse, con lo scopo sia di descrivere le caratteristiche del fenomeno a cui i dati si riferiscono, sia di dedurre, ove possibile, le leggi generali che lo regolano. La statistica si suddivide in statistica descrittiva o deduttiva e in statistica induttiva o inferenza statistica. Noi ci occuperemo di approfondire la seconda, nella quale si studiano le condizioni per cui le conclusioni dedotte dall'analisi statistica di un campione sono valide in casi più generali. In particolare l'inferenza statistica si pone l'obiettivo di indurre o inferire le proprietà di una popolazione (parametri) sulla base dei dati conosciuti relativi ad un campione. Lo scopo principale di questa tesi è analizzare il Teorema di Cochran e illustrarne le possibili applicazioni nei problemi di stima in un campione Gaussiano. In particolare il Teorema di Cochran riguarda un'importante proprietà delle distribuzioni normali multivariate, che risulta fondamentale nella determinazione di intervalli di fiducia per i parametri incogniti.

Nel primo capitolo vengono presentati gli strumenti necessari per affrontare i problemi della statistica, in particolare i problemi di stima. Vengono richiamate, infatti, la definizione di distribuzione Normale con alcune proprietà e altre leggi note; quali la legge Gamma, la legge del Chi Quadrato e la legge t di Student. Le leggi appena menzionate sono accomunate da una importante proprietà: la simmetria, che si trova alla base di qualsiasi calcolo di intervalli di confidenza per un parametro sconosciuto.

Il secondo capitolo è articolato in tre parti. Inizialmente vengono introdotti alcuni concetti generali che riguardano i problemi di stima per creare un background adeguato a presentare l'argomento centrale dell'elaborato. In seguito vengono fatti alcuni esempi di

stimatori corretti per media e varianza. Infine si affronta nello specifico il Teorema di Cochran e la sua dimostrazione.
Il terzo capitolo è dedicato all'esposizione di alcune applicazioni del Teorema di Cochran: la stima della media e della varianza per campioni Gaussiani, la regressione lineare semplice e la regressione lineare multipla. Nel problema della stima della media vengono differenziati due casi: nel primo caso la varianza è nota, nel secondo anche la varianza risulta sconosciuta. Questa situazione è senza dubbio più adattabile ai problemi reali, infatti viene applicata ad un esempio pratico. Infine vengono esaminati il modello di regressione lineare semplice e il modello di regressione lineare multipla con relativi esempi.

Capitolo 1

Nozioni preliminari

In questo capitolo vengono analizzati gli strumenti necessari per comprendere l'argomento trattato nel presente elaborato.

1.1 Alcuni concetti di probabilità

1.1.1 Distribuzione Normale

Si dice che una densità di probabilità su $\mathbb{R}$ è *normale* (o *Gaussiana*) di parametri μ e σ^2 (oppure che è $N(\mu, \sigma^2)$), se è della forma

$$g(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

In particolare, la densità f di una $N(0,1)$ è

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Una proprietà fondamentale della legge $N(0,1)$ è la simmetria, quindi X e $-X$ hanno la stessa legge. Da questa proprietà si ricava la relazione

$$\begin{aligned}\Phi(x) = P\{X \leq x\} = P\{-X \leq x\} = P\{X \geq -x\} = \\ = 1 - P\{X \leq -x\} = 1 - \Phi(-x)\end{aligned} \tag{1.1}$$

Sempre la simmetria della legge $N(0,1)$ permette di ricavare alcune relazioni molto utili per i quantili; di solito si indica con il simbolo ϕ_α il quantile di ordine α della $N(0,1)$. I quantili di uso più frequente sono:

$$\begin{aligned}\phi_{0.95} &= 1.644854 \\ \phi_{0.975} &= 1.959964\end{aligned} \tag{1.2}$$

Siccome il numero ϕ_α è caratterizzato dalla relazione $P\{X \leq \phi_\alpha\} = \alpha$, abbiamo infatti

$$P\{X \leq -\phi_\alpha\} = P\{-X \leq -\phi_\alpha\} = P\{X \geq \phi_\alpha\} = 1 - P\{X \leq \phi_\alpha\} = 1 - \alpha$$

da cui la relazione

$$-\phi_\alpha = \phi_{1-\alpha} \tag{1.3}$$

Dalla (1.3) si ottiene un'altra relazione importante:

$$\begin{aligned}P\{|X| \leq \phi_{1-\frac{\alpha}{2}}\} = P\{-\phi_{1-\frac{\alpha}{2}} \leq X \leq \phi_{1-\frac{\alpha}{2}}\} = \\ = P\{X \leq \phi_{1-\frac{\alpha}{2}}\} - P\{X \leq -\phi_{1-\frac{\alpha}{2}}\} = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha\end{aligned} \tag{1.4}$$

Osserviamo che per ottenere le (1.3) e (1.4) si sfrutta solo la proprietà di simmetria della legge $N(0,1)$. Quindi queste due relazioni sono valide per tutte le leggi simmetriche.

1.1.2 Leggi Gamma

Si chiama *funzione gamma* la funzione $\Gamma : \mathbb{R}^+ \to \mathbb{R}^+$ definita da

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$$

L'integrale non è, per valori generici di α, calcolabile elementarmente; si vede però (con una integrazione per parti) che per ogni $\alpha > 0$ è

$$\alpha\Gamma(\alpha) = \Gamma(\alpha + 1)$$

Inoltre $\Gamma(1) = 1$; quindi per ogni n intero positivo è

$$\Gamma(n) = (n-1)!$$

Infine la sostituzione $t = x^{\frac{1}{2}}$ dà $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
Si chiama *Legge Gamma* di parametri α e λ (entrambi positivi), abbreviata in $\Gamma(\alpha, \lambda)$, la legge di una variabile aleatoria con valori reali positivi e densità

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x > 0 \\ 0 & \text{altrimenti} \end{cases}$$

Una importante proprietà delle leggi gamma è espressa nel seguente teorema:

Teorema 1.1.1. *Se X_1 e X_2 sono variabili aleatorie indipendenti, con leggi rispettivamente $\Gamma(\alpha_1, \lambda)$ e $\Gamma(\alpha_2, \lambda)$, allora la legge di $X_1 + X_2$ è $\Gamma(\alpha_1 + \alpha_2, \lambda)$.*

Il teorema si estende alla somma di quante variabili si desidera perché se X_1, X_2, X_3 sono indipendenti allora anche $(X_1 + X_2)$ e X_3 lo sono. Quindi $X_1 + X_2 + X_3$ ha legge $\Gamma(\alpha_1 + \alpha_2 + \alpha_3, \lambda)$ e similmente $X_1 + \ldots + X_n$ ha legge $\Gamma(\alpha_1 + \ldots + \alpha_n, \lambda)$.

1.1.3 La legge di X^2

Mostriamo che, se $X \sim N(0, \sigma^2)$ allora X^2 è una $\Gamma(\frac{1}{2}, \frac{1}{2\sigma^2})$.
La densità di X è $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$. Sia G la funzione di ripartizione di X^2, dunque $G(y) = P(X^2 \leq y)$. Se $y \leq 0$ ovviamente $G(y) = 0$; se $y > 0$, allora

$$G(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx = \\ = 2\int_0^{\sqrt{y}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx \tag{1.5}$$

Pongo $x = \sqrt{t}$; in questo modo si ha

$$G(y) = \int_0^y \frac{1}{\sigma\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-\frac{t}{2\sigma^2}} dt = \int_0^y \frac{1}{\sigma\sqrt{2\pi}} t^{\frac{1}{2}-1} e^{-\frac{1}{2\sigma^2}t} dt$$

e quindi la densità è nulla per $y \leq 0$ e per $y \geq 0$ vale

$$g(y) = \frac{1}{\sigma\sqrt{2\pi}} y^{\frac{1}{2}-1} e^{-\frac{1}{2\sigma^2}y}$$

Questa nella parte non costante è la densità di una $\Gamma(\frac{1}{2}, \frac{1}{2\sigma^2})$. Siccome l'integrale su $\mathbb{R}$ di g deve dare 1, la costante è necessariamente determinata in modo da corrispondere a quella di una $\Gamma(\frac{1}{2}, \frac{1}{2\sigma^2})$ che è

$$\frac{(\frac{1}{2\sigma^2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} = \frac{1}{\sigma\sqrt{2\pi}}$$

In particolare se $X \sim N(0,1)$ allora $X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$.

1.1.4 Legge del Chi Quadrato con n gradi di libertà

Si chiama *Legge del Chi Quadrato* con n gradi di libertà, indicata con $\chi^2(n)$, la legge della variabile somma dei quadrati di n variabili aleatorie indipendenti $X_1, \ldots, X_n$, ciascuna con legge $N(0,1)$.
Per ciascuna X_i abbiamo visto che X_i^2 è una $\Gamma(\frac{1}{2}, \frac{1}{2})$; per le proprietà delle distribuzioni Gamma segue allora

$$\chi^2(n) \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

cioè la densità di probabilità di $\chi^2(n)$ è

$$g(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{se } x > 0 \\ 0 & \text{se } x \leq 0 \end{cases}$$

1.1.5 Legge t di Student

Si chiama legge *t di Student* con n gradi di libertà (e si scrive $t(n)$) la legge di una v.a. Z della forma

$$Z = \frac{X}{\sqrt{Y}}\sqrt{n}$$

dove X e Y sono v.a. indipendenti di legge rispettivamente $N(0,1)$ e $\chi^2(n)$. Non è difficile calcolare la densità di una v.a. di legge $t(n)$, ma per i nostri scopi è sufficiente conoscerne numericamente i quantili (che indicheremo con $t_\alpha(n)$), e per questo vi sono delle tavole come quella riportata nell'Appendice B. Si osserva che i valori tabulati si avvicinano per n grande a quelli corrispondenti della legge $N(0,1)$. I quantili delle leggi di Student godono di proprietà interessanti. Intanto queste leggi sono simmetriche: se la v.a. Z è di Student allora Z e $-Z$ hanno la stessa distribuzione; infatti poiché le v.a. $N(0,1)$ sono simmetriche si ha

$$-Z = \frac{-X}{\sqrt{Y}}\sqrt{n} \sim \frac{X}{\sqrt{Y}}\sqrt{n} = Z$$

Dalla proprietà di simmetria seguono le relazioni

$$P\{Z \leq -t_{1-\alpha}(n)\} = \alpha \tag{1.6}$$

$$P\{|Z| \geq t_{1-\frac{\alpha}{2}}(n)\} = \alpha \tag{1.7}$$

Capitolo 2

Teorema di Cochran

2.1 Problemi di stima

La statistica induttiva si pone l'obiettivo di pervenire a una stima dei parametri caratteristici di una v.a. a partire da valutazioni campionarie della stessa.
È consuetudine indicare il parametro incognito (scalare o vettoriale) con θ, e con Θ l'insieme dei valori che θ può assumere. Talvolta ciò che si desidera stimare è θ medesimo, in particolare quando θ è uno scalare; in altri casi si è interessati a una funzione $\psi(\theta)$ del parametro θ.

Definizione 2.1. Si chiama *modello statistico* una famiglia di spazi di probabilità $(\Omega, \mathcal{A}, (P^\theta)_{\theta\in\Theta})$ con P^θ dipendente dal parametro θ.

Definizione 2.2. Una *osservazione* è un vettore $X = (X_1, \ldots, X_n)$ di v.a. definite su Ω; X ha densità di probabilità $x \to f_n(x|\theta)$, condizionata al valore θ del parametro, ossia (nel caso di densità continua),

$$P^\theta\{X \in A\} = \int_A f_n(x|\theta)dx$$

Definizione 2.3. Consideriamo un modello statistico ed una funzione $\psi : \Theta \to \mathbb{R}^m$ del parametro (eventualmente vettoriale) θ. Sia $t : \mathbb{R}^n \to \mathbb{R}^m$ una funzione sufficientemente regolare. La funzione dell' osservazione $X = (X_1, \ldots, X_n)$ descritta da $T = t(X_1, \ldots, X_n)$ si chiama *statistica*.

Se i valori di t appartengono al codominio di ψ T si chiama *stimatore* di $\psi(\theta)$. Intuitivamente questo equivale ad assumere T come approssimazione di $\psi(\theta)$. Uno stimatore di $\psi(\theta)$ è una qualunque funzione delle osservazioni, non necessariamente legata al significato che di volta in volta assume $\psi(\theta)$. Per questa ragione risulta necessario individuare dei criteri per stabilire quali sono i buoni stimatori.

Definizione 2.4. Si dice che $T = t(X_1, \ldots, X_n)$ è uno *stimatore corretto* di $\psi(\theta)$ se per ogni $\theta \in \Theta$ risulta $E^\theta(T) = \psi(\theta)$ ($E^\theta(T)$ rappresenta la media di T relativamente alla probabilità P^θ).

Uno stimatore non corretto si dice *distorto*. Tale proprietà di uno stimatore T è in generale una qualità apprezzabile per utilizzare T come approssimazione di $\psi(\theta)$, tuttavia esistono esempi di buoni stimatori che non godono di questa proprietà.

Definizione 2.5. Sia $X_1, \ldots, X_n$ una sequenza di v.a. indipendenti ed equidistribuite, cioè con la stessa legge; se f_θ è la densità di ciascuna X_i (per un dato θ in Θ), la densità congiunta di $X = (X_1, \ldots, X_n)$ è

$$f_n((x_1, \ldots, x_n)|\theta) = f_\theta(x_1) \ldots f_\theta(x_n)$$

In questa situazione si dice che un vettore aleatorio $X = (X_1, \ldots, X_n)$ è un *campione* di *rango* (o *numerosità*) n.

Definizione 2.6. Sia $0 < \alpha < 1$. Siano $T_1 = t_1(X_1, \ldots, X_n)$ e $T_2 = t_2(X_1, \ldots, X_n)$ due statistiche per campioni di rango n; sia $I_x = [T_1, T_2]$. Si dice che I_x è un *intervallo di fiducia* o di *confidenza* per $\psi(\theta)$ di livello $1 - \alpha$, se per ogni $\theta \in \Theta$ si ha

$$P^\theta\{\psi(\theta) \in I_x\} \geq 1 - \alpha$$

È evidente che l'interesse di un intervallo di confidenza è tanto maggiore quanto più α è piccolo, quindi è vicino a 1 il livello di confidenza $1 - \alpha$. Assai frequentemente si usa individuare intervalli di confidenza al 95%, vale a dire con $\alpha = 0.05$.

2.2 Stimatori per media e varianza

In questa sezione faremo degli esempi di stimatori corretti della media e della varianza, che utilizzeremo in seguito. Sia $X = (X_1, \ldots, X_n)$ un campione di rango n di variabili indipendenti ed equidistribuite, con speranza matematica finita per ogni $\theta \in \Theta$. La media campionaria $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ è uno stimatore corretto della media comune di ciascuna X_i. Infatti, per la linearità della speranza matematica, qualunque sia $\theta \in \Theta$ si ha

$$E^\theta(\bar{X}) = \frac{1}{n}(E^\theta(X_1) + \ldots + E^\theta(X_n)) = E^\theta(X_i)$$

Supponiamo ora le X_i con media e varianza finite, per ogni $\theta \in \Theta$, e andiamo alla ricerca di uno stimatore corretto per la varianza $var_\theta(X_i)$. È opportuno distinguere due casi: il primo, in cui la media è nota, e vale μ; il secondo, più vicino alle situazioni concrete, nel quale né la media né la varianza di X_i sono conosciute. Nel secondo caso, più interessante,

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

è uno stimatore corretto per la $var_\theta(X_i)$.

2.3 Teorema di Cochran

Teorema 2.3.1 (Teorema di Cochran)**.** *Sia X una variabile aleatoria m-dimensionale di legge $N(0, I)$ (vale a dire, ciascuna X_i ha legge $N(0,1)$ e $X_1, \ldots, X_m$ sono indipendenti). Siano $E_1, \ldots, E_k$ sottospazi vettoriali di $\mathbb{R}^m$ a 2 a 2 ortogonali. Per $i = 1, \ldots, k$ indichiamo con n_i la dimensione di E_i e con P_i il proiettore ortogonale su E_i. Allora le variabili aleatorie P_iX, $i = 1, \ldots, k$, sono indipendenti e la variabile aleatoria $|P_iX|^2$ ha distribuzione $\chi^2(n_i)$.*

Dimostrazione. Supponiamo per semplicità $k = 2$ e consideriamo il caso che E_1 sia il sottospazio relativo alle prime n_1 coordinate e E_2 quello relativo alle successive n_2. Dunque, le proiezioni su E_1 e E_2 sono rispettivamente

$$P_1X = (X_1, \ldots, X_{n_1}, 0, \ldots, 0)$$
$$P_2X = (0, \ldots, 0, X_{n_1+1}, \ldots, X_{n_1+n_2}, 0, \ldots, 0)$$

Se poniamo $Y = P_1X$, $Z = P_2X$ è chiaro che $Cov(Y_i, Z_j) = 0$ per ogni $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, poiché le v.a. $X_1, \ldots, X_m$ sono indipendenti. P_1X e P_2X sono v.a. non correlate con distribuzione congiunta normale, quindi possiamo concludere che sono indipendenti.
Inoltre, ricordando che la somma dei quadrati di k v.a. $N(0,1)$ indipendenti segue una legge $\chi^2(k)$

$$|P_1X|^2 = X_1^2 + \ldots + X_{n_1}^2 \sim \chi^2(n_1)$$
$$|P_2X|^2 = X_{n_1+1}^2 + \ldots + X_{n_1+n_2}^2 \sim \chi^2(n_2)$$

Se E_1 ed E_2 non sono come in questo caso particolare, si può però mostrare che esiste una trasformazione ortogonale O tale che OE_1 e OE_2 siano appunto i sottospazi generati rispettivamente dalle prime n_1 coordinate e dalle successive n_2. Sfruttando il fatto che una trasformazione ortogonale muta una variabile normale multivariata $N(0, I)$ ancora in $N(0, I)$ ci si può ricondurre al caso appena trattato. La seconda parte della prova risulta elementare. □

Una conseguenza del Teorema di Cochran, fondamentale per le applicazioni ai problemi di stima in statistica, è la seguente:

Corollario 2.3.2. *Siano $Z_1, \ldots, Z_m$ v.a. indipendenti e tutte di legge $N(\mu, \sigma^2)$. Poniamo*

$$\bar{Z} = \frac{1}{m}(Z_1 + \ldots + Z_m)$$
$$S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Z_i - \bar{Z})^2$$

Allora $\bar{Z}$ e S^2 sono indipendenti e si ha

$$\frac{m-1}{\sigma^2}S^2 \sim \chi^2(m-1) \tag{2.1}$$

$$\frac{\sqrt{m}(\bar{Z}-\mu)}{S} \sim t(m-1) \tag{2.2}$$

Dimostrazione. Innanzitutto consideriamo il caso più semplice di v.a. $N(0,1)$. Indichiamo con E il sottospazio di $\mathbb{R}^m$ generato dal vettore $e = (1,\ldots,1)$ (cioè il sottospazio dei vettori aventi tutte le coordinate uguali come nell'Esempio A.1) e con $E^\perp$ il suo ortogonale, ovvero l'insieme dei vettori aventi la somma delle coordinate uguale a 0. Sappiamo dall'Esempio A.1 che il proiettore ortogonale $P_E : \mathbb{R}^m \to E$ è dato da $P_E x = (\bar{x},\ldots,\bar{x})$. Se $X \sim N(0,I)$ e $\bar{X} = \frac{1}{m}(X_1 + \ldots + X_m)$, allora la proiezione ortogonale di X su E è $P_E X = (\bar{X},\ldots,\bar{X})$ e quindi per la (A.2), la proiezione ortogonale di X su $E^\perp$ è

$$P_{E^\perp}X = (I - P_E)X = X - P_E X = (X_1 - \bar{X},\ldots,X_m - \bar{X})$$

Per il Teorema 2.3.1 $\bar{X}$ e $|P_{E^\perp}X|^2$ sono v.a. indipendenti. Inoltre

$$|P_{E^\perp}X|^2 = \sum_{i=1}^{m}(X_i - \bar{X})^2 = (m-1)S^2 \tag{2.3}$$

Per il Teorema di Cochran dunque $(m-1)S^2 \sim \chi^2(m-1)$ e S^2 è indipendente da $\bar{X}$. Infine poiché $\bar{X}$ ha legge $N(0,\frac{1}{m})$, $\sqrt{m}\bar{X}$ è $N(0,1)$ e per la definizione della legge di Student si ha che

$$T = \frac{\sqrt{m}\bar{X}}{S} \sim t(m-1) \tag{2.4}$$

Per dimostrare il corollario basta ricondursi a questo caso. Poniamo

$$X_i = \frac{Z_i - \mu}{\sigma}$$

e dunque il vettore $X = (X_1, \ldots, X_m)$ è $N(0, I)$ e per le considerazioni appena fatte $\bar{X}$ e $S_X^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X})^2$ sono indipendenti. Tenendo conto che $Z_i = \sigma X_i + \mu$ si ha

$$\bar{Z} = \sigma \bar{X} + \mu \tag{2.5}$$

$$\sum_{i=1}^m (X_i - \bar{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^m (Z_i - \bar{Z})^2 = \frac{m-1}{\sigma^2} S^2 \tag{2.6}$$

e ne segue che anche $\bar{Z}$ e S^2 sono indipendenti come funzioni di variabili indipendenti. Infine $\frac{m-1}{\sigma^2} S^2 \sim \chi^2(m-1)$ per la (2.6), e poiché

$$\frac{\sqrt{m}(\bar{Z} - \mu)}{S} = \frac{\sqrt{m}\bar{X}}{S_X}$$

la (2.2) segue dalla (2.4). □

Capitolo 3

Applicazioni

3.1 Stima della media per campioni Gaussiani

Consideriamo un campione di v.a. $X_1, \ldots, X_n$ di legge Gaussiana. Il nostro obiettivo è costruire intervalli di confidenza per la media μ di questo campione. In primo luogo è necessario distinguere due casi.

3.1.1 Primo caso: la varianza σ^2 è nota

Un primo caso semplice (ma raramente utile nella pratica) consiste nel supporre che le osservazioni abbiano ciascuna legge $N(\mu, \sigma^2)$, dove σ^2 è un numero fissato e conosciuto. La media e la varianza di $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ sono rispettivamente μ e $\frac{\sigma^2}{n}$ e, di conseguenza, la variabile

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

si distribuisce come una $N(0,1)$. Indicando con $\phi_{1-\frac{\alpha}{2}}$ il quantile di ordine $1 - \frac{\alpha}{2}$ della legge $N(0,1)$, si ha

$$\begin{aligned} 1 - \alpha = P^\mu \left\{ \left| \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \right| \leq \phi_{1-\frac{\alpha}{2}} \right\} = P^\mu \left\{ -\phi_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \leq \phi_{1-\frac{\alpha}{2}} \right\} = \\ = P^\mu \left\{ \bar{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \right\} \end{aligned} \tag{3.1}$$

Dunque,

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right]$$

è un intervallo di fiducia di livello $1-\alpha$ per μ.

3.1.2 Secondo caso: la varianza σ^2 non è nota

Nei problemi reali, quando si vuole costruire un intervallo di confidenza per la media di un campione Normale, raramente si conosce la varianza del campione. Sia pertanto $X = (X_1, \ldots, X_n)$ un campione di legge Normale $N(\mu, \sigma^2)$ con μ e σ^2 ignoti; ci occupiamo del problema di stimare μ. Siano $\bar{X}$ e S^2 gli stimatori corretti di μ e σ^2, che abbiamo ricavato sopra. È noto che la media normalizzata, esplicitata nella v.a.

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$$

e, per il Corollario al Teorema di Cochran

$$W = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$$

ossia la varianza normalizzata segue una distribuzione χ^2 con $n-1$ gradi di libertà. Inoltre, sempre per il Corollario (2.3.2) Z e W sono indipendenti. Allora la v.a.

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S} = \frac{Z}{\sqrt{W}}\sqrt{n-1} \sim t(n-1)$$

Questo risultato è molto importante perché nell'espressione di T non compare σ^2 che non è conosciuta. Dunque, mediante i quantili della legge di Student possiamo determinare intervalli di confidenza per μ. Fissato $\alpha \in]0,1[$, abbiamo

$$\begin{aligned} 1-\alpha = P^\mu\{|T| \leq t_{1-\frac{\alpha}{2}}(n-1)\} = P^\mu\left\{\left|\frac{\sqrt{n}(\bar{X}-\mu)}{S}\right| \leq t_{1-\frac{\alpha}{2}}(n-1)\right\} = \\ = P^\mu\left\{\bar{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right\} \end{aligned} \tag{3.2}$$

Rispetto al primo caso σ è stata sostituita dal suo stimatore S, come era ragionevole aspettarsi; la differenza più significativa sta nel fatto che i quantili della legge normale sono stati sostituiti da quelli delle leggi di Student (che sono un po' più grandi). È bene osservare che l'intervallo di confidenza è più ampio di quello ottenuto nel caso di σ^2 nota. Anche questo è ragionevole: una minore informazione in partenza porta come conseguenza una minore precisione nel risultato che si ottiene. Notiamo che si possono ricavare altri intervalli di confidenza al livello $1-\alpha$ per μ ; per esempio dall'uguaglianza $P^\mu\{T \geq -t_{1-\alpha}(n-1)\} = 1-\alpha$ si ottiene

$$]-\infty, \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1)]$$

L'intervallo centrato è tra tutti quello di minima lunghezza, perché si colloca dove la densità di probabilità di t è più concentrata. Per questo motivo esso fornisce (fissato il livello di fiducia) la stima con maggiore precisione, nel senso che minimizza la lunghezza dell'intervallo di confidenza. In determinate applicazioni serve tuttavia una stima unilaterale.

Esempio 3.1. In una città circola un certo numero n di taxi (che non conosciamo, ma sappiamo essere dell'ordine di qualche centinaio), numerati con un numero progressivo, da 1 a n. Un giornalista prende nota dei numeri di 20 taxi che vede passare in una strada del centro

$$\{233, 74, 317, 110, 114, 104, 222, 394, 226, 149, 358, 259, 106, 79, 73, 357, 16, 81, 267, 287\}$$

Come può utilizzare questi valori per stimare il numero n di taxi della sua città?
Risoluzione.
L'osservazione X_i del numero di taxi i-esimo che ci passa davanti è una variabile aleatoria con valori interi equidistribuiti tra 1 e n; la sua media è $\mu = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} =$
$= \frac{n+1}{2}$, quindi $n = 2\mu - 1$. Dunque stimare n equivale a stimare μ. Si osserva che le variabili X_i non hanno legge normale, tuttavia la media campionaria di un numero abbastanza grande di osservazioni rende questo difetto meno rilevante. La media campionaria sul campione osservato assume il valore $\bar{X} = 191.3$; la varianza stimata è

$S^2 = \frac{1}{19}\sum_{i=1}^{20}(X_i - \bar{X})^2 = 13741$; questi valori consentono di determinare un intervallo di confidenza per μ. Se, per esempio scegliamo $\alpha = 0.05$, otteniamo il seguente intervallo di fiducia per μ

$$\left[\bar{X} - \frac{S}{\sqrt{20}}t_{0.975}(19), \bar{X} + \frac{S}{\sqrt{20}}t_{0.975}(19)\right] = [136.439, 246.161]$$

che porta all'intervallo di confidenza per n: $[271.877, 491.323]$. È utile constatare che una limitazione inferiore certa del numero di taxi è ovvia, ed è il massimo dei numeri che abbiamo osservato; nel nostro caso 394. Dunque conviene cercare un intervallo di confidenza inferiormente illimitato per μ. Applicando il ragionamento esposto sopra si ha

$$\left]-\infty, \bar{X} + \frac{S}{\sqrt{20}}t_{0.95}(19)\right]$$

che corrisponde all'intervallo di confidenza per $n = 2\mu - 1$

$$\left]-\infty, 2\left(\bar{X} + \frac{S}{\sqrt{20}}t_{0.95}(19)\right) - 1\right]$$

il quale con gli attuali dati sperimentali è $]-\infty, 472.247]$. Questo risultato combinato con l'informazione certa $n \geq 394$, dà l'intervallo nel quale n si trova con probabilità 95%: $[394, 472.47]$.

I numeri di questo esempio sono stati generati con *Mathematica 10*, sfruttando le considerazioni appena fatte.

$\boldsymbol{n = \text{RandomInteger}[\{300, 500\}];}$

$\boldsymbol{m = 20;}$

$\boldsymbol{x = \text{RandomInteger}[\{1, n\}, m];}$

$\boldsymbol{\text{media} = \text{Mean}[x];}$

$\boldsymbol{\text{s2} = \frac{m\text{Variance}[x]}{m-1}; s = \sqrt{\text{s2}};}$

$\boldsymbol{\text{alfa} = 0.05;}$

$t = \text{Quantile}\left[\text{StudentTDistribution}[m-1], 1-\frac{\text{alfa}}{2}\right];$

$\text{t1} = \text{Quantile}[\text{StudentTDistribution}[m-1], 1-\text{alfa}];$

$a = \text{media} - \frac{st}{\sqrt{m}}; b = \frac{st}{\sqrt{m}} + \text{media};$

$c = \frac{st1}{\sqrt{m}} + \text{media};$

$\text{Print}[x]$

$\text{Print}[\text{“media campionaria osservata e stimatore della varianza”}]$

$\text{Print}[\{N[\text{media}], N[\text{s2}]\}]$

$\text{Print}[\text{“intervallo di confidenza per la media”}]$

$\text{Print}[N[\{a, b\}]]$

$\text{Print}[\text{“intervallo di confindenza per n”}]$

$\text{Print}[N[\{2a-1, 2b-1\}]]$

$\text{Print}[\text{“migliore intervallo di confidenza per n”}]$

$\text{Print}[\{\max(x), 2c-1\}]$

$\text{Print}[\text{“valore effettivo di n”}]$

$\text{Print}[n]$

$\{233, 74, 317, 110, 114, 104, 222, 394, 226, 149, 358, 259, 106, 79, 73, 357, 16, 81, 267, 287\}$

media campionaria osservata e stimatore della varianza

$\{191.3, 13741.\}$

intervallo di confidenza per la media

$\{136.439, 246.161\}$

intervallo di confindenza per n

$\{271.877, 491.323\}$

migliore intervallo di confidenza per n

$\{394, 472.247\}$

valore effettivo di n

397

Il valore vero di n calcolato alla fine appartiene all'intervallo sopra determinato.

3.2 Stima della varianza per campioni Gaussiani

Definizione 3.1. Si chiama *quantità pivotale* una v.a. $Q(X, \theta)$ tale che la sua legge rispetto a P^θ non dipenda da θ.

Ciò equivale a dire che la quantità $P^\theta\{Q(X, \theta) \in A\}$ non dipende da θ per ogni intervallo $A \subset \mathbb{R}$. Se la legge di Q è nota, allora per ogni $\alpha \in]0, 1[$ se ne possono determinare i quantili, ossia si possono trovare due numeri q_1 e q_2 tali che

$$P^\theta\{q_1 \leq Q(X, \theta) \leq q_2\} = 1 - \alpha$$

Supponiamo ora che la relazione

$$q_1 \leq Q(X, \theta) \leq q_2$$

possa essere risolta rispetto a θ, in modo che si abbia $t_1(X) \leq \theta \leq t_2(X)$ con t_1, t_2 opportune funzioni. Segue allora che

$$P^\theta\{t_1(X) \leq \theta \leq t_2(X)\} = 1 - \alpha$$

vale a dire $[t_1(X), t_2(X)]$ è un'intervallo di confidenza per θ al livello $1 - \alpha$.
Per stimare la varianza di un campione Gaussiano si può utilizzare il metodo della quantità pivotale che abbiamo appena introdotto. Supponiamo che la media μ non sia nota; come conseguenza del Corollario (2.3.2) si ha che

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

e quindi $\frac{(n-1)S^2}{\sigma^2}$ è una quantità pivotale, per cui

$$1-\alpha = P^{\sigma^2}\left\{\frac{n-1}{\sigma^2}S^2 \geq \chi^2_\alpha(n-1)\right\} = P^{\sigma^2}\left\{\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_\alpha(n-1)}\right\}$$

Ciò significa che

$$[0, \frac{n-1}{\chi^2_\alpha(n-1)}S^2]$$

è un'intervallo di fiducia di livello $1-\alpha$ per σ^2. Per un intervallo bilaterale si può ripetere il ragionamento precedente:

$$\begin{aligned} 1-\alpha &= P^{\sigma^2}\left\{\chi^2_{\frac{\alpha}{2}}(n-1) \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{1-\frac{\alpha}{2}}(n-1)\right\} = \\ &= P^{\sigma^2}\left\{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right\} \end{aligned} \tag{3.3}$$

da cui si ottiene che l'intervallo

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

è di fiducia di livello $1-\alpha$ per σ^2.

3.3 Regressione lineare semplice

Un problema statistico molto frequente è quello in cui si considera una possibile relazione lineare tra una variabile y e altre variabili $x_1, \ldots, x_n$ più una perturbazione aleatoria. Il più semplice modello di regressione è il modello di regressione lineare semplice, in cui si assume che

$$y_i = \beta_0 + \beta_1 x_i + w_i$$

dove $w_1, \ldots, w_n$ sono v.a. indipendenti $N(0, \sigma^2)$ e σ^2 è indipendente da i. La variabile y si chiama *variabile dipendente* mentre x è il *predittore*. Questo modello dipende dai parametri incogniti β_0, β_1, σ^2. Ora ci occuperemo di stimare questi parametri. Per

determinare uno stimatore di β_0 e β_1 si può ragionare nel modo seguente: se consideriamo i punti $(x_1, y_1), \ldots, (x_n, y_n)$ nel piano, la retta di regressione $y = \beta_0 + \beta_1 x$ deve essere tale che la distanza della retta dai punti sia minima. Cerchiamo cioè i valori di β_0, β_1 per cui la quantità

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

sia minima. Si tratta di risolvere un problema di minimo di una funzione (di tipo quadratico) di due variabili (β_0 e β_1) che si risolve cercando i valori che annullano le derivate parziali.

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{3.4}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{3.5}$$

Se indichiamo con b_0, b_1 le soluzioni otteniamo

$$b_0 = \bar{y} - b_1 \bar{x}$$

dove

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Sostituendo il valore ottenuto per b_0 in (3.5) si ottiene

$$b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

Si può semplificare la notazione indicando il denominatore di b_1 con $\bar{\sigma}_x^2$.
Vediamo ora alcune proprietà di questi stimatori. Intanto essi sono non distorti; approfondiremo questo concetto successivamente quando parleremo della regressione multipla. Veniamo al calcolo delle varianze: poniamo per semplicità

$$v_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

e sfruttiamo le seguenti equivalenze

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} y_i(x_i - \bar{x})$$
$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

per scrivere

$$b_1 = \sum_{i=1}^{n} v_i(y_i - \bar{y})$$

Inoltre vale la relazione

$$\sum_{i=1}^{n} v_i^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Allora per le proprietà della varianza e ricordando che $Var(y_i) = \sigma^2$

$$Var(b_1) = \sum_{i=1}^{n} v_i^2 Var(y_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{\bar{\sigma}_x^2} \tag{3.6}$$

Per quanto riguarda la varianza di b_0 si ha

$$Var(b_0) = Var(\bar{y}) + \bar{x}^2 Var(b_1) - 2\bar{x}Cov(\bar{y}, b_1) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\bar{\sigma}_x^2}\right) \tag{3.7}$$

Infatti $Var(\bar{y}) = \frac{\sigma^2}{n}$, mentre $\bar{y}$ e b_1 sono non correlate poiché

$$Cov(\bar{y}, b_1) = \sum_{i=1}^{n} v_i Cov(\bar{y}, y_i - \bar{y}) =$$
$$\sum_{i=1}^{n} v_i\big(Cov(y_i, \bar{y}) - Cov(\bar{y}, \bar{y})\big) = \sum_{i=1}^{n} v_i\left(\frac{\sigma^2}{n} - \frac{\sigma^2}{n}\right) = 0$$

Se teniamo conto del fatto che le v.a. w_i sono Gaussiane anche le y_i risultano Gaussiane; poiché esse sono anche indipendenti, la loro legge congiunta è normale e poiché gli stimatori b_0 e b_1 sono funzioni lineari-affini delle y_i, anch'essi hanno legge normale. In

conclusione

$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\bar{\sigma}_x^2}\right)\right)$$
$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\bar{\sigma}_x^2}\right)$$

Per determinare degli intervalli di fiducia per β_0 e β_1 resta da stimare σ^2. Si chiamano *valori stimati* le quantità

$$\hat{y} = b_0 + b_1 x_i$$

e *residui* le quantità

$$r_i = y_i - \hat{y}_i$$

Poniamo infine

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2$$

Risulta, come dimostreremo in seguito nel capitolo dedicato alla regressione multipla, che

$$X := \frac{s^2}{\sigma^2}(n-2) \sim \chi^2(n-2)$$

ed inoltre che s^2 è indipendente da b_0 e b_1. Dunque s^2 è uno stimatore non distorto di σ^2, infatti la media di una v.a. $\chi^2(n-2)$ vale appunto $n-2$. Applicando la definizione delle leggi di Student si ha

$$Z := \frac{b_0 - \beta_0}{\sqrt{Var(b_0)}} = \frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\bar{\sigma}_x^2}}} \sim N(0,1)$$

Allora

$$T_0 := \frac{b_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\bar{\sigma}_x^2}}} = \frac{Z}{\sqrt{X}}\sqrt{n-2} \tag{3.8}$$

segue una legge di Student con $n-2$ gradi di libertà. Allo stesso modo si ha

$$T_1 := \frac{b_1 - \beta_1}{s}\bar{\sigma}_x \sim t(n-2) \tag{3.9}$$

T_0 e T_1 sono dunque delle quantità pivotali, grazie alle quali possiamo calcolare intervalli di fiducia per β_0 e β_1. Applicando il solito ragionamento si ricava facilmente che

$$\left[b_0 - s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\bar{\sigma}_x^2}}t_{1-\frac{\alpha}{2}}(n-2), b_0 + s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\bar{\sigma}_x^2}}t_{1-\frac{\alpha}{2}}(n-2)\right]$$
$$\left[b_1 - \frac{s}{\bar{\sigma}_x}t_{1-\frac{\alpha}{2}}(n-2), b_1 + \frac{s}{\bar{\sigma}_x}t_{1-\frac{\alpha}{2}}(n-2)\right]$$

sono intervalli di fiducia di livello $1-\alpha$ rispettivamente per β_0 e β_1. È opportuno mettere in evidenza alcune proprietà dei residui che ci porteranno a introdurre una nuova quantità:

$$\begin{aligned} &\sum_{i=1}^{n} r_i = 0 \\ &\sum_{i=1}^{n} r_i \hat{y}_i = 0 \end{aligned} \tag{3.10}$$

La prima delle (3.10) segue dalla definizione dei residui e dalla (3.4). Per la seconda basta osservare che

$$\sum_{i=1}^{n} r_i \hat{y}_i = b_0 \sum_{i=1}^{n} r_i + b_1 \sum_{i=1}^{n} r_i x_i$$

e abbiamo appena visto che la prima somma a destra è nulla, mentre la seconda è uguale a 0 grazie alla (3.5). Dalla prima delle (3.10) si ricava

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

A questo punto possiamo introdurre una nuova quantità che è spesso utile considerare

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

In questa espressione il denominatore è più grande del numeratore perché

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \big((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\big)^2 = \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \\
&= \underbrace{\sum_{i=1}^n r_i^2}_{\geq 0} + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \underbrace{2\sum_{i=1}^n r_i(\hat{y}_i - \bar{y})}_{=0} \geq \sum_{i=1}^n (\hat{y}_i - \bar{y})^2
\end{aligned}$$

e dunque $0 \leq R^2 \leq 1$. La quantità R^2 è la proporzione di variazione di y che viene spiegata attraverso il modello di regressione. Essa è uno degli indici che conviene considerare quando si vuole valutare la bontà del modello o la sua correttezza, infatti al crescere del valore di R^2 diminuiscono le distanze dei punti osservati dalla retta di regressione.

3.3.1 Test statistici applicati alla regressione lineare semplice

Nelle applicazioni del modello di regressione lineare semplice si è spesso condotti a considerare dei test statistici riguardanti le tre quantità $\beta_0, \beta_1, \sigma^2$. In particolare da questo momento in poi ci occuperemo di stabilire se il parametro incognito θ è di un certo tipo oppure no, ossia se un determinato modello è accettabile. In un problema di test si è in presenza di una partizione $\{\Theta_H, \Theta_A\}$ di Θ e si vuole stabilire se $\theta \in \Theta_H$. L'insieme Θ_H si chiama *ipotesi* mentre Θ_A è l'alternativa. L'obiettivo di un test è scegliere se respingere o no l'ipotesi, quindi bisogna stabilire quali siano i valori delle osservazioni che conducono al rigetto di quest'ultima. Chiameremo questo insieme di valori delle osservazioni la *regione critica* o di *rigetto* del test. In generale qualunque sia la scelta della regione

critica, se l'ipotesi è vera vi è una probabilità positiva di avere un'osservazione nella regione di rigetto e quindi di respingere a torto l'ipotesi (*errore di prima specie*). Allo stesso modo vi è una probabilità positiva di non respingere una ipotesi falsa (*errore di seconda specie*).

Definizione 3.2. Si chiama *potenza* del test di regione critica D la funzione $\pi_D : \Theta \to [0,1]$ definita da

$$\pi_D(\theta) = P^\theta\{X \in D\}$$

Dunque se $\theta \in \Theta_H$ $\pi_D(\theta)$ è la probabilità di respingere a torto l'ipotesi (cioè l'errore di prima specie) se il vero valore del parametro è θ.

Definizione 3.3. Si chiama *livello* del test di regione critica D la quantità

$$\alpha_D = \sup_{\theta\in\Theta_H} P^\theta\{X \in D\} = \sup_{\theta\in\Theta_H} \pi_D(\theta)$$

È chiaro che il livello α_D è l'estremo superiore delle probabilità di errore di prima specie. In genere poiché l'errore di prima specie è considerato il più grave, si cerca sempre di determinare una regione di rigetto per un valore del livello pari ad un prefissato α (tipicamente i valori sono $\alpha = 10\%, 5\%, 1\%$).
Nell'esempio seguente prenderemo in esame uno dei più interessanti test statistici che coinvolgono il modello di regressione lineare semplice.

Esempio 3.2. In un esperimento diretto allo studio della relazione tra il numero di pulsazioni sotto sforzo (per minuto) e l'età (in anni). Ora eseguiamo un interessante test statistico che consiste nel verificare se la variabile dipendente (le pulsazioni) dipende dal predittore (l'età). In particolare si assume $\alpha = 0.05$ e si fa l'ipotesi

$$H : \beta_1 = 0 \text{ contro } A : \beta_1 \neq 0 \tag{3.11}$$

Si deduce facilmente, grazie a (3.9), che

$$\left\{ \left| \frac{b_1}{s} \bar{\sigma}_x \right| \geq t_{1-\frac{\alpha}{2}}(n-2) \right\}$$

è una regione di rigetto di livello α per il test considerato. Lo stesso ragionamento vale anche per l'ipotesi $H : \beta_0 = 0$.
Sono stati rilevati i seguenti dati su 10 soggetti di sesso maschile:

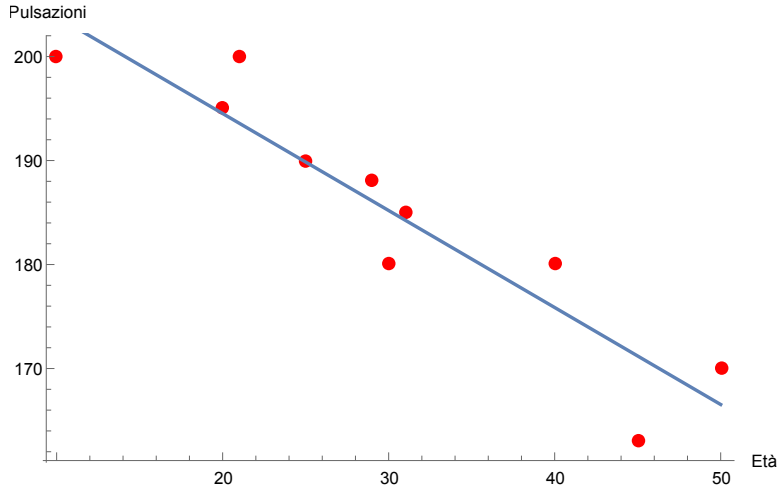
Età	Pulsazioni
10	200
20	195
21	200
25	190
29	188
30	180
31	185
40	180
45	63
50	170

I calcoli riportati nella seguente tabella sono stati effettuati utilizzando il software *Mathematica 10*:

Predittore	Coef	DevSt	TStat	p
b_0	213.172	4.22964	50.3995	2.66005×10^{-11}
b_1	-0.932628	0.1312	-7.10845	0.000101145
$s = 3.203$	$R^2 = 0.863318$			

Nella prima colonna vengono esplicitati gli stimatori b_0 e b_1 e nella seconda ne viene data la deviazione standard (cioè la radice quadrata della varianza, calcolata tramite la (3.6) e la (3.7), sostituendo al valore sconosciuto σ^2 la sua stima s^2). La terza colonna contiene il quoziente tra la prima e la seconda, ossia riporta le statistiche T_0 e T_1 definite nella (3.8) e nella (3.9), calcolate ponendo $\beta_0 = 0$ e $\beta_1 = 0$ rispettivamente. Poiché queste statistiche seguono una legge $t(n-2)$ nelle ipotesi $\beta_0 = 0$ e $\beta_1 = 0$, e abbiamo visto che regioni di rigetto in questi casi al livello α sono rispettivamente $\{|T_0| \geq t_{1-\frac{\alpha}{2}}\}$ e $\{|T_1| \geq t_{1-\frac{\alpha}{2}}\}$, avremo che queste ipotesi sono respinte se per il valore di $TStat$ corrispondente si ha $|TStat| \geq t_{1-\frac{\alpha}{2}}$. Se si confrontano i valori di $TStat$ con i corrispondenti quantili della Tabella B.1 si potrebbe già dedurre che le ipotesi vengono rigettate. L'ultima colonna

riporta appunto la probabilità $P\{|X| \geq TStat\}$, dove $X \sim t(n-2)$, ossia il livello di significatività del test. Dunque concludiamo che le ipotesi $H : \beta_1 = 0$ e $H : \beta_0 = 0$ vengono respinte poiché la quantità nell'ultima colonna è più piccola del livello α stabilito ($\alpha = 0.05$). Nell'ultima riga si trovano i valori di s e della quantità R^2. Il valore di R^2 abbastanza vicino a 1, come nel nostro caso, è caratteristico di un buon adattamento del modello ai dati posseduti. Infine mostriamo la rappresentazione grafica della regressione.



Si osserva che al crescere dell'età tendono a diminuire le pulsazioni al minuto, dunque i dati sono caratterizzati da una forte correlazione negativa.

3.4 Regressione lineare multipla

Un modello di regressione lineare multipla si differenzia da un modello di regressione semplice per la presenza di più di un predittore. In questo caso il modello è

$$y = \beta_1 x_1 + \ldots + \beta_k x_k + w$$

dove la variabile dipendente y è una funzione lineare dei predittori $x_1, \ldots, x_k$ più una perturbazione w. Le v.a. w_i sono indipendenti e hanno legge $N(0, \sigma^2)$ (la varianza σ^2, come prima, non dipende da i). A meno di precisare il contrario supporremo sempre che il primo predittore, x_1, assuma il valore 1, cioè

$$x_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

x_1 si chiama anche il fattore costante. Inoltre, supporremo che i vettori $x_1, \ldots, x_k$ siano linearmente indipendenti. Tale ipotesi deriva dal fatto che, se i vettori $x_1, \ldots, x_k$ fossero dipendenti ve ne sarebbe uno che si potrebbe ottenere come combinazione lineare degli altri e non sarebbe quindi possibile distinguere l'effetto di quest'ultimo da quello degli altri predittori. Si osserva che, rispetto a P^θ la v.a. w ha una legge $N(0, \sigma^2 I)$ e dunque l'osservazione y ha legge $N(\beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2 I)$. La condizione di varianza costante del termine di errore w corrisponde all'ipotesi di omoschedaticità.
Il primo problema, come per la regressione lineare semplice, consiste nella stima del parametro $\theta = (\beta_1, \ldots, \beta_k, \sigma^2)$ che varia in $\Theta = \mathbb{R}^k \times \mathbb{R}^+$. Sappiamo che al variare di $\beta_1, \ldots, \beta_k$ il vettore $\beta_1 x_1 + \ldots + \beta_k x_k$ descrive l'iperpiano di $\mathbb{R}^n$ (n al solito è il numero delle osservazioni) generato dai vettori $x_1, \ldots, x_k$. Possiamo dunque cercare i valori di $\beta_1, \ldots, \beta_k$ in corrispondenza dei quali la distanza $|y - \beta_1 x_1 - \ldots - \beta_k x_k|$ sia minima, il che equivale a calcolare la proiezione ortogonale di y sull'iperpiano E generato da $x_1, \ldots, x_k$. Se indichiamo con X la matrice $n \times k$ di cui i vettori $x_1, \ldots, x_k$ sono le colonne e poniamo

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

allora ogni vettore appartenente all'iperpiano E generato da $x_1, \ldots, x_k$ si può scrivere nella forma $X\beta$ al variare di $\beta \in \mathbb{R}^k$. Riprendendo l'Esempio A.2 il valore di β per cui $X\beta$ è la proiezione di y su E è dato da

$$b = (X^* X)^{-1} X^* y$$

Si può dimostrare che b è uno stimatore non distorto di β. Intanto, poiché $y \sim N(X\beta, \sigma^2 I)$ e b è una funzione lineare di y, anche b segue una legge normale. Sapendo che $E(y) = X\beta$,

la media di b vale

$$E(b) = (X^*X)^{-1}X^*E(y) = (X^*X)^{-1}X^*X\beta = \beta$$

Per calcolare la matrice di covarianza di b occorre ricordare che se C_Y è la matrice di covarianza di una v.a. Y, allora la matrice di covarianza di AY è AC_YA^*. Quindi la matrice di covarianza di b è (in questo caso $A = (X^*X)^{-1}X^*$))

$$(X^*X)^{-1}X^*\sigma^2 I\big((X^*X)^{-1}X^*\big)^* = \sigma^2(X^*X)^{-1}X^*X(X^*X)^{-1} = \sigma^2(X^*X)^{-1}$$

(abbiamo sfruttato il fatto che $(AB)^* = B^*A^*$ e che $(X^*X)^{-1}$ è una matrice simmetrica). Si può concludere che lo stimatore b segue una legge $N(\beta, \sigma^2(X^*X)^{-1})$ e le sue componenti b_i hanno legge $N(\beta_i, \sigma^2 m_{ii})$, dove con m_{ij} indichiamo l'elemento di posto ij della matrice $(X^*X)^{-1}$.
Come per la regressione semplice indichiamo con $\hat{y} = Xb = X(X^*X)^{-1}X^*y$ il vettore dei valori stimati, ovvero la proiezione di y sul sottospazio E. Definiamo il vettore dei residui

$$r = y - \hat{y} = (I - P_E)y$$

Resta da stimare il parametro σ^2. Per fare ciò ci serviremo di un lemma.

Lemma 3.4.1. *Le v.a. n-dimensionali r e $\hat{y}$ sono indipendenti. Inoltre posto*

$$s^2 = \frac{1}{n-k}|r|^2 = \frac{1}{n-k}\sum_{i=1}^{n} r_i^2$$

allora si ha

$$\frac{s^2}{\sigma^2}(n-k) \sim \chi^2(n-k)$$

Dimostrazione. È una semplice applicazione del Teorema di Cochran: poiché $w \sim N(0, \sigma^2 I)$, possiamo supporre che sia $w = \sigma W$, dove $W \sim N(0,1)$. Per ipotesi $y = X\beta + \sigma W$ e dunque, siccome $P_E X\beta = X\beta$ ($X\beta$ è un vettore di E) mentre $P_{E^\perp} X\beta = 0$

$$\hat{y} = P_E(X\beta + \sigma W) = X\beta + \sigma P_E W$$
$$r = P_{E^\perp}(X\beta + \sigma W) = \sigma P_{E^\perp} W$$

Per il Teorema di Cochran le v.a. $P_E W$ e $P_{E^\perp} W$ sono indipendenti, allora lo stesso è vero per $\hat{y}$ e r. Inoltre, sempre per il Teorema 2.3.1, poiché $dim E^\perp = n - dim E = n - k$ la v.a. $|P_{E^\perp} W|^2$ ha legge $\chi^2(n-k)$; basta ora osservare che

$$\frac{s^2}{\sigma^2}(n-k) = |P_{E^\perp} W|^2 \sim \chi^2(n-k)$$

e il valore atteso di una v.a. con distribuzione $\chi^2(n-k)$ è $n-k$. □

Abbiamo dunque determinato degli stimatori non distorti dei parametri β e σ^2 con i quali si possono affrontare i vari problemi di stima e di test. Poiché infatti $b_i \sim N(\beta_i, \sigma^2 m_{ii})$ e b_i è indipendente da s^2, si ha facilmente

$$\frac{b_i - \beta_i}{s\sqrt{m_{ii}}} \sim t(n-k) \tag{3.12}$$

Dalla relazione

$$P\left\{\left|\frac{b_i - \beta_i}{s\sqrt{m_{ii}}}\right| \leq t_{1-\frac{\alpha}{2}}(n-k)\right\} = 1 - \alpha$$

si ricava che

$$I = [b_i - s\sqrt{m_{ii}} t_{1-\frac{\alpha}{2}}(n-k), b_i + s\sqrt{m_{ii}} t_{1-\frac{\alpha}{2}}(n-k)]$$

è un intervallo di fiducia per β_i di livello $1-\alpha$. Per lo stesso motivo

$$\left\{\left|\frac{b_i - z}{s\sqrt{m_{ii}}}\right| \geq t_{1-\frac{\alpha}{2}}(n-k)\right\}$$

è una regione di rigetto di livello α per il test

$$H : \beta_1 = z \text{ contro } A : \beta_i \neq z$$

Anche per la regressione multipla si può considerare il fattore

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

La quantità R^2 ha lo stesso significato che nel caso della regressione semplice, a condizione però che vi sia un fattore costante, altrimenti non si potrebbe affermare che la somma dei residui valga 0 e non vale più la relazione $0 \leq R^2 \leq 1$.

Esempio 3.3. L'obiettivo di questo esempio è porre in relazione l'aspettativa di vita (in anni) in 10 nazioni con alcune variabili come la frazione della popolazione adulta in grado di leggere e di scrivere, il tasso di crescita del PIL (prodotto interno lordo di un paese), la mortalità infantile e il recente numero di morti causate dal virus dell'HIV in un anno. I seguenti dati sono stati rilevati utilizzando il software *Mathematica 10*:

	Literacy fraction (x_1)	GDP (x_2)	Infant mortality (x_3)	HIV deaths (x_4)	Life expectancy (y)
Algeria	0.699	3%	0.02773	1000.	71.
Iran	0.77	3.5%	0.03578	4300.	74.048
Italy	0.984	-1.04%	0.00551	1900.	82.385
Spain	0.977	1.16%	0.00421	2300.	82.1
South Africa	0.864	3.06%	0.0442	350000	56.916
Perù	0.929	9.77%	0.02862	3300.	74.826
Rwanda	0.711	11.2%	0.08161	7800	64.066
Kenya	0.874	2.01%	0.0547	150000	61.716
United States	0.99	1.1%	0.00626	22000	78.941
Brazil	0.886	5.23%	0.023	15000	73.937

	Coef	DevSt	TStat	p
b_0	66.0539	9.04332	7.30417	0.000753286
b_1	16.0955	9.35683	1.72019	0.146026
b_2	-9.82462	26.5765	-0.369673	0.726767
b_3	-160.527	57.8378	-2.77546	0.0391114
b_4	-0.0000476822	7.69462×10^{-6}	-6.19683	0.00159707
$s = 4.02207$	$R^2 = 0.970038$			

Nella prima colonna della seconda tabella vengono elencati gli stimatori e nella seconda le deviazioni standard. La terza colonna contiene le statistiche della (3.12) calcolate ponendo $\beta_i = 0$, con $i = 0, \ldots, 4$ rispettivamente. Se confrontiamo i valori di questa co-

lonna con l'opportuno quantile della Tabella B.1 notiamo che l'osservazione non si trova nella regione critica, quindi i predittori x_1 e x_2 non sono rilevanti. In base ai valori di p (livello di significatività del test) per b_1 e b_2 anche in questo caso si può concludere che le ipotesi $\beta_1 = 0$ e $\beta_2 = 0$ non possono essere respinte al livello di α stabilito ($\alpha = 0.05$). Questa conclusione risulta sensata poiché paesi ricchi come l'Italia possono essere interessati da un abbassamento del PIL, ma possiedono ugualmente un'elevata speranza di vita, mentre paesi emergenti come il Brasile hanno un'aspettativa di vita piuttosto bassa. Infine nell'ultima riga il valore di R^2 abbastanza vicino a 1 indica un buon adattamento del modello ai dati.

Appendice A

Richiami di algebra lineare

Su $\mathbb{R}^m$ è definito il prodotto scalare

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

Due vettori $x, y \in \mathbb{R}^m$ si dicono *ortogonali* se $\langle x, y \rangle = 0$.
Due sottospazi vettoriali E e F si dicono ortogonali se ogni vettore di E è ortogonale a ogni vettore di F. Se E è un sottospazio vettoriale di $\mathbb{R}^m$ si indica con $E^\perp$ il suo ortogonale, cioè l'insieme di tutti i vettori x di $\mathbb{R}^m$ tali che $\langle x, z \rangle = 0$ per ogni $z \in E$. $E^\perp$ è anch'esso un sottospazio vettoriale di $\mathbb{R}^m$ ed ha dimensione $m - k$, se k è la dimensione di E. Inoltre ogni $x \in \mathbb{R}^m$ si può scrivere in maniera unica nella forma $x = x_1 + x_2$, dove $x_1 \in E$, $x_2 \in E^\perp$.
Indichiamo con P_E il *proiettore ortogonale* su E, cioè l'applicazione $P_E : x \to x_1$ che ad ogni $x \in \mathbb{R}^m$ associa la sua componente su E. P_E è un operatore lineare.
Sono immediate le relazioni

$$P_E P_E = P_E \tag{A.1}$$

$$I - P_E = P_{E^\perp} \tag{A.2}$$

La (A.1) segue dal fatto che $P_E x = x$ se $x \in E$. La (A.2) è particolarmente utile perché permette di calcolare immediatamente $P_{E^\perp}$ a partire da P_E. Inoltre si ha che:

Lemma A.0.2. *$P_E x$ è il vettore di E che si trova a distanza minima da x.*

Dimostrazione. Se y è un generico vettore in E e $x = x_1 + x_2$ con $x_1 \in E$, $x_2 \in E^\perp$ allora

$$\begin{aligned} |y-x|^2 = |(y-x_1)-x_2|^2 = \langle (y-x_1)-x_2, (y-x_1)-x_2\rangle = \\ = |y-x_1|^2 - 2\underbrace{\langle y-x_1, x_2\rangle}_{0} + |x_2|^2 = |y-x_1|^2 + |x_2|^2 \end{aligned} \tag{A.3}$$

($\langle y-x_1, x_2\rangle = 0$ perché $y - x_1 \in E$ mentre $x_2 \in E^\perp$). La quantità $|y-x|^2$ è dunque sempre $\geq |x_2|^2$ ed è esattamente uguale a $|x_2|^2$ se e solo se $y = x_1 = P_E x$. □

Esempio A.1. Sia E il sottospazio (di dimensione 1) di $\mathbb{R}^m$ dei vettori aventi tutte le componenti uguali. Per il Lemma A.0.2 $P_E x$ è il vettore $z \in E$ per cui la quantità $|x-z|^2$ è minima. Poiché z è della forma $z = (t, t, \ldots, t)$ si tratta di determinare il punto di minimo di

$$\psi : t \to |x-z|^2 = \sum_{i=1}^{m} (x_i - t)^2$$

Derivando e calcolando i punti critici si ottiene

$$\psi\prime(t) = -2\sum_{i=1}^{m}(x_i - t)$$

$$0 = \sum_{i=1}^{m}(x_i - t) = \sum_{i=1}^{m} x_i - mt$$

cioè

$$t = \frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x}$$

e dunque $z = (\bar{x}, \ldots, \bar{x})$. La proiezione ortogonale di x su E è dunque il vettore le cui coordinate sono tutte uguali alla media delle coordinate di x. In questi esempi si può omettere la verifica che il punto critico è effettivamente un punto di minimo poiché si tratta di una questione facile visto che la funzione da minimizzare è quadratica.

Esempio A.2. Supponiamo, più in generale, che E sia generato dai vettori $z_1, \ldots, z_k$, $k < m$, che supporremo linearmente indipendenti. Un generico vettore $z \in E$ si scrive

$z = \theta_1 z_1 + \ldots + \theta_k z_k$. Ovvero, se Z è la matrice $m \times k$ di cui i vettori $z_1, \ldots, z_k$ sono le colonne e poniamo $\theta = (\theta_1, \ldots, \theta_k)$, un generico vettore di E si scrive $z = Z\theta$ al variare di θ in $\mathbb{R}^k$. Per il Lemma A.0.2 per calcolare P_E basta determinare il valore θ_0 che minimizzi la quantità

$$\theta \to \psi(\theta) = |x - Z\theta|^2$$

dopo di che sarà $P_E x = Z\theta_0$. Il problema è quindi ricondotto al calcolo del punto di minimo di una funzione su $\mathbb{R}^k$.

$$\begin{aligned} \psi(\theta) = |x - Z\theta|^2 = |x|^2 + |Z\theta|^2 - 2\langle x, Z\theta\rangle = |x|^2 + \langle Z\theta, Z\theta\rangle - 2\langle x, Z\theta\rangle = \\ = |x|^2 + \langle Z^* Z\theta, \theta\rangle - 2\langle Z^* x, \theta\rangle \end{aligned} \tag{A.4}$$

A meno della costante $|x|^2$ ψ è la somma della funzione lineare $\theta \to -2\langle Z^* x, \theta\rangle$ e della funzione quadratica $\theta \to \langle Z^* Z\theta, \theta\rangle$; dunque il suo gradiente vale

$$grad\psi(\theta) = 2Z^* Z\theta - 2Z^* x \tag{A.5}$$

Quindi l'equazione $grad\psi(\theta) = 0$ si risolve facilmente se la matrice $Z^* Z$ è invertibile. Mostriamo che, nelle ipotesi fatte, ciò è vero, cioè che se i vettori $z_1, \ldots, z_k$ sono linearmente indipendenti, allora la matrice $Z^* Z$ (che è $k \times k$) è invertibile. Infatti se $\theta \in \mathbb{R}^k$ fosse un vettore tale che $Z^* Z\theta = 0$, a maggior ragione si avrebbe

$$0 = \langle Z^* Z\theta, \theta\rangle = \langle Z\theta, Z\theta\rangle = |Z\theta|^2$$

Ma ciò è possibile solo se $Z\theta = \theta_1 z_1 + \ldots + \theta_k z_k = 0$ e, poiché supponiamo che i vettori $z_1, \ldots, z_k$ siano linearmente indipendenti, ciò implicherebbe $\theta_1 = \ldots = \theta_k = 0$. Dunque $Z^* Z\theta = 0$ se e solo se $\theta = 0$ e quindi $Z^* Z$ è invertibile. Riprendendo (A.5) vediamo che $grad\psi$ si annulla per $\theta = \theta_0$, dove

$$\theta_0 = (Z^* Z)^{-1} Z^* x$$

Quindi

$$P_E x = Z\theta_0 = Z(Z^*Z)^{-1}Z^*x \tag{A.6}$$

La (A.6) riconduce il calcolo di P_E a quello di un prodotto di matrici, che può essere effettuato numericamente in maniera abbastanza semplice da un calcolatore.

Appendice B

Quantili delle leggi $t(n)$ di Student

n	0.95	0.975
1	6.31375	12.7062
2	2.91999	4.3027
3	2.35336	3.1824
4	2.13187	2.7764
5	2.01505	2.5706
6	1.94318	2.4469
7	1.89459	2.3646
8	1.85955	2.3060
9	1.83311	2.2622
10	1.81246	2.2281
11	1.79589	2.2010
12	1.78229	2.1788
13	1.77093	2.1604
14	1.76131	2.1448
15	1.75305	2.1315
16	1.74589	2.1199
17	1.73961	2.1098
18	1.73407	2.1009
19	1.72914	2.0930
20	1.72473	2.0860

Tabella B.1: Quantili delle leggi $t(n)$ di Student

Nella tabella riportata sopra troviamo i quantili $(t_\alpha(n))$ delle leggi t di Student con

n gradi di libertà utilizzati più frequentemente, ossia quelli corrispondenti a $\alpha = 0.95$ e $\alpha = 0.975$.

Bibliografia

[1] Paolo Baldi, *Calcolo delle probabilità e statistica*, McGraw-Hill, 1998

[2] Simone Borra, Agostino Di Ciaccio, *STATISTICA: metodologie per le scienze economiche e sociali*, McGraw-Hill, 2008

[3] A. Pascucci, *Note delle Lezioni, dispense del corso 'Probabilità e Statistica 1'*, Bologna, 2014