Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

La Statistica Inferenziale e alcune applicazioni classiche

Relatore:
Chiar.mo Prof.
PAOLO NEGRINI

Presentata da:
GIADA MORETTI

Sessione Unica
2017/18

Indice

Introduzione

Nella presente tesi ho approfondito lo studio di alcuni argomenti di Statistica inferenziale che non ho avuto occasione di studiare nel mio curriculum universitario, tra cui la stima di parametri e i test di ipotesi statistiche.
Nella prima parte riassumo una serie di nozioni necessarie alla comprensione del lavoro, in particolare analizzo la modellizzazione di un evento aleatorio di cui si conosce a priori la funzione di probabilità.
Successivamente, espongo il metodo con cui si può ricavare una stima dei parametri associati alla probabilità di un evento aleatorio di cui si sono solo osservati i risultati. In seguito, definisco le leggi gamma, di chi-quadro e t di Student che sono necessarie per stabilire un intervallo di confidenza per media e varianza di campioni gaussiani e per la probabilità di successo in una sequenza di prove indipendenti.
Nel terzo capitolo introduco i test di ipotesi statistiche che vengono utilizzati per confrontare dati sperimentali fra di loro o con valori assegnati: in particolare prendo in considerazione prove bernoulliane, media e varianza di campioni gaussiani e distribuzioni di cui si vuole stabilire l'indipendenza.
Nel quarto capitolo analizzo due applicazioni storiche dei metodi esposti nella parte iniziale della tesi. La prima tratta della famosa disputa fra Fisher e Mendel sui dati raccolti dal secondo durante i suoi esperimenti di genetica coi piselli odorosi: secondo Fisher infatti, nonostante l'esposizione chiara, ordinata e razionale dei dati raccolti, Mendel ha falsificato o omesso parte dei risultati al fine di avere un riscontro sperimentale delle sue teorie sulla ereditarietà dei caratteri. Esamino approfonditamente l'insieme di risultati che più aveva convinto Fisher della fondatezza della sua idea, svolgendo il test del chi-quadro sia con la proporzione teorica proposta da Mendel che con quella suggerita da Fisher. Da questa disputa, si è sviluppata una serie di lavori che, cercando di demolire o di sostenere il lavoro di Mendel, hanno portato alla nascita di nuovi modelli biologici, genetici e statistici.
La seconda applicazione storica tratta dei risultati ottenuti da Gosset nel campo

della statistica applicata: mentre è alla ricerca della miglior varietà d'orzo da coltivare nei campi del Regno Unito, sviluppa alcuni metodi statistici usati ancora oggi. I suoi esperimenti sono volti alla ricerca del metodo più economicamente vantaggioso per produrre birra: si impegna così non solo a studiare una funzione che gli permetta di valutare campioni limitati ma anche a sviluppare un metodo di semina dei campi che consenta di minimizzare le variazioni nel raccolto delle diverse varietà studiate dovute a cause esterne.

Capitolo 1

Nozioni preliminari

Nei problemi di statistica inferenziale, considereremo quantità che, essendo il risultato di misurazioni di eventi aleatori, dipendono da fenomeni casuali: introduciamo quindi in questo capitolo le nozioni di base necessarie e alcuni risultati che verranno usati nei prossimi capitoli.

Esempio. Consideriamo un dado a sei facce, non bilanciato. Vogliamo sapere, prima di lanciare il dado, qual è la probabilità dei vari eventi che possono succedere come, ad esempio, la probabilità che esca un numero pari, che esca il numero 6 o che esca il numero 12. Dato che i vari risultati del lancio del dado sono equiprobabili, è logico pensare che la probabilità che esca il numero 6 sia $\frac{1}{6}$, la probabilità che esca un numero pari sia $\frac{1}{2}$ e la probabilità che esca 12 sia 0.

I risultati dell'esempio sono intuitivi: vogliamo però poter costruire un modello simile per un qualsiasi fenomeno che sappiamo, a priori, come si comporterà.

Definizione 1.1. Chiamiamo *spazio degli eventi* un insieme Ω non vuoto che rappresenta l'insieme di tutti i possibili risultati di un fenomeno aleatorio. Un elemento $\omega \in \Omega$ si dice *evento elementare* mentre un sottoinsieme $E \subseteq \Omega$ si chiama *evento*.

Definizione 1.2. Si chiama σ-*algebra* una famiglia $\mathscr{F}$ di sottoinsiemi di Ω per cui valgono le seguenti proprietà:

- $\Omega \in \mathscr{F}$,
- se $\{A_i\}_{i\in\mathbb{N}} \in \mathscr{F}$ allora $\bigcup_{i\in\mathbb{N}} A_i \in \mathscr{F}$,
- se $A \in \mathscr{F}$ allora $A^C \in \mathscr{F}$.

Definizione 1.3. Una coppia ordinata $(\Omega, \mathscr{F})$ con $\mathscr{F}$ σ-algebra di sottoinsiemi di Ω si chiama *spazio misurabile*.

Definizione 1.4. Sia $(\Omega, \mathscr{F})$ uno spazio misurabile. Una funzione

$$\mu : \mathscr{F} \to [0; +\infty]$$

si dice *misura* su Ω se

- $\mu(\emptyset) = 0$,
- se $\{A_i\}_{i\in\mathbb{N}} \in \mathscr{F}$ con $A_i \cap A_j = \emptyset$ se $i \neq j$ allora

$$\mu(\bigcup_{i\in\mathbb{N}} A_i) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

La terna $(\Omega, \mathscr{F}, \mu)$ si chiama *spazio di misura*.
Se una misura μ è tale che $\mu(\Omega) = 1$ allora si dice che μ è una *probabilità* e che la terna $(\Omega, \mathscr{F}, \mu)$ è uno *spazio di probabilità*.

Esempio. Riprendiamo l'esempio del dado a sei facce non bilanciato: uno spazio degli eventi può essere $\Omega = \{1, 2, 3, 4, 5, 6\}$. Ω ha cardinalità finita quindi una possibile σ-algebra è data dall'insieme delle parti $\mathcal{P}(\Omega)$. Ora, sappiamo che il dado è bilanciato quindi tutti i risultati hanno la stessa probabilità; inoltre la probabilità che cerchiamo deve soddisfare $P(\Omega) = 1$ quindi

$$1 = P(\Omega) = \sum_{\omega\in\Omega} P(\{\omega\}) = p \cdot \#\Omega$$

dove p è la probabilità $P(\{\omega\})$. Abbiamo quindi che $p = \frac{1}{\#\Omega}$. Inoltre possiamo vedere ogni $A \in \mathcal{P}(\Omega)$ come l'unione disgiunta degli eventi elementari $\{\omega_1\}$,$\{\omega_2\}$,... per gli $\omega_i \in A$: sarà allora

$$P(A) = \frac{\#A}{\#\Omega}$$

definita per ogni $A \subseteq \Omega$. Abbiamo così descritto il comportamento di P per ogni sottoinsieme A di Ω.
Osserviamo che il modello usato non è unico ma è quello che naturalmente viene da associare a questo genere di fenomeni aleatori in cui tutti gli eventi elementari sono equiprobabili.

1.1 Variabili aleatorie

Definizione 1.5. Siano $(\Omega, \mathscr{F})$ e $(E, \mathscr{E})$ due spazi di misura. Una funzione

$$X : \Omega \to E$$

si dice *misurabile* se per ogni $A \in \mathscr{E}$ si ha $X^{-1}(A) \in \mathscr{F}$.
Inoltre, se lo spazio $(\Omega, \mathscr{F})$ è uno spazio di probabilità, allora X si dice *variabile aleatoria*.

Spesso ci interesseranno variabili aleatorie reali quindi useremo, equivalentemente, la seguente definizione:

Definizione 1.6. Sia $(\Omega, \mathcal{A}, P)$ uno spazio di probabilità. Diremo che un'applicazione

$$X : \Omega \to \mathbb{R}$$

è una *variabile aleatoria* se per ogni $t \in \mathbb{R}$ l'insieme $\{\omega : X(\omega) \leq t\}$ è in $\mathcal{A}$.

Definizione 1.7. Sia X una variabile aleatoria: chiamiamo *legge* o *distribuzione* di X la funzione

$$A \to P(\{\omega : X(\omega) \in A\})$$

dove $A \subseteq \mathbb{R}$.

Per indicare che una variabile aleatoria X segue una certa legge P_X si scrive $X \sim P_X$.
Per comodità di notazione, indicheremo l'insieme $\{\omega : X(\omega) \in A\}$ con $\{X \in A\}$: non sempre questo tipo di insiemi è un evento. Sappiamo però che, per definizione di variabile aleatoria, $\{X \leq t\}$ è sempre un evento: inoltre $\{X > t\} = \{X \leq t\}^C$ è un evento; anche

$$\{t_1 < X \leq t_2\} = \{X \leq t_2\} \cap \{X > t_1\}$$

e

$$\{X = t\} = \cap_{n \in \mathbb{N}} \left\{ t - \frac{1}{n} < X \leq t \right\}$$

sono eventi.

Definizione 1.8. Una variabile aleatoria X che assume un numero finito o infinito ma numerabile di valori si dice *discreta* mentre se una variabile aleatoria prende valori in un intervallo di $\mathbb{R}$ o, eventualmente, tutto $\mathbb{R}$ si dice *continua*.

Definizione 1.9. Si chiama *funzione di ripartizione* la funzione

$$\begin{array}{rccc} F: & \mathbb{R} & \to & [0,1] \\ & t & \mapsto & F(t) = P(\{X \leq t\}) \end{array}$$

Le funzioni di ripartizione sono funzioni non decrescenti in quanto, all'aumentare di t, l'evento $\{X \leq\}$ aumenta. Inoltre $0 \leq F(t) \leq 1$. Per le variabili aleatorie continue, la funzione di ripartizione è anche continua da destra e tale che $\lim_{t\to-\infty} F(t) = 0$ e $\lim_{t\to+\infty} F(t) = 1$.

Definizione 1.10. Si chiama *densità discreta* di una variabile aleatoria discreta X la funzione

$$\begin{array}{rccc} p: & \mathbb{R} & \to & \mathbb{R}^+ \\ & x & \mapsto & p(x) = P(\{X = x)\} \end{array}$$

Possiamo così determinare la legge di X a partire dalla densità discreta: l'evento $\{X \in A\}$ sarà l'unione di tutti gli eventi $\{X = x_i\}$ per cui $x_i \in A$ e, quindi,

$$P(\{X \in A\}) = \sum_{x_i \in A} P(\{X = x_i\}) = \sum_{x_i \in A} p(x_i)$$

dove la somma è una somma finita se l'insieme dei valori assunti da X è finito o una serie se ha cardinalità numerabile.

Definizione 1.11. Sia X una variabile aleatoria continua e F la sua funzione di ripartizione: una funzione f integrabile su $\mathbb{R}$ e tale che $f \geq 0$ si dice *densità* di X se

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

Definizione 1.12. Sia $X = (X_1, \ldots, X_n)$ un vettore in cui X_i sono variabili aleatorie reali: X si chiama *variabile aleatoria n-dimensionale.*

Per comodità di notazione, nel seguito assumiamo che $n = 2$ e che $Z = (X, Y)$. Sappiamo già che gli insiemi $\{X \leq x\}$ e $\{Y \leq y\}$ sono eventi per definizione di variabile aleatoria reale: anche $\{X \leq x, Y \leq y\}$ è un evento come intersezione di $\{X \leq x\}$ e $\{Y \leq y\}$: scriveremo $\{X \leq x, Y \leq y\} = \{Z \in A_{x,y}\}$ dove $A_{x,y} = \{(u, v) : u \leq x, v \leq y\}$.

Definizione 1.13. Si chiama *funzione di ripartizione congiunta* di X e Y la funzione

$$F(x, y) = P(\{X \leq x, Y \leq y\}) = P(\{Z \in A_{x,y}\}).$$

Definizione 1.14. Si dice che X e Y hanno *densità congiunta* f se vale

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\mathrm{d}u = \int_{A_{x,y}} f(u,v)\,\mathrm{d}v\mathrm{d}u.$$

Proposizione 1.1.1. *Siano X e Y due variabili aleatorie che hanno densità congiunta f. Allora $X+Y$ ha densità*

$$g(x) = \int_{-\infty}^{+\infty} f(x, z-x)\,\mathrm{d}x.$$

Dimostrazione. Consideriamo la variabile aleatoria 2-dimensionale $Z = (X,Y)$: osserviamo che $X+Y = \phi(Z)$ dove $\phi((x,y)) = x+y$. Vorremmo poter calcolare la funzione di ripartizione di $\phi(Z)$:

$$\begin{aligned} G(t) &= P(\{\phi(Z) \le t\}) = P(\{\phi(Z) \in]-\infty, t]\}) = \\ &= P\left(\left\{X \in \phi^{-1}(]-\infty,t])\right\}\right) = \int_{\phi^{-1}(]-\infty,t])} f(x)\,\mathrm{d}x. \end{aligned}$$

I punti (x,y) in $\phi^{-1}(]-\infty,t])$ sono quelli per cui $x+y \le t$: allora abbiamo

$$G(t) = \int_{\phi^{-1}(]-\infty,t])} f(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty}\int_{-\infty}^{t-x} f(x,y)\,\mathrm{d}y\mathrm{d}x.$$

Usiamo il cambio di variabile $z = y+x$ e otteniamo

$$\int_{-\infty}^{t-x} f(x,y)\,\mathrm{d}y = \int_{-\infty}^{t} f(x, z-x)\,\mathrm{d}z.$$

Allora

$$\begin{aligned} G(t) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{t-x} f(x,y)\,\mathrm{d}y\mathrm{d}x = \int_{-\infty}^{+\infty}\int_{-\infty}^{t} f(x,z-x)\,\mathrm{d}z\mathrm{d}x = \\ &= \int_{-\infty}^{t}\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\mathrm{d}z. \end{aligned}$$

Poniamo quindi $g(z) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x$: otteniamo per la funzione di ripartizione di $X+Y$ l'espressione

$$G(t) = \int_{-\infty}^{t} g(z)\,\mathrm{d}z$$

cioè g è la densità che cerchiamo. □

Definizione 1.15. Siano $X_1, \ldots, X_n$ n variabili aleatorie definite sullo stesso spazio di probabilità: diciamo che sono *indipendenti* se per ogni $A_1, \ldots, A_n \subseteq \mathbb{R}$ si ha

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(\{X_1 \in A_1\}) \cdot \ldots \cdot P(\{X_n \in A_n\}).$$

Siano, invece, $X_1, \ldots, X_n, \ldots$ un numero infinito di variabili aleatorie definite sullo stesso spazio: si dicono *indipendenti* se, per ogni $m > 0$, $X_1, \ldots, X_m$ sono indipendenti.

Definizione 1.16. Sia X una variabile aleatoria discreta che assume valori x_k con probabilità $p_X(x_k)$. Allora si chiama *media* di X il valore $\mu = E[X] = \sum x_k p_X(x_k)$.
Sia, invece, X una variabile aleatoria continua con densità f. Allora si chiama *media* di X il valore $\mu = E[X] = \int_{-\infty}^{+\infty} x f(x)\, \mathrm{d}x$.

Proposizione. *Siano X e Y variabili aleatorie e $c \in \mathbb{R}$ allora valgono*

- $E[cX] = cE[X]$;
- $E[X+Y] = E[X] + E[Y]$;
- *se X e Y sono indipendenti allora* $E[XY] = E[X]E[Y]$.

Definizione 1.17. Si chiama *varianza* della variabile aleatoria X, discreta o continua, la quantità $\sigma^2 = Var[X] = E\left[(X - E[X])^2\right]$.

Proposizione. *Sia X una variabile aleatoria e $a \in \mathbb{R}$ allora valgono*

- $\sigma^2 = E[X^2] - E[X]^2$;
- $Var[aX] = a^2 Var[X]$;
- $Var(a + X) = Var(X)$.

Proposizione 1.1.2. ***Disuguaglianza di Chebyshev*** *Sia X una variabile aleatoria, discreta o continua, e $\delta > 0$, allora*

$$P(|X - E[X]| > \delta) \leq \frac{Var[X]}{\delta^2}.$$

Definizione 1.18. Siano X e Y due variabili aleatorie. Chiamiamo *covarianza* delle due variabili aleatorie la quantità

$$Cov(X, Y) = E[XY] - E[X]E[Y].$$

Osserviamo che se X e Y sono indipendenti allora $Cov(X, Y) = 0$.

1.2 Alcune leggi

Definizione 1.19. Si chiama *legge di Bernoulli* di parametro p la legge di una variabile aleatoria X tale che

$$X = \begin{cases} 1 & \text{con probabilità } p \\ 0 & \text{con probabilità } 1-p \end{cases}$$

dove $0 \leq p \leq 1$. Si scrive $X \sim B(1,p)$.

Definizione 1.20. Si chiama *legge binomiale* la legge di una variabile aleatoria X che assume i valori $k = 0, 1, 2, \ldots, n$ con probabilità

$$p_k = P(\{X = k\}) = \binom{n}{k} p_k (1-p)^{n-k}$$

dove $0 \leq p \leq 1$. Scriveremo $X \sim B(n,p)$.

Osserviamo che, come evidenziato dalla notazione, la legge di Bernoulli non è altro che una legge binomiale con $n = 1$.
Inoltre, è facile vedere che $E[B(n,p)] = np$ e $Var[B(n,p)] = np(1-p)$.

Proposizione 1.2.1. *Siano $X_1, \ldots, X_n$ n variabili aleatorie tali che $X_i \sim B(1,p)$ per ogni i con $0 \leq p \leq 1$. Allora $X = X_1 + \ldots + X_n$ sarà distribuita come $B(n,p)$.*

Definizione 1.21. Si chiama *legge di Poisson* di parametro $\lambda > 0$ la legge di una variabile aleatoria X che assume i valori $k \in \mathbb{N}$ con probabilità

$$p_k = P(\{X = k\}) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Possiamo vedere che $E[P(\lambda)] = Var[P(\lambda)] = \lambda$.

Definizione 1.22. Si chiama *legge normale* o *legge gaussiana* la legge di una variabile aleatoria X la cui densità è data da

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

con $\mu \in \mathbb{R}$ e $\sigma > 0$. Scriveremo $X \sim N(\mu, \sigma^2)$; inoltre la sua funzione di ripartizione è

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, \mathrm{d}t:$$

questo integrale non ha un'espressione analitica elementare.

Nel caso in cui $X \sim N(0,1)$, si dice che la legge è normale standard e si è soliti usare le seguenti notazioni per densità e funzione di ripartizione:

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

$$\Phi(x) = F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}t.$$

Proposizione. *Sia X una variabile aleatoria reale con densità f e $a, b \in \mathbb{R}$: allora la densità di $aX + b$ è*

$$g(t) = \frac{1}{|a|} f\left(\frac{t-b}{a}\right).$$

Quindi, se $\mu, \sigma \in \mathbb{R}$ e $X \sim N(0,1)$ allora $Y = \sigma X + \mu$ ha densità

$$g(y) = \frac{1}{|\sigma|} f\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Abbiamo quindi una relazione fra $X \sim N(0,1)$ e $Y \sim N(\mu, \sigma^2)$ che ci permetterà di svolgere i calcoli con leggi normali standard e poi ricollegarci a una legge normale qualsiasi.

Proposizione. *Sia $X \sim N(\mu, \sigma^2)$ allora valgono:*

- *se $Y \sim N(\nu, \tau^2)$ e X e Y sono indipendenti allora $X+Y \sim N(\mu+\nu, \sigma^2+\tau^2)$;*
- *se $a, b \in \mathbb{R}$ allora $P(\{a \leq X \leq b\}) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$.*

Inoltre, possiamo vedere che $E[N(\mu, \sigma^2)] = \mu$ e $Var[N(\mu, \sigma^2)] = \sigma^2$.

1.3 Legge dei grandi numeri e Teorema del limite centrale

Definizione 1.23. Sia $(X_n)_n$ una successione di variabili aleatorie. Diciamo che la successione converge alla variabile aleatoria X in probabilità se fissato $\delta > 0$ allora

$$\lim_{x \to +\infty} P(|X_n - X| > \delta) = 0.$$

Scriveremo $X_n \xrightarrow{\mathrm{P}} X$.

Teorema 1.3.1. ***Legge dei grandi numeri.*** *Sia $(X_n)_n$ una successione di variabili aleatorie indipendenti con la stessa legge. Supponiamo che $E[X_i] = \mu$ e $Var[X_i] = \sigma^2$ per ogni i. Definiamo $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$: allora*

$$\bar{X} \to \mu$$

per $n \to +\infty$.

Dimostrazione. Innanzitutto vediamo che anche $\bar{X}$ ha media μ: infatti

$$E\left[\bar{X}\right] = \frac{1}{n}E[X_1 + \ldots + X_n] = \frac{1}{n}(E[X_1] + \ldots + E[X_n]) = \mu.$$

Inoltre la sua varianza vale

$$\begin{aligned} Var\left[\bar{X}\right] &= \frac{1}{n^2}Var[X_1 + \ldots + X_n] = \frac{1}{n^2} \cdot \cdot Var[X_1] = \\ &= \frac{1}{n^2}(Var[X_1] + \ldots + Var[X_n]) = \frac{\sigma^2}{n}. \end{aligned}$$

Possiamo applicare la disuguaglianza di Chebyshev 1.1.2 e otteniamo

$$P\left(|\bar{X} - \mu| > \eta\right) \leq \frac{Var\left[\bar{X}\right]}{\eta^2} = \frac{\sigma^2}{n\eta^2} \xrightarrow[n\to+\infty]{} 0.$$

□

Definizione 1.24. Siano $X_1, X_2, \ldots$ variabili aleatorie reali e $F_1, F_2, \ldots$ le loro funzioni di ripartizione; sia inoltre X un'ulteriore variabile aleatoria e F la sua funzione di ripartizione. Si dice che la successione $(X_n)_n$ *converge in legge* a X se e solo se

$$\lim_{n\to+\infty} F_n(x) = F(x)$$

per ogni $x \in \mathbb{R}$ in cui F è continua. Scriveremo $X_n \xrightarrow{\mathscr{L}} X$.

Teorema 1.3.2. ***Teorema del limite centrale.*** *Sia $(X_n)_n$ una successione di variabili aleatorie indipendenti, con la stessa legge, con media μ e varianza $\sigma^2 > 0$. Sia inoltre*

$$S_n = X_1 + \ldots + X_n \text{ e } S_n' = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Allora

$$\lim_{n\to+\infty} P\left(S_n' \leq t\right) = \Phi(t).$$

Il teorema del limite centrale ci permette di approssimare la legge di una variabile aleatoria che si può scrivere come

$$S_n' = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

con una legge $N(0,1)$ qualunque siano le leggi delle X_i a patto che n sia abbastanza grande: non esistono risultati teorici sul valore che n deve assumere ma ci si basa solo sull'esperienza empirica. Si potrebbe, anzi, vedere che per qualsiasi n esistono delle X_i che rispettano le condizioni del teorema del limite centrale ma per cui la legge di S_n' non è approssimabile con quella di $N(0,1)$. Useremo in particolare la seguente approssimazione: se X_i soddisfano le condizioni del teorema del limite centrale, allora

$$P\{X_1 + \ldots + X_n \leq t\} = P\left\{S_n' \leq \frac{t - n\mu}{\sigma\sqrt{n}}\right\} \simeq \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right).$$

Sia $X_1, \ldots, X_n$ una successione di variabili aleatorie che soddisfano le richieste del teorema del limite centrale cioè indipendenti, equidistribuite, di media μ e varianza $\sigma^2 > 0$: possiamo usare l'approssimazione che abbiamo appena visto per stimare la probabilità con cui la media di queste variabili aleatorie differisce dalla media μ. Poniamo $\bar{X}_n = \frac{1}{n}(X_1, \ldots, X_n)$: possiamo allora vedere che, se $\alpha > 0$, si ha

$$\begin{aligned}
&P\left(\left\{|\bar{X}_n - \mu| > \alpha\right\}\right) = P\left(\left\{\bar{X}_n\alpha + \mu\right\} \cup \left\{\bar{X}_n < \alpha + \mu\right\}\right) = \\
&= P\left(\{X_1 + \ldots + X_n > n(\alpha + \mu)\}\right) + P\left(\{X_1 + \ldots + X_n < n(\alpha + \mu)\}\right) \simeq \\
&\simeq 1 - \Phi\left(\frac{n(\alpha + \mu) - n\mu}{\sigma\sqrt{n}}\right) + \Phi\left(\frac{n(\alpha + \mu) - n\mu}{\sigma\sqrt{n}}\right) = 2\Phi\left(-\frac{\alpha}{\sigma}\sqrt{n}\right)
\end{aligned}$$

Capitolo 2

Stima di parametri

Esempio 2.1. Consideriamo un'urna contenente palline bianche e palline nere per un totale di 10 palline, non riconoscibili al tatto e in una proporzione non nota. Si estrae una pallina, si segna il risultato e si reinserisce la pallina nell'urna: questa operazione viene ripetuta 1000 volte, ottenendo 302 volte una pallina di colore bianco. Non possiamo costruire uno spazio di probabilità come avevamo fatto nel primo capitolo in quanto non sappiamo la proporzione fra palline bianche e palline nere e manca, quindi, la probabilità p di estrarre palline bianche.

Il metodo di analisi di un fenomeno casuale esposto nel primo capitolo può essere utilizzato solo nel caso in cui si conoscano a priori i parametri che lo caratterizzano; non possiamo costruire uno spazio di probabilità adeguato a descrivere un evento aleatorio nel caso in cui, invece, siano noti solo i risultati. Dovremo quindi usare un metodo alternativo: ricavare informazioni generali a partire da un numero grande ma limitato di dati sperimentali è lo scopo della statistica inferenziale. Vorremmo costruire comunque uno spazio di probabilità che descriva le osservazioni dell'esperimento aleatorio: dovremo ricorrere a un modello simile a quello del primo capitolo in cui la probabilità P però non è nota. Un modello conveniente per questo tipo di problemi è il seguente:

Definizione 2.1. Si chiama *modello statistico* una famiglia di spazi di probabilità $(\Omega, \mathcal{A}, \left(P^\theta\right)_{\theta\in\Theta})$ dove Θ è un opportuno insieme.

Il parametro θ incognito, a seconda delle osservazioni prese in esame, può essere sia scalare che vettoriale e sarà l'oggetto della nostra analisi: vorremo stimare un suo valore oppure una funzione $\Psi(\theta)$ che dipende da esso.

Esempio. Nell'esempio 2.1, il parametro che vogliamo stimare sarà $\theta = p$ dove p è la probabilità con cui si estrae una pallina bianca, quindi si tratta di uno scalare.

Consideriamo invece un esperimento in cui la misura di una grandezza fisica non sia accurata in quanto è affetta da errori casuali non imputabili a errori dello strumento con cui viene effettuata: per ottenere una lettura corretta, la misura viene effettuata più volte e si ottengono valori diversi. In questo caso vorremmo poter stimare sia la media μ che la varianza σ^2 delle misure: il parametro è quindi bidimensionale $\theta = (\mu, \sigma^2)$; inoltre potremmo non essere interessati a una stima esatta della varianza quindi ci troveremmo di fronte al caso in cui vogliamo stimare solo una funzione $\Psi(\theta) = \mu$ del parametro.

Esempio. Un modello statistico adatto allo studio dell'esempio 2.1 può essere il seguente: nel momento in cui si registra il risultato dell'estrazione, si segna 1 quando la pallina estratta è bianca e 0 quando la pallina estratta è nera; si ottiene così una sequenza di 0 e di 1 che indicheremo con $\omega = (\omega_1, \ldots, \omega_{1000})$. Si ha poi

$$\Omega = \{0, 1\}^{1000}$$

$$\mathcal{A} = \text{l'insieme delle parti di } \Omega$$

$$\Theta = [0, 1]$$

$$P^\theta(\omega) = \theta^k (1 - \theta)^{1-k}$$

dove k è il numero di volte che 1 appare nella sequenza ω: P^θ è quindi la probabilità di ottenere una sequenza ω quando la probabilità di osservare 1 in un singolo lancio vale θ.

Definizione 2.2. Si chiama *osservazione* un vettore di variabili aleatorie $X = (X_1, \ldots, X_n)$ definite su Ω.
Nel caso in cui le variabili aleatorie X_i siano tra loro indipendenti per ogni θ e abbiano la stessa legge, si parla di *campione* di rango n.

Definizione 2.3. Sia $\Psi(\theta) : \Omega \to \mathbb{R}^m$ la funzione del parametro θ che vogliamo stimare. Si chiama *statistica* ogni funzione dell'osservazione $T = t(X_1, \ldots, X_n)$ dove $t : \mathbb{R}^n \to \mathbb{R}^m$ è una funzione sufficientemente regolare. Se $t(X_1, \ldots, X_n) \subseteq \Psi(\Theta)$, T si chiama *stimatore* di $\Psi(\theta)$.

Intuitivamente dare uno stimatore T significa fissare la regola che, se si osserva ω, allora si stima $\Psi(\theta)$ con la quantità $T(\omega)$. Con questa definizione, però, ogni variabile aleatoria a valori in $\mathbb{R}^m$ è uno stimatore: una prima questione quindi consiste nello stabilire dei criteri per decidere quali stimatori diano effettivamente una buona approssimazione del parametro.

Definizione 2.4. Sia $T = t(X_1, \ldots, X_n)$ uno stimatore di $\Psi(\theta)$. Diremo che T è uno *stimatore corretto* o *non distorto* se $E^{\theta}(T) = \Psi(\theta)$;

Vogliamo anche poter confrontare due diversi stimatori per capire quale è il migliore.

Definizione 2.5. Si chiama *rischio quadratico* dello stimatore X la funzione $R_X(\theta) = E^{\theta}\left[(X - \Psi(\theta))^2\right]$.

Osserviamo che se X è non distorto, allora $R_X(\theta)$ non è altro che la varianza di X.

Definizione 2.6. Siano X e Y due diversi stimatori di $\Psi(\theta)$, diremo che X è *preferibile* a Y se $R_X(\theta) \leq R_Y(\theta)$ per ogni θ e *strettamente preferibile* se esiste θ per cui $R_X(\theta) < R_Y(\theta)$. Inoltre X si dice *ammissibile* se non esistono stimatori Y che siano strettamente preferibili di X.

Esempio. Definiamo due stimatori per l'esempio 2.1: sia $X(\omega) = \frac{1}{1000}\sum_{i=1}^{1000}\omega_i$ e $Y(\omega) = \frac{\omega_1+\omega_{1000}}{2}$. Chiaramente il primo stimatore usa tutte le informazioni ricavate dalle estrazioni delle palline mentre il secondo si limita a prendere in considerazione solo la prima e ultima estrazione: per la legge dei grandi numeri sappiamo che, per $m \to \infty$, la frazione $\frac{b}{m}$ dove b indica il numero di palline bianche estratte e m il numero totale di lanci tende al valore $\frac{B}{M}$ dove B indica il numero di palline bianche nell'urna e M il numero totale di palline. Possiamo quindi supporre che lo stimatore X sia "migliore" dello stimatore Y.
Vogliamo ora però vedere se questi risultati intuitivi vengono anche confermati dalle definizioni appena date: abbiamo

$$E[X] = \frac{1}{1000}\sum_{i=1}^{1000} E[\omega_i] = \frac{1}{1000}1000\theta = \theta$$

$$E[Y] = E\left[\frac{1}{2}(\omega_1 + \omega_{1000})\right] = \frac{1}{2}E[\omega_1] + \frac{1}{2}E[\omega_{1000}] = \theta$$

quindi entrambi gli stimatori sono corretti. Dato che sono corretti, possiamo calcolare il rischio quadratico di entrambi semplicemente calcolandone la varianza:

$$R_X(\theta) = Var\left(\frac{1}{1000}\sum_{i=1}^{1000}\omega_i\right) = \frac{1}{1000^2}\sum_{i=1}^{1000} Var(\omega_i) = \frac{1}{1000}\theta(1-\theta)$$

$$R_Y(\theta) = Var\left(\frac{1}{2}(\omega_1 + \omega_{1000})\right) = \frac{1}{4}Var(\omega_1) + \frac{1}{4}Var(\omega_{1000}) = \frac{1}{2}\theta(1-\theta)$$

cioè, come avevamo pensato, lo stimatore X è preferibile allo stimatore Y.

Non possiamo ancora stabilire quanto uno stimatore sia vicino al valore effettivo di $\Psi(\theta)$: per farlo introduciamo gli intervalli di confidenza.

Definizione 2.7. Si chiama *intervallo di confidenza* per $\Psi(\theta)$ di livello α con $0 < \alpha < 1$ un'applicazione $\omega \to B_\omega$ che ad ogni $\omega \in \Omega$ fa corrispondere un intervallo $B_\omega \subseteq \mathbb{R}$ tale che

a. $\{\omega : \Psi(\theta) \in B_\omega\} \in \mathcal{A}$ per ogni $\theta \in \Theta$;

b. per ogni $\theta \in \Theta$ allora $P^\theta(\{\omega : \Psi(\theta) \in B_\omega\}) \geq 1 - \alpha$.

Possiamo allora dire che, con probabilità $1 - \alpha$, il valore di $\Psi(\theta)$ appartiene a un intervallo di $\mathbb{R}$, indipendentemente da quale sia il valore di θ.

2.1 Stimatori per media e varianza

Vogliamo ora vedere stimatori corretti per media e varianza nel caso in cui $X = (X_1, \ldots, X_n)$ sia un campione di rango n con X_i indipendenti, equidistribuite e 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 si ha

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

per qualunque $\theta \in \Theta$.
Per trovare uno stimatore corretto per la varianza $Var_\theta(X_i)$, nel caso in cui media e varianza siano finite per ogni $\theta \in \Theta$, distinguiamo due diversi casi: nel primo la media è nota e vale μ mentre nel secondo sia la media che la varianza non sono note.
Nel primo caso uno stimatore corretto di $Var_\theta(X_i)$ è $\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$: infatti

$$E^\theta\left(\bar{\sigma}^2\right) = E^\theta\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}E^\theta\left((X_i - \mu)^2\right) = Var_\theta(X_i)$$

Nel secondo caso, invece, non conosciamo μ quindi sostituiamo in $\bar{\sigma}^2$ lo stimatore di μ: avremo allora $\bar{\bar{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. Per vedere se è uno stimatore per $Var_\theta(X_i)$ calcoliamo $E^\theta(\bar{\bar{\sigma}}^2)$. Sappiamo che $\sum_{i=1}^{n} X_i = n\bar{X}$ quindi

$$\begin{aligned}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 &= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\bar{X}^2 = \\ &= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\cdot n\bar{X} + n\bar{X}^2 = \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.\end{aligned} \tag{2.1}$$

Dalla definizione di varianza, $Var_\theta(X_i) = E^\theta(X_i^2) - E^\theta(X_i)^2$ abbiamo

$$E^\theta\left(X_i^2\right) = Var_\theta(X_i) + E^\theta(X_i)^2 \tag{2.2}$$

e, analogamente, $E^\theta\left(\bar{X}^2\right) = Var_\theta\left(\bar{X}\right) + E^\theta\left(\bar{X}\right)^2$. Inoltre vale

$$\begin{aligned}Var_\theta\left(\bar{X}\right) &= Var_\theta\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2}Var_\theta(X_1 + \ldots + X_n) = \\ &= \frac{1}{n^2}nVar_\theta(X_i) = \frac{1}{n}Var_\theta(X_i)\end{aligned}$$

e sappiamo che $E^\theta\left(\bar{X}\right) = E^\theta(X_i)$. Quindi abbiamo

$$E^\theta\left(\bar{X}^2\right) = \frac{1}{n}Var_\theta(X_i) + E^\theta(X_i)^2: \tag{2.3}$$

possiamo così sostituire nell'equazione 2.1 le espressioni 2.2 e 2.3 ottenendo

$$\begin{aligned}&E^\theta\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = E^\theta\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \\ &= E^\theta\left(\sum_{i=1}^{n} X_i^2\right) - E^\theta\left(n\bar{X}^2\right) = nE^\theta\left(X_i^2\right) - nE^\theta\left(\bar{X}^2\right) = \\ &= n\left(Var_\theta(X_i) + E^\theta(X_i)^2\right) - n\left(\frac{1}{n}Var_\theta(X_i) + E^\theta(X_i)^2\right) = \\ &= nVar_\theta(X_i) + nE^\theta(X_i)^2 - Var_\theta(X_i) - nE^\theta(X_i)^2 = (n-1)\,var_\theta(X_i).\end{aligned}$$

Quindi $E^\theta\left(\bar{\bar{\sigma}}^2\right) = \frac{1}{n}E^\theta\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\right) = \frac{n-1}{n}Var_\theta\left(X_i\right)$ cioè $\bar{\bar{\sigma}}^2$ non è uno stimatore corretto per $Var_\theta\left(X_i\right)$ ma, da questo possiamo anche dedurre che

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

è uno stimatore corretto.

2.2 Le leggi gamma, chi-quadro e t di Student

Definizione 2.8. Chiamiamo *funzione gamma* la funzione

$$\begin{aligned}\Gamma : \mathbb{R}^+ &\to \mathbb{R}^+\\ \alpha &\mapsto \int_0^{+\infty} x^{\alpha-1}e^{-x}\,\mathrm{d}x.\end{aligned}$$

Osserviamo che questo integrale non è calcolabile in maniera elementare per qualsiasi valore di α ma che, integrando per parti, si può vedere che vale

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

e $\Gamma\left(1\right) = 1$. Quindi, per ogni numero naturale, $\Gamma\left(n\right) = (n-1)!$. Inoltre, usando la sostituzione $t = x^{\frac{1}{2}}$, si ottiene

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^{+\infty} e^{-t^2}\,\mathrm{d}x = \sqrt{\pi}$$

Definizione 2.9. Si dice *legge gamma* di parametri $\alpha > 0$ e $\lambda > 0$ la legge di una variabile aleatoria che ha per densità la funzione

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

Osserviamo che, se $\alpha = 1$, la legge $\Gamma\left(1,\lambda\right)$ è la *distribuzione esponenziale* di parametro λ.

Osservazione 2.1. Se una densità è della forma

$$g\left(x\right) = \begin{cases} cx^{\alpha-1}e^{-\lambda x} & \text{se } x > 0\\ 0 & \text{altrimenti}\end{cases}$$

dove c è una costante positiva allora g è una densità del tipo $\Gamma(\alpha, \lambda)$ e $c = \frac{\lambda^\alpha}{\Gamma(\alpha)}$. Infatti, se g è una densità di probabilità, il suo integrale su $\mathbb{R}$ vale 1 e, con il cambio di variabile $\lambda x = y$ si ha

$$1 = \int_0^{+\infty} g(x)\, \mathrm{d}x = c \int_0^{+\infty} x^{\alpha-1} e^{-\lambda x}\, \mathrm{d}x = \frac{c}{\lambda^\alpha} \int_0^{+\infty} y^{\alpha-1} e^{-y}\, \mathrm{d}x = c\, \frac{\Gamma(\alpha)}{\lambda^\alpha}$$

e, quindi, $c = \frac{\lambda^\alpha}{\Gamma(\alpha)}$.

Vale il seguente teorema:

Teorema 2.2.1. *Se X_1 e X_2 sono due variabili aleatorie indipendenti con leggi rispettivamente $\Gamma(\alpha_1, \lambda)$ e $\Gamma(\alpha_2, \lambda)$, allora la legge della variabile aleatoria $X_1 + X_2$ è $\Gamma(\alpha_1 + \alpha_2, \lambda)$.*

Dimostrazione. Indichiamo con $g(y)$ la densità di $X_1 + X_2$ e con f_1 e f_2 rispettivamente le densità di X_1 e X_2: per la proposizione 1.1.1, vogliamo calcolare

$$g(y) = \int_{-\infty}^{+\infty} f_1(x)\, f_2(y-x)\, \mathrm{d}x = \int_0^y f_1(x)\, f_2(y-x)\, \mathrm{d}x$$

dove l'integrazione è estesa solo all'intervallo $[0, y]$ in quanto f_1 e f_2 sono nulle per valori negativi dei loro argomenti. Ora

$$\begin{aligned}
g(y) &= \int_0^y f_1(x)\, f_2(y-x)\, \mathrm{d}x && = \\
&= \int_0^y \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} x^{\alpha_1-1} e^{-\lambda x} (y-x)^{\alpha_2-1} e^{-\lambda(y-x)}\, \mathrm{d}x && = \\
&= \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} e^{-\lambda y} \int_0^y x^{\alpha_1-1} (y-x)^{\alpha_2-1}\, \mathrm{d}x && = \\
&= \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} e^{-\lambda y} \int_0^1 (ty)^{\alpha_1-1} (y-ty)^{\alpha_2-1}\, y\, \mathrm{d}t && = \\
&= \left(\frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} \int_0^1 t^{\alpha_1-1} (1-t)^{\alpha_2-1}\, \mathrm{d}t \right) \cdot y^{\alpha_1+\alpha_2-1} e^{-\lambda y}.
\end{aligned}$$

Quindi, per l'osservazione 2.1, la parte fra parentesi nell'ultima riga è la costante compatibile con $\Gamma(\alpha_1 + \alpha_2, \lambda)$ e ciò completa la dimostrazione. □

Il teorema può essere esteso a un numero finito di variabili aleatorie $X_i \sim \Gamma(\alpha_i, \lambda)$ a due a due indipendenti: si ha che $X_1 + \ldots + X_n \sim \Gamma(\alpha_1 + \ldots + \alpha_n, \lambda)$. Vogliamo ora vedere che, se $X \sim N(0, \sigma^2)$, allora $X^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$: la densità

di X è $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$. Sia allora G la funzione di ripartizione di X^2 cioè $G(y) = P(X^2 \leq y)$: sarà $G(y) = 0$ se $y \leq 0$ mentre, se $y > 0$ allora

$$G(y) = P\left(X^2 \leq y\right) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \\ = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \, \mathrm{d}x = 2\int_0^{\sqrt{y}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \, \mathrm{d}x.$$

Facciamo il cambio di variabile $x = \sqrt{t}$ e abbiamo

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

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

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

per $y > 0$: la parte non costante è la densità di $\Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$ e, per 2.1, sappiamo che questa è la legge gamma che cerchiamo.

Definizione 2.10. Siano $X_i \sim N(0,1)$ n variabili aleatorie indipendenti: si chiama *legge di chi-quadro* con n gradi di libertà $\chi^2(n)$ la legge della variabile aleatoria ottenuta come $X_1^2 + \ldots + X_n^2$.

Osserviamo che, per quanto appena visto, ogni X_i ha legge $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$ e, per il teorema 2.2.1, $\chi^2(n) \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$.
Possiamo inoltre vedere che vale $E[\chi^2(n)] = n$ e $Var[\chi^2(n)] = 2n$.

Definizione 2.11. Siano $X \sim N(0,1)$ e $Y \sim \chi^2(n)$ due variabili aleatorie indipendenti: allora si chiama *legge t di Student* con n gradi di libertà $t(n)$ la legge di una variabile aleatoria Z della forma

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

Vogliamo calcolare la densità di probabilità di $t(n)$: ci basterà trovare la funzione di ripartizione di Z. Infatti, per definizione è

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} h(u) \, \mathrm{d}u$$

e, nella forma con un integrale, la funzione integranda $h(u)$ è proprio la densità di $t(n)$. Le densità di probabilità di X e Y sono rispettivamente

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

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

Dato che X e Y sono indipendenti, la densità della variabile aleatoria bidimensionale (X, Y) è

$$\begin{aligned} f(x)\, g(y) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} e^{-\frac{1}{2}y} = \\ &= \frac{1}{2^{\frac{n+1}{2}}\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{1}{2}\left(y+x^2\right)} = a \cdot x^{\frac{n}{2}-1} e^{-\frac{1}{2}\left(y+x^2\right)} \end{aligned}$$

se y è positivo e 0 altrimenti. Per qualsiasi $z \in \mathbb{R}$ vale

$$\begin{aligned} F(z) &= P\left(\frac{X}{\sqrt{Y}}\sqrt{n} \leq z\right) = P\left(X \leq \frac{z\sqrt{Y}}{\sqrt{n}}\right) = \\ &= a \int_0^{+\infty} \left(\int_{-\infty}^{\frac{z\sqrt{y}}{\sqrt{n}}} y^{\frac{n}{2}-1} e^{-\frac{1}{2}\left(y+x^2\right)} \, \mathrm{d}x \right) \mathrm{d}y. \end{aligned}$$

Cambiamo variabile in modo che l'intervallo di integrazione del secondo integrale diventi $]-\infty, z]$ cioè $\frac{x\sqrt{n}}{\sqrt{y}} = u$ cioè $x = \frac{\sqrt{y}}{\sqrt{n}}u$: abbiamo quindi

$$\begin{aligned} F(z) &= a \int_0^{+\infty} \left(\int_{-\infty}^{z} y^{\frac{n}{2}-1} e^{-\frac{1}{2}\left(y+\frac{y}{n}u^2\right)} \frac{\sqrt{y}}{\sqrt{n}} \, \mathrm{d}u \right) \mathrm{d}y = \\ &= \frac{a}{\sqrt{n}} \int_0^{+\infty} \left(\int_{-\infty}^{z} y^{\frac{n-1}{2}} e^{-\frac{1}{2}y\left(1+\frac{u^2}{n}\right)} \mathrm{d}u \right) \mathrm{d}y = \\ &= \frac{a}{\sqrt{n}} \int_{-\infty}^{z} \left(\int_0^{+\infty} y^{\frac{n-1}{2}} e^{-\frac{1}{2}y\left(1+\frac{u^2}{n}\right)} \mathrm{d}y \right) \mathrm{d}u. \end{aligned}$$

Ora cambiamo la variabile nell'integrale interno, in modo che l'esponenziale diventi della forma e^{-v} senza cambiare intervallo di integrazione: sarà $\frac{1}{2}y\left(1+\frac{u^2}{n}\right) = v$

cioè $y = 2v\left(1+\frac{u^2}{n}\right)^{-1}$. Abbiamo così

$$F(z) = \frac{a}{\sqrt{n}}\int_{-\infty}^{z}\left(\int_0^{+\infty} 2^{\frac{n-1}{2}} v^{\frac{n-1}{2}}\left(1+\frac{u^2}{n}\right)^{-\frac{n-1}{2}} e^{-v}\cdot 2\left(1+\frac{u^2}{n}\right)^{-1}\mathrm{d}v\right)\mathrm{d}u =$$
$$= 2^{\frac{n+1}{2}}\frac{a}{\sqrt{n}}\int_{-\infty}^{z}\left(1+\frac{u^2}{n}\right)^{-\frac{n+1}{2}}\left(\int_0^{+\infty} v^{\frac{n-1}{2}}e^{-v}\,\mathrm{d}v\right)\mathrm{d}u =$$
$$= 2^{\frac{n+1}{2}}\frac{a}{\sqrt{n}}\Gamma\left(\frac{n+1}{2}\right)\int_{-\infty}^{z}\left(1+\frac{u^2}{n}\right)^{-\frac{n+1}{2}}\mathrm{d}u.$$

Sostituiamo, ora, a con il suo valore

$$F(z) = \frac{1}{2^{\frac{n+1}{2}}\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}2^{\frac{n+1}{2}}\frac{1}{\sqrt{n}}\Gamma\left(\frac{n+1}{2}\right)\int_{-\infty}^{z}\left(1+\frac{u^2}{n}\right)^{-\frac{n+1}{2}}\mathrm{d}u =$$
$$= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\int_{-\infty}^{z}\left(1+\frac{u^2}{n}\right)^{-\frac{n+1}{2}}\mathrm{d}u$$

e otteniamo così il valore della densità di probabilità di Z cioè

$$h(u) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{u^2}{n}\right)^{-\frac{n+1}{2}}.$$

Questa funzione è pari e il suo grafico risulta essere simile a quello della densità di $N(0,1)$ e, anzi, per n abbastanza grande si può sostituire $t(n)$ con $N(0,1)$.
Sia X una variabile aleatoria e F_X la sua funzione di ripartizione: supponiamo che sia continua. Allora l'equazione $F_X(x) = \alpha$ ha sicuramente soluzione per $0 < \alpha < 1$; inoltre, se supponiamo che F_X sia anche strettamente crescente, allora questa soluzione è unica. Indichiamo questa soluzione con q_α: sarà l'unico numero per cui

$$P(\{X \le q_\alpha\}) = \alpha.$$

Definizione 2.12. Si chiama *quantile* di ordine α con $0 < \alpha < 1$ di una variabile aleatoria X il più grande numero q_α tale che

$$F_X(q_\alpha) = P\{X \le q_\alpha\} \le \alpha.$$

Sia $0 < \alpha < 1$, indichiamo con $t_\alpha(n)$ il quantile di ordine α della legge $t(n)$ e con $\chi^2_\alpha(n)$ il quantile di ordine α della legge $\chi^2(n)$. Il quantile è quindi definito dalla relazione $P((n) \le t_\alpha(n))$ e, se F è la funzione di ripartizione di $t(n)$, allora

possiamo scrivere $F(t_\alpha(n)) = \alpha$ cioè $t_\alpha(n) = F^{-1}(\alpha)$; analogamente lo stesso vale per $\chi^2_\alpha(n)$.
Osserviamo che la distribuzione di $t(n)$ è simmetrica quindi i suoi quantili seguono la relazione

$$t_{1-\alpha}(n) = t_{-\alpha}(n).$$

2.3 Teorema di Cochran

Leggi normali multivariate

Definizione 2.13. Una funzione $Z = (Z_1, Z_2) = Z_1 + iZ_2$ a valori complessi si dice *variabile aleatoria complessa* se e solo se Z_1 e Z_2 sono variabili aleatorie reali.

Definizione 2.14. Sia X una variabile aleatoria n-dimensionale: si chiama *funzione caratteristica* di X la funzione

$$\phi(\theta) = E\left[e^{i\langle\theta,X\rangle}\right] = E\left[\cos\langle\theta,X\rangle\right] + iE\left[\sin\langle\theta,X\rangle\right]$$

dove $\langle\cdot,\cdot\rangle$ indica il prodotto scalare di $\mathbb{R}^n$.

Siano X e Y due variabili aleatorie indipendenti e $Z = (X, Y)$:

$$\phi_Z(\theta) = E\left[e^{i\langle\theta,Z\rangle}\right] = E\left[e^{i\theta_1 X}e^{i\theta_2 Y}\right] = E\left[e^{i\theta_1 X}\right]E\left[e^{i\theta_2 Y}\right] = \phi_X(\theta_1)\,\phi_Y(\theta_2).$$

Inoltre, si può vedere anche che, se $\phi_Z(\theta) = \phi_X(\theta_1)\,\phi_Y(\theta_2)$, allora X e Y sono indipendenti. Questo risultato si può generalizzare al caso di n variabili:

Proposizione 2.3.1. *Siano $X_1, \ldots, X_n$ e $X = (X_1, \ldots, X_n)$ variabili aleatorie: allora X_i sono indipendenti se e solo se*

$$\phi_X(\theta) = \phi_{X_1}(\theta_1) \cdot \ldots \cdot \phi_{X_n}(\theta_n).$$

Siano $X_1, \ldots, X_n$ variabili indipendenti di legge $N(0,1)$: allora la variabile aleatoria n-dimensionale avrà densità data da

$$f(x) = \frac{1}{2\pi}e^{-\frac{x_1^2}{2}} \ldots \frac{1}{2\pi}e^{-\frac{x_n^2}{2}} = \frac{1}{(2\pi)^{\frac{m}{2}}}e^{-\frac{|x|^2}{2}}.$$

Inoltre, per la proposizione 2.3.1,

$$\phi_X(\theta) = e^{-\frac{\theta_1^2}{2}} \ldots e^{-\frac{\theta_n^2}{2}} = e^{-\frac{|\theta|^2}{2}}.$$

Definizione 2.15. Sia $z \in \mathbb{R}^m$ e C una matrice $m \times m$ simmetrica e semi-definita positiva: sia

$$\phi_X(\theta) = e^{i\langle\theta,z\rangle} e^{-\frac{1}{2}\langle C\theta,\theta\rangle}$$

una funzione caratteristica. Esiste sempre una variabile aleatoria m-dimensionale che ha ϕ_X come funzione caratteristica: chiameremo la legge di X legge normale multivariata di media z e matrice di covarianza C cioè $N(z, C)$.

Consideriamo $X \sim N(z, C)$ e supponiamo che C sia diagonale con elementi diagonali λ_k: allora

$$\begin{aligned}\phi_X(\theta) &= e^{i\langle\theta,z\rangle} e^{-\frac{1}{2}\langle C\theta,\theta\rangle} = e^{i\langle\theta,z\rangle} e^{-\frac{1}{2}\sum_{k=1}^n \lambda_k\theta_k^2} = \\ &= e^{i\theta_1 z_1} e^{-\frac{1}{2}\lambda_1\theta_1^2} \cdot \ldots \cdot e^{i\theta_n z_n} e^{-\frac{1}{2}\lambda_n\theta_n^2} = \\ &= \phi_{X_1}(\theta_1) \cdot \ldots \cdot \phi_{X_n}(\theta_n).\end{aligned}$$

Per la proposizione 2.3.1 le variabili aleatorie $X_1, \ldots, X_n$ sono indipendenti. Osserviamo che, dato che la matrice C è la matrice di covarianza di $X = (X_1, \ldots, X_n)$, abbiamo appena visto che se X_i non sono correlate allora sono indipendenti se la loro distribuzione congiunta è normale multivariata.

Richiami di algebra lineare . Consideriamo $\mathbb{R}^n$: è naturalmente definito il prodotto scalare

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

Diremo inoltre che due vettori $x, y \in \mathbb{R}^n$ sono *ortogonali* se $\langle x, y\rangle = 0$; due sottospazi $E, F \subseteq \mathbb{R}^n$ si dico *ortogonali* se ogni vettore di E è ortogonale a ogni vettore di F. Sia E un sottospazio di $\mathbb{R}^n$, indichiamo quindi con $E^\perp$ lo spazio ortogonale a E cioè lo spazio composto da tutti i vettori ortogonali a quelli di E: anche $E^\perp$ è un sottospazio di $\mathbb{R}^n$. Possiamo allora scrivere ogni vettore $x \in \mathbb{R}^n$ come $x = u + v$ dove $u \in E$ e $v \in E^\perp$. Inoltre, definiamo il *proiettore ortogonale* su E così

$$\begin{array}{rccc} P_E: & \mathbb{R}^n & \to & E \\ & x & \mapsto & u \end{array}$$

cioè la funzione che associa ad ogni vettore $x \in \mathbb{R}^n$ la sua componente su E.

Proposizione. *Sia $x \in \mathbb{R}^n$ e E un sottospazio di $\mathbb{R}^n$, allora $P_E(x)$ è il vettore di E che si trova a distanza minima da x.*

Dimostrazione. Sia $y \in E$ e scriviamo $x = u + v$ con $u \in E$ e $v \in E^{\perp}$. Allora

$$|y - x|^2 = |\,(y - u) - v|^2 = \langle (y - u) - v, (y - u) - v \rangle = \\ = |y - u|^2 - 2\,\langle y - u, v \rangle + |v|^2 = |y - u|^2 + |v|^2$$

dove $\langle y - u, v \rangle = 0$ perchè $y - u \in E$ e $v \in E^{\perp}$ sono ortogonali. Questo vuol dire che $|y - x|^2$ è sempre maggiore o uguale a $|v|^2 = |x - u|^2$ ed è uguale se e solo se $y = u$. □

Useremo, inoltre, il seguente risultato: sia $B = \{b_1, \ldots, b_s\}$ una base ortogonale di E, allora possiamo calcolare la proiezione di un vettore $x \in \mathbb{R}^n$ così

$$P_E\,(x) = \sum_{i=1}^{s} \frac{\langle x, b_i \rangle}{\langle b_i, b_i \rangle} b_i.$$

Teorema 2.3.2. ***Teorema di Cochran*** *Sia $X = (X_1, \ldots, X_n)$ una variabile aleatoria n-dimensionale di legge $N\,(0, I)$, cioè tale che $X_i \sim N\,(0, 1)$ per ogni i sono indipendenti. Siano $E_1, \ldots, E_k$ sottospazi vettoriali di $\mathbb{R}^n$ a due a due ortogonali; siano $n_i = dim\,(E_i)$ e P_i il proiettore ortogonale su E_i. Allora le variabili aleatorie $P_i\,(X)$ sono indipendenti e la variabile $|||P_i\,(X)\,||^2$ ha legge $\chi^2\,(n_i)$.*

Dimostrazione. Supponiamo che $k = 2$ e che E_1 sia il sottospazio relativo alle prime n_1 coordinate e che E_2 sia il sottospazio relativo alle successive n_2 coordinate: è sempre possibile assumere che E_1 ed E_2 siano fatti così in quanto possiamo sempre trovare una matrice ortogonale M per cambiare le coordinate definite da E_1, E_2 e dal completamento alla base canonica; una trasformazione ortogonale muta una variabile normale multivariata $N\,(0, I)$ sempre in una normale multivariata $N\,(0, I)$ quindi possiamo ricondurci al caso che tratteremo.
Siano $Y = P_1\,(X)$ e $Z = P_2\,(X)$: poiché $X_1, \ldots, X_n$ sono indipendenti, $cov\,(Y_j, Z_k) = 0$ per $1 \leq j \leq n_1$ e $1 \leq k \leq n_2$ dato che Y_j e Z_k sono semplicemente due delle $X_1, \ldots, X_n$. Questo implica che Y e Z sono indipendenti in quanto hanno covarianza nulla e distribuzione congiunta normale. Vogliamo ora vedere che $||P_i\,(X)\,||^2 \sim \chi^2\,(n_i)$: sappiamo già che $\chi^2\,(s)$ è la legge della somma di s quadrati di variabili aleatorie indipendenti e ciascuna con legge $N\,(0, 1)$ e, inoltre,

$$||P_1\,(X)\,||^2 = X_1^2 + \ldots + X_{n_1}^2 \sim \chi^2\,(n_1)\,;\; ||P_2\,(X)\,||^2 = X_{n_1+1}^2 + \ldots + X_{n_1+n_2}^2 \sim \chi^2\,(n_2)$$

cioè è verificato il teorema per $k = 2$. □

Corollario 2.3.3. *Siano $Z_1, \ldots, Z_m$ variabili aleatorie indipendenti di legge $N(\mu, \sigma^2)$ e siano*

$$\bar{Z} = \frac{1}{m}(Z_1 + \ldots + Z_m)$$

e

$$S^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Z_i - \bar{Z}\right)^2.$$

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

$$\frac{m-1}{\sigma^2}S^2 \sim \chi^2(m-1) \quad \frac{\sqrt{m}(\bar{z}-\mu)}{S} \sim t(m-1)$$

Dimostrazione. Vediamo innanzitutto il caso in cui $Z_i \sim N(0,1)$ per ogni i. Consideriamo il sottospazio di $\mathbb{R}^m$ dato da $E = \langle(1,1,\ldots,1)\rangle$: il proiettore ortogonale è

$$\begin{aligned} P_E(x) &= \frac{\langle x,(1,1,\ldots,1)\rangle}{\langle(1,1,\ldots,1),(1,1,\ldots,1)\rangle}(1,1,\ldots,1) = \\ &= \frac{x_1+\ldots+x_m}{m}(1,1,\ldots,1) = (\bar{x},\bar{x},\ldots,\bar{x}) \end{aligned}$$

dove $\bar{x} = \frac{1}{m}(x_1 + \ldots + x_m)$. Sia ora $Z = (Z_1, \ldots, Z_m)$: la proiezione ortogonale di Z su E è $P_E(Z) = \left(\bar{Z}, \bar{Z}, \ldots, \bar{Z}\right)$ mentre la proiezione ortogonale di Z su $E^\perp$ è $P_{E^\perp}(Z) = (1 - P_E)(Z) = \left(Z_1 - \bar{Z}, \ldots, Z_m - \bar{Z}\right)$. E e $E^\perp$ sono sottospazi tra loro ortogonali di $\mathbb{R}^m$: possiamo applicare il teorema di Cochran e avremo che $P_E(Z)$ e $P_{E^\perp}(Z)$ sono variabili aleatorie indipendenti; quindi saranno indipendenti $\bar{Z}$ e $P_{E^\perp}(Z)$ poiché $\bar{Z}$ è una componente di $P_E(Z)$. Osserviamo ora che

$$||P_{E^\perp}(Z)||^2 = \sum_{i=1}^{m}\left(Z_i - \bar{Z}\right)^2 = (m-1)S^2$$

quindi S^2 è indipendente da $\bar{Z}$ e, per il teorema di Cochran, $(m-1)S^2 \sim \chi^2(m-1)$. Sappiamo che la legge di $\bar{Z}$ è $N\left(0, \frac{1}{m}\right)$ e quindi $\sqrt{m}\bar{Z} \sim N(0,1)$; per la definizione della legge di Student, avremo quindi che

$$T = \frac{\sqrt{m}\bar{Z}}{S} \sim t(m-1).$$

Abbiamo così dimostrato il caso in cui $Z_i \sim N(0,1)$ per ogni i: per il caso generale definiamo $X_i = \frac{Z_i - \mu}{\sigma}$ in modo da avere $X_i \sim N(0,1)$. Per quanto abbiamo visto

nella prima parte, $\bar{X} = \frac{1}{m}(X_1 + \ldots + X_m)$ e $(m-1)S_X^2 = \sum_{i=1}^{m}\left(X_i - \bar{X}\right)^2$ sono indipendenti. Inoltre, per ogni i, $Z_i = \sigma X_i + \mu$ quindi

$$\bar{Z} = \sigma\bar{X}\mu \quad \sum_{i=1}^{m}\left(X_i - \bar{X}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{m}\left(Z_i - \bar{Z}\right)^2 = \frac{m-1}{\sigma^2}S^2.$$

$\bar{Z}$ e S^2 sono indipendenti in quanto funzioni di variabili indipendenti; per quanto visto nella prima parte, $\frac{m-1}{\sigma^2}S^2 \sim \chi^2(m-1)$ e, dato che $\frac{\sqrt{m}\left(\bar{Z}-\mu\right)}{S} = \frac{\sqrt{m}\bar{X}}{S_X}$, $\frac{\sqrt{m}\left(\bar{Z}-\mu\right)}{S} \sim t(m-1)$. □

2.4 Quantità pivotali

Definizione 2.16. Sia $X = (X_1, \ldots, X_n)$ un'osservazione da cui si vuole stimare un parametro scalare θ e $Q(X, \theta)$ una funzione dell'osservazione e del parametro: se la legge di Q rispetto a P^θ è indipendente da θ (ovvero per ogni a, $b \in \mathbb{R}$ con $a < b$ la probabilità $P^\theta(a \leq Q(X,\theta) \leq b)$ non è funzione di θ allora $Q(X,\theta)$ si dice *quantità pivotale*.

Osserviamo che, nonostante la sua legge non dipende da θ, una quantità pivotale dipende comunque da θ e non è quindi una statistica.
Inoltre, se conosciamo la legge di Q, possiamo fissare $\alpha \in]0,1[$ e trovare due numeri q_1 e q_2 tali che $P^\theta(q_1 \leq Q(X,\theta) \leq q_2) = 1-\alpha$: se la relazione $q_1 \leq Q(X,\theta) \leq q_2$ è risolvibile rispetto a θ cioè si possono trovare $t_1(X)$ e $t_2(X)$ tali per cui $t_1(X) \leq \theta \leq t_2(X)$ allora

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

Questo vuol dire che l'intervallo $[t_1(X), t_2(X)]$ è un intervallo di fiducia per θ di livello $1-\alpha$.
Ad esempio, per la quantità pivotale $T(X,\mu) = \sqrt{n}\frac{\bar{X}-\mu}{S}$, un intervallo di fiducia ricavato con questo metodo sarà

$$\bar{x} - \frac{S}{\sqrt{n}}q_2 \leq \mu \leq \bar{x} - \frac{S}{\sqrt{n}}q_1$$

con q_1 e q_2 adeguati.

2.5 Stime per media e varianza di campioni gaussiani

Siano X_i n variabili aleatorie equidistribuite con legge $N(\mu, \sigma^2)$ e $X = (X_1, \ldots, X_n)$: vogliamo costruire un intervallo di confidenza per μ.
Supponiamo innanzitutto di conoscere la varianza σ^2: la media e la varianza di $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ sono rispettivamente μ e $\frac{1}{n}\sigma^2$ quindi $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)$ ha legge $N(0,1)$. Sia allora $\delta > 0$: vorremo che

$$P\left(|\bar{X} - \mu| \leq \delta\right) = P\left(|Z| \leq \frac{\sqrt{n}}{\sigma}\delta\right) = 2\Phi\left(\frac{\sqrt{n}}{\sigma}\delta\right) - 1.$$

valga almeno $1 - \alpha$ quindi

$$2\Phi\left(\frac{\sqrt{n}}{\sigma}\delta\right) - 1 \geq 1 - \alpha$$

$$\Phi\left(\frac{\sqrt{n}}{\sigma}\delta\right) \geq 1 - \frac{\alpha}{2}$$

$$\frac{\sqrt{n}}{\sigma}\delta \geq \phi_{1-\frac{\alpha}{2}}$$

$$\delta \geq \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}.$$

Abbiamo quindi visto che $P\left(|\bar{X} - \mu| \leq \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$ cioè la probabilità che

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

è $1 - \alpha$: questo è un intervallo di confidenza per μ.
Il caso più vicino alla realtà è quello in cui la varianza σ^2 non è nota: idealmente vorremmo ottenere un risultato simile a quello appena visto con σ^2 sostituito dal suo stimatore S^2 ma questo accade solo in parte e abbiamo alcune approssimazioni da fare che non sono banali. Consideriamo allora $\bar{X}$ e S^2 cioè gli stimatori corretti di μ e σ^2: sappiamo che

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

e, per il corollario del teorema di Cochran,

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

e W e Z sono indipendenti. Allora

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

e nell'espressione di T non compare σ^2: tramite i quantili di $t\left(n-1\right)$ possiamo trovare intervalli di confidenza per μ. Abbiamo

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

cioè un intervallo di confidenza per μ al livello $1-\alpha$ è $\left[\bar{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}, \bar{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}\right]$: la differenza più significativa che abbiamo col caso in cui σ^2 è nota è data dal fatto che i quantili della legge normale $\phi_{1-\frac{\alpha}{2}}$ sono sostituiti dai quantili della legge di Student $t_{1-\frac{\alpha}{2}}\left(n-1\right)$ e questo fa in modo che a parità di livello di fiducia i quantili per t siano maggiori di quelli per la normale quindi l'intervallo di confidenza è più ampio. Vogliamo ora ottenere intervalli di confidenza anche per la varianza σ^2: per farlo ci serviremo di una opportuna quantità pivotale. Sappiamo già che $W\left(X,\sigma\right) = \left(n-1\right)\frac{S^2}{\sigma^2}$ ha legge $\chi^2\left(n-1\right)$ che è indipendente da σ: quindi W è una quantità pivotale per la varianza. Possiamo allora trovare un espressione per un intervallo di fiducia per σ^2 sia unilaterale che bilatero: nel primo caso avremo una maggiorazione per la varianza

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

cioè $\left[0, \frac{(n-1)S^2}{\chi^2_\alpha(n-1)}\right]$ è un intervallo di confidenza unilaterale di livello $1-\alpha$ per la varianza; nel secondo caso, invece

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

cioè $\left[\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$ è un intervallo di confidenza bilatero di livello $1-\alpha$ per la varianza.

2.6 Stima della probabilità di successo in una sequenza di prove indipendenti

Sia $X = (X_1 + \ldots + X_n)$ un'osservazione in cui $X_i \in \{0, 1\}$ cioè un'osservazione di uno schema successo-insuccesso, indichiamo con $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ la media campionaria: se n è abbastanza grande allora la variabile aleatoria $Y = \sqrt{n}\frac{\bar{X}-\theta}{\sqrt{\theta(1-\theta)}}$ ha approssimativamente distribuzione $N(0,1)$. Fissiamo $\alpha \in]0,1[$ e con probabilità $1-\alpha$ avremo

$$-\phi_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}-\theta}{\sqrt{\theta(1-\theta)}} \leq \phi_{1-\frac{\alpha}{2}}. \tag{2.4}$$

Y è una quantità pivotale in quanto la sua legge è indipendente da θ: vorremmo poter usare le disequazioni 2.4 come avevamo già fatto per stimare la varianza ma in questo caso otteniamo

$$\bar{X} - \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \leq \theta \leq \bar{X} + \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}$$

in cui abbiamo ancora θ sia nel primo che nell'ultimo membro; osserviamo che, per ogni $\theta \in]0,1[$ si ha che $\theta(1-\theta) \leq \frac{1}{4}$ possiamo sostituirlo nella disequazione e ottenere

$$\bar{X} - \frac{1}{2\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \leq \theta \leq \bar{X} + \frac{1}{2\sqrt{n}}\phi_{1-\frac{\alpha}{2}}$$

cioè $\left[\bar{X} - \frac{1}{2\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \bar{X} + \frac{1}{2\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right]$ è un intervallo di confidenza per θ al livello $1-\alpha$. In realtà, per valori di θ vicini a 0 o a 1, la maggiorazione che abbiamo usato per $\theta(1-\theta)$ è abbastanza ampia e questo ci dà un intervallo di confidenza molto più ampio di quello che vorremmo avere: svolgiamo in modo più accurato il calcolo partendo da 2.4.

$$-\phi_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}-\theta}{\sqrt{\theta(1-\theta)}} \leq \phi_{1-\frac{\alpha}{2}} \Leftrightarrow \frac{n\left(\bar{X}-\theta\right)^2}{\theta(1-\theta)} \leq \phi^2_{1-\frac{\alpha}{2}} \Leftrightarrow$$

$$\Leftrightarrow n\bar{X}^2 - 2n\bar{X}\theta + n\theta^2 \leq \phi^2_{1-\frac{\alpha}{2}}\theta - \phi^2_{1-\frac{\alpha}{2}}\theta^2 \Leftrightarrow$$

$$\Leftrightarrow \left(n + \phi^2_{1-\frac{\alpha}{2}}\right)\theta^2 - \left(2n\bar{X} + \phi^2_{1-\frac{\alpha}{2}}\right) + n\bar{X}^2 \leq 0$$

dove abbiamo eliminato il denominatore in quanto per $\theta \in]0,1[$ si ha che $\theta(1-\theta) > 0$. Quindi

$$\theta = \frac{2n\bar{X} + \phi^2_{1-\frac{\alpha}{2}} \pm \sqrt{4n^2\bar{X}^2 + 4n\bar{X}\phi^2_{1-\frac{\alpha}{2}} + \phi^4_{1-\frac{\alpha}{2}} - 4n^2\bar{X}^2 - 4n\bar{X}^2\phi^2_{1-\frac{\alpha}{2}}}}{2\left(n + \phi^2_{1-\frac{\alpha}{2}}\right)} =$$

$$= \frac{2n\bar{X} + \phi^2_{1-\frac{\alpha}{2}} \pm \sqrt{4n\bar{X} + \phi^2_{1-\frac{\alpha}{2}} - 4n\bar{X}^2}}{2\left(n + \phi^2_{1-\frac{\alpha}{2}}\right)} = W^{\pm}$$

cioè 2.4 equivale a $W^- \leq \theta \leq W^+$ e $[W^-, W^+]$ è un intervallo di confidenza per θ al livello $1-\alpha$: questa espressione per l'intervallo di confidenza ha calcoli abbastanza complicati e può essere sostituita da una sua versione approssimata che non modifica molto l'intervallo. Infatti, $\phi_{0,975} = 1,96$ quindi se tutte le altre grandezze presenti in $W^{\pm}$ sono sensibilmente più grandi possiamo allora toglierlo dall'espressione di $W^{\pm}$ e usare l'approssimazione

$$W^{\pm} \simeq \bar{X} \pm \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}}$$

per avere $\left[\bar{X} - \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}}, \bar{X} + \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}}\right]$ come intervallo di confidenza approssimato per θ al livello $1-\alpha$.

Capitolo 3

Test statistici

Nel primo capitolo abbiamo visto come si possono stimare parametri incogniti usando le informazioni ricavate da risultati di esperimenti. Lo scopo di questo capitolo sarà invece quello di illustrare come si svolgono i test statistici ovvero come si può decidere, con un preciso livello di affidabilità, fra due possibili alternative di cui non conosciamo con assoluta certezza quale sia vera: è importante sottolineare quindi che un test statistico non stabilisce quale delle due ipotesi sia vera ma suggerisce quale delle due accettare, valutando la probabilità di possibili errori. Ricordiamo inoltre che per poter svolgere un test statistico, si dovranno sia approssimare i parametri in esso coinvolti che scegliere un insieme di risultati di un esperimento aleatorio: cambiare, anche di poco, uno di questi valori può dare un risultato diverso e sarà importante saper valutare anche questi cambiamenti.
In generale, l'ipotesi che vogliamo decidere se accettare o meno verrà tradotta, a livello matematico, nello stabilire se un parametro θ appartiene o meno a un certo sottoinsieme, deciso a priori, dell'insieme Θ in cui il parametro può variare: dividiamo quindi l'insieme Θ in due sottoinsiemi disgiunti Θ_H e Θ_A e tali che $\Theta = \Theta_H \sqcup \Theta_A$. L'insieme Θ_H si chiama *ipotesi* o *ipotesi nulla* mentre Θ_A si dice *alternativa* o *ipotesi alternativa*: si dirà quindi che l'ipotesi è vera se $\theta \in \Theta_H$ e che l'ipotesi è falsa se $\theta \in \Theta_A$. Tradizionalmente lo scopo del test statistico è quello di confutare l'ipotesi. Non sappiamo però quale delle due ipotesi sia quella effettivamente vera: nello svolgimento del test potremmo essere condotti ad accettare l'ipotesi anche senza che essa sia vera. Vogliamo quindi poter trovare un procedimento che ci dica con quale probabilità stiamo commettendo questi errori: per fare ciò costruiamo un modello statistico e stabiliamo una regola che, in base ai valori delle osservazioni X, ci permette di decidere quale ipotesi accettare.

Definizione 3.1. Sia $X = (X_1, \ldots, X_n)$ un'osservazione: si chiama *regione di*

rigetto o *regione critica* un sottoinsieme D dell'insieme Ω dei valori che X può assumere per cui se $X \in D$ allora $\theta \notin \Theta_H$ cioè se l'osservazione assume valori in questa regione, non accettiamo l'ipotesi.

La scelta della regione di rigetto di un test è un momento delicato dello studio in quanto non c'è una regola vincolante: se si pensa che l'ipotesi sia vera, allora si cercherà di scegliere D in modo che la probabilità di $X \in D$ sia grande mentre se si pensa che l'ipotesi sia falsa, la scelta di D sarà fatta in modo da avere una bassa probabilità di $X \in D$. Nella maggior parte dei casi D è un intervallo del tipo $]-\infty, k]$ o $[k, +\infty[$ che viene intersecato con Ω; inoltre, idealmente, la scelta di D dovrebbe avvenire prima delle osservazioni, in modo che non sia influenzata dai risultati di X.

Definizione 3.2. Chiamiamo

- *errore di prima specie* quello che si commette se rigettiamo l'ipotesi quando essa è vera cioè quando $X \in D$ ma $\theta \in \Theta_H$;
- *errore di seconda specie* quello che si commette se non respingiamo l'ipotesi quando essa è falsa cioè se $X \notin D$ ma $\theta \in \Theta_A$.

Chiaramente la probabilità di un errore sia di prima che di seconda specie è non nulla: per come vengono svolti generalmente i test statistici, l'errore di prima specie risulta più grave in quanto porta ad accettare un'ipotesi non vera mentre l'errore di seconda specie fa in modo che venga rifiutata un'alternativa non falsa. Vorremmo quindi poter trovare il modo di minimizzare entrambe le possibilità di questi due errori: non sarà possibile farlo in quanto vedremo che sono indipendenti l'una dall'altra e ci limiteremo a cercare di minimizzare quella dell'errore di prima specie.

Definizione 3.3. Si chiama *potenza* di un test con regione critica D la funzione

$$\pi_D : \Theta \to [0,1]\,; \quad \pi_D(\theta) = P^\theta(X \in D)\,.$$

Sia θ_0 il valore vero del parametro θ. Se $\theta_0 \in \Theta_H$, $X \in D$ fa in modo che rifiutiamo l'ipotesi e quindi abbiamo un errore di prima specie allora $\pi_D(\theta_0)$ è la probabilità dell'errore di prima specie; se, invece, $\theta_0 \notin \Theta_H$ allora $X \notin D$ ci porta a commettere l'errore di seconda specie e quindi $1 - \pi_D(\theta_0)$ sarà la probabilità dell'errore di seconda specie.

Osserviamo che se conoscessimo il valore vero del parametro θ allora potremmo sapere quale dei due errori possono essere commessi: infatti un errore di prima specie può avvenire solo se $\theta_0 \in \Theta_H$ mentre un errore di seconda specie può avvenire solo se $\theta_0 \in \Theta_A$ quindi non si possono verificare entrambi contemporaneamente. Non conosciamo però θ_0 quindi non sappiamo esattamente le probabilità di commettere i due tipi di errori dato che P^θ dipende dal valore stimato di θ_0: dato che l'errore di prima specie è quello che vogliamo maggiormente evitare, vorremmo poter avere una stima della probabilità di commetterlo.

Definizione 3.4. Sia D la regione critica di un test: si chiama *livello* del test il numero $\alpha_D = \sup_{\theta \in \Theta_H} P^\theta (X \in D)$.

Quindi, per minimizzare la probabilità di un errore di prima specie, sarà opportuno scegliere un α piccolo, dato che si tratta dell'estremo superiore della probabilità di commettere un errore di prima specie.

3.1 Confronto fra due probabilità in una prova bernoulliana

Consideriamo due variabili aleatorie X e Y che assumono valore 1 con probabilità, rispettivamente, p_1 e p_2 e valore 0 con probabilità $1 - p_1$ e $1 - p_2$. In base a n osservazioni indipendenti di X e m osservazioni indipendenti di Y si vuole stabilire se si può accettare o no l'ipotesi bilaterale $p_1 = p_2$ o una delle due ipotesi unilaterali $p_1 \leq p_2$ o $p_1 \geq p_2$.

Ipotesi bilaterale Siano $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ rispettivamente le n osservazioni di X e le m di Y. Allora le due variabili aleatorie

$$B_X = X_1 + \ldots + X_n \sim B(n, p_1) \text{ e } B_Y = Y_1 + \ldots + Y_m \sim B(m, p_2)$$

hanno distribuzione binomiale: se n, m, np_1 e mp_2 sono abbastanza grandi allora possiamo approssimare $B_X \sim N(np_1, np_1(1-p_1))$ e $B_Y \sim N(mp_2, mp_2(1-p_2))$; avremo quindi che

$$\bar{X} = \frac{B_X}{n} \sim N\left(p_1, \frac{p_1(1-p_1)}{n}\right) \text{ e } \bar{Y} = \frac{B_Y}{m} \sim N\left(p_2, \frac{p_2(1-p_2)}{m}\right)$$

$\bar{X}$ e $\bar{Y}$ sono le frequenze relative di ottenere 1 rispettivamente nelle sequenze $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_m)$: sono gli stimatori di p_1 e p_2. L'ipotesi bilaterale non è accettabile se la differenza $p_1 - p_2$ risulta essere più grande di $\delta > 0$: dato che non conosciamo p_1 e p_2 potremo stimare la differenza con

$$\bar{X} - \bar{Y} = \bar{X} + \left(-\bar{Y}\right) \sim N\left(p_1 - p_2, \frac{p_1\left(1-p_1\right)}{n} + \frac{p_2\left(1-p_2\right)}{m}\right)$$

Abbiamo quindi che

$$\frac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \sim N(0,1)$$

e, se si verifica l'ipotesi bilaterale, avremo $p_1 - p_2 = 0$ quindi

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \sim N(0,1)$$

dove, non conoscendo p_1 e p_2, il denominatore non è noto. Quindi, se il campione è sufficientemente ampio, possiamo ottenere un'approssimazione di $\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$ sostituendo p_1 e p_2 rispettivamente con i loro stimatori $\bar{X}$ e $\bar{Y}$. Vogliamo allora confrontare

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n} + \frac{\bar{Y}\left(1-\bar{Y}\right)}{m}}}$$

con i quantili di $N(0,1)$: una regione di rigetto per l'ipotesi bilaterale al livello α per T sarà quindi

$$D = \left]-\infty, -\phi_{1-\frac{\alpha}{2}}\right] \cup \left]\phi_{1-\frac{\alpha}{2}}, +\infty\right[.$$

Ipotesi unilaterale Consideriamo innanzitutto l'ipotesi unilaterale $p_1 \geq p_2$: poichè $\bar{X}$ e $\bar{Y}$ sono stimatori per p_1 e p_2, accetteremo l'ipotesi se $\bar{X} - \bar{Y}$ è non negativo. Come prima, abbiamo

$$\bar{X} - \bar{Y} \sim N\left(p_1 - p_2, \frac{p_1\left(1-p_1\right)}{n} + \frac{p_2\left(1-p_2\right)}{m}\right)$$

ma, dato che l'ipotesi non è più $p_1 - p_2 = 0$, T come abbiamo definito nel paragrafo precedente non è più distribuita come $N(0,1)$ ma

$$Z = \frac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n} + \frac{\bar{Y}\left(1-\bar{Y}\right)}{m}}} \sim N(0,1).$$

Fissiamo allora il livello del test $\alpha \in]0,1[$ e avremo

$$\alpha = P\left(Z \leq -\phi_{1-\alpha}\right) = P\left(T - \frac{p_1 - p_2}{\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n} + \frac{\bar{Y}\left(1-\bar{Y}\right)}{m}}}\right) \leq -\phi_{1-\alpha}.$$

Dato che stiamo supponendo $p_1 - p_2 \geq 0$,

$$P\left(T - \frac{p_1 - p_2}{\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n} + \frac{\bar{Y}\left(1-\bar{Y}\right)}{m}}}\right) \geq P\left(T \leq -\phi_{1-\alpha}\right)$$

e quindi questo vuol dire che, se $p_1 \geq p_2$, allora la probabilità che $T \leq -\phi_{1-\alpha}$ è minore o uguale ad α cioè $D =]-\infty, -\phi_{1-\alpha}[$ è una regione di rigetto per l'ipotesi unilaterale che abbiamo preso in considerazione.
Nel caso in cui abbiamo l'ipotesi unilaterale $p_1 \leq p_2$ possiamo ripetere lo stesso ragionamento appena fatto, semplicemente scambiando i ruoli di X e Y: otterremo che una regione critica in questo caso è data da $D =]\phi_{1-\alpha}, +\infty[$.

3.2 Test su medie e varianza di popolazioni gaussiane

Fino ad ora abbiamo trattato casi in cui vogliamo stimare la media di una variabile bernoulliana: avremo ora a che fare con una variabile gaussiana in cui sono ignote sia la media che la varianza. Sia allora $X_1, \ldots, X_n$ un campione di variabili indipendenti ciascuna con legge $N(\mu, \sigma^2)$. Consideriamo un valore μ_0 con cui vogliamo confrontare μ: potremo avere un'ipotesi bilaterale quando vogliamo vedere se $\mu = \mu_0$ o unilaterale nel caso in cui vogliamo verificare se $\mu \leq \mu_0$ o $\mu \geq \mu_0$. Osserviamo innanzitutto che il parametro incognito, in questo caso, è $\theta = (\mu, \sigma^2)$ quindi, oltre all'ipotesi che scegliamo di verificare, abbiamo come condizione anche $\sigma^2 > 0$.

Ipotesi unilaterale Consideriamo innanzitutto l'ipotesi unilaterale $\mu \leq \mu_0$: corrisponde all'insieme $\Theta_H = \{(\mu, \sigma^2) : \mu \leq \mu_0 \text{ e } \sigma^2 > 0\}$. Siano allora $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}$ e $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)$ gli stimatori corretti di μ e σ^2; definiamo inoltre $T = \sqrt{n}\frac{\bar{X}-\mu_0}{S}$. Dato che μ è il vero valore della media di X_i allora sappiamo

che $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1)$. Supponiamo che l'ipotesi sia vera, allora $\sqrt{n}\frac{\mu-\mu_0}{S} \leq, 0$. Fissiamo allora $\alpha \in]0,1[$: sarà

$$P^\theta(T > t_{1-\alpha}) = P^\theta\left(\sqrt{n}\frac{\bar{X}-\mu}{S} + \sqrt{n}\frac{\mu-\mu_0}{S} > t_{1-\alpha}\right) \leq$$
$$\leq P^\theta\left(\sqrt{n}\frac{\bar{X}-\mu}{S} > t_{1-\alpha}\right) = \alpha$$

dove abbiamo usato $t_{1-\alpha}$ per indicare il quantile di ordine $1-\alpha$ di $t(n-1)$. Quindi T appartiene all'intervallo $D =]t_{1-\alpha}, +\infty[$ con probabilità α cioè D può essere preso come regione critica di livello α per accettare l'ipotesi unilaterale.
Nel caso in cui $\mu \geq \mu_0$ possiamo ripetere il ragionamento appena visto, definendo $\bar{X}$, S^2 e T allo stesso modo. In questo caso abbiamo di nuovo che $\sqrt{n}\frac{\bar{X}-\mu}{S} \sim t(n-1)$ ma ora, se l'ipotesi è vera, $\sqrt{n}\frac{\mu-\mu_0}{S} > 0$. Quindi

$$P^\theta(T \leq t_\alpha) = P^\theta\left(\sqrt{n}\frac{\bar{X}-\mu}{S} + \sqrt{n}\frac{\mu-\mu_0}{S} \leq t_\alpha\right) \leq$$
$$\leq P^\theta\left(\sqrt{n}\frac{\bar{X}-\mu}{S} \leq t_\alpha\right) = \alpha$$

cioè $D =]-\infty, t_\alpha[$ è una regione critica di livello α per questa ipotesi.

Ipotesi bilaterale Analogamente al caso unilaterale, usiamo $\bar{X}$, S^2 e T come sono già stati definiti; inoltre avremo che, se $\mu = \mu_0$ allora $T = \sqrt{n}\frac{\bar{X}-\mu_0}{S} \sim t(n-1)$. Quindi

$$P^\theta\left(|T| \geq t_{1-\frac{\alpha}{2}}\right) = \alpha$$

cioè una regione critica in questo caso è $D = \left]-\infty, -t_{1-\frac{\alpha}{2}}\right[\cup\left]t_{1-\frac{\alpha}{2}}, \infty\right[$.

Osservazione 3.1. Possiamo utilizzare questi risultati anche nel caso in cui le variabili aleatorie X_i non siano distribuite come $N(\mu, \sigma^2)$: in questo caso, infatti, se n è sufficientemente grande, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ è ben approssimato da una variabile aleatoria gaussiana.

3.2.1 Confronto fra le medie di due campioni

Abbiamo appena visto come confrontare la media di un campione con un valore scelto a priori, un problema un po' più complicato è dato dal confronto fra le medie di due diversi campioni.

Campioni accoppiati Siano $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$ i due campioni, entrambi formati da n variabili aleatorie normali con media rispettivamente μ_X e μ_Y ed equidistribuite: supponiamo che le X_i siano indipendenti e, allo stesso modo, che le Y_i siano indipendenti ma non che le X_i siano indipendenti in relazione con le Y_i. Ci riconduciamo al caso precedente semplicemente definendo $Z_i = X_i - Y_i$: infatti le Z_i così definite sono indipendenti e la loro media è data da $\mu_Z = \mu_X - \mu_Y$. Qualsiasi ipotesi vorremo testare sulle medie μ_X e μ_Y potremo quindi testarla semplicemente confrontando μ_Z e 0.

Campioni indipendenti Consideriamo $X_1, \ldots, X_n$ un campione di variabili aleatorie gaussiane equidistribuite con media μ_X e varianza σ^2 e $Y_1, \ldots, Y_m$ un campione di variabili aleatorie gaussiane equidistribuite con media μ_Y e varianza σ^2 tutte indipendenti: stiamo supponendo che la varianza delle X_i e delle Y_i sia uguale. Vogliamo allora testare sia l'ipotesi bilaterale $\mu_X = \mu_Y$ che l'ipotesi unilaterale $\mu_X \leq \mu_Y$ oppure $\mu_X \geq \mu_Y$. Vediamo innanzitutto l'ipotesi bilaterale: chiamiamo $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$, allora sappiamo che

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{n-1}{\sigma^2}S_X^2 \sim \chi^2(n-1) \tag{3.1}$$

e

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)^2 = \frac{m-1}{\sigma^2}S_Y^2 \sim \chi^2(m-1). \tag{3.2}$$

Inoltre queste due variabili sono indipendenti da $\bar{X}$ e $\bar{Y}$. Chiamiamo

$$S_{tot}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)^2\right)$$

o anche

$$S_{tot}^2 = \frac{1}{n+m-2}\left((n-1)S_X^2 + (m-1)S_Y^2\right)$$

questa è la media pesata dei due stimatori delle varianze, con pesi proporzionali alla grandezza dei campioni: usando 3.1 e 3.2 possiamo ottenere

$$\frac{n+m-2}{\sigma^2}S_{tot}^2 \sim \chi^2(n-1) + \chi^2(m-1) \sim \chi^2(n+m-2).$$

Quest'ultimo passaggio viene dal fatto che $\chi^2(k) = \Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$ e $\Gamma(\alpha_1, \beta) + \Gamma(\alpha_2, \beta) \sim \Gamma(\alpha_1 + \alpha_2, \beta)$. Abbiamo, inoltre, $\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right)$ e $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{m}\right)$. Se

l'ipotesi $\mu_X = \mu_Y$ è vera allora

$$\bar{X} - \bar{Y} \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

e quindi

$$\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1).$$

Useremo, per la verifica dell'ipotesi, la statistica

$$T = \frac{\bar{X} - \bar{Y}}{S_{tot}\sqrt{\frac{1}{n} + \frac{1}{m}}}:$$

possiamo scriverla così

$$T = \frac{\bar{X} - \bar{Y}}{S_{tot}\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \frac{\sigma}{S_{tot}} \frac{\sqrt{n+m-2}}{\sqrt{n+m-2}} =$$
$$= \sqrt{n+m-2} \frac{\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{(n+m-2)\frac{S_{tot}^2}{\sigma^2}}}$$

e questo mostra che

$$T = \frac{\bar{X} - \bar{Y}}{S_{tot}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2).$$

Possiamo quindi usare i quantili di $t(n+m-2)$ per definire una regione critica per l'ipotesi bilaterale al livello $\alpha \in]0,1[$ fissato: infatti

$$\alpha = P\left(|T| \geq t_{1-\frac{\alpha}{2}}(n+m-2)\right)$$

quindi la regione critica sarà $D = \left]-\infty, -t_{1-\frac{\alpha}{2}}(n+m-2)\right[\cup \left]t_{1-\frac{\alpha}{2}}, +\infty\right[$.
Consideriamo ora l'ipotesi unilaterale $\mu_X \geq \mu_Y$: le ipotesi sui due campioni sono uguali a quelle del caso dell'ipotesi bilaterale. In questo caso però avremo che

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1).$$

Useremo, inoltre, la stessa statistica T che avevamo usato per il caso bilaterale ma non possiamo più dire che è distribuita come $t(n+m-2)$: avremo invece che

$$T' = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_{tot}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2).$$

Scegliamo di cercare una regione critica per il test della forma $]-\infty, -k]$ con $k > 0$ in quanto, se l'ipotesi è vera, valori negativi di T sono poco probabili. Fissiamo $\alpha \in]0,1[$ e avremo quindi che

$$\begin{aligned}
\alpha &= P\left(t(n+m-2) \leq -t_{1-\alpha}(n+m-2)\right) = \\
&= P\left(\frac{\bar{X} - \bar{Y}}{S_{tot}\sqrt{\frac{1}{n} + \frac{1}{m}}} - \frac{(\mu_X - \mu_Y)}{S_{tot}\sqrt{\frac{1}{n} + \frac{1}{m}}} \leq -t_{1-\alpha}(n+m-2)\right) \leq \\
&\leq P\left(\frac{\bar{X} - \bar{Y}}{S_{tot}\sqrt{\frac{1}{n} + \frac{1}{m}}} \leq -t_{1-\alpha}(n+m-2)\right)
\end{aligned}$$

dove, l'ultima disuguaglianza viene dal fatto che stiamo supponendo $\mu_X - \mu_Y \geq 0$. Quindi una regione critica per il test è data da $D =]-\infty, -t_{1-\alpha}(n+m-2)]$.

3.2.2 Test di Fisher

Abbiamo visto, nei paragrafi precedenti, come poter confrontare la media di un campione sia con un valore teorico che con la media di un campione diverso: vorremmo poter avere un test che ci permetta di analizzare anche la varianza di un campione allo stesso modo.

Confronto con un valore assegnato

Siano $X_1, \ldots, X_n$ n variabili gaussiane indipendenti ed equidistribuite con media μ e varianza σ^2 incognite e σ_0^2 un valore assegnato. Vogliamo quindi un test per decidere se accettare o no l'ipotesi $\sigma^2 \leq \sigma_0^2$. Definiamo

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

e fissiamo il livello di fiducia $\alpha \in]0,1[$: cercheremo una regione di rigetto per l'ipotesi della forma $D = [a, +\infty[$ cioè respingeremo l'ipotesi se $S^2 \geq a$. Cerchiamo

quindi di determinare a in funzione di α: per il corollario del teorema di Cochran, sappiamo che $\bar{X}$ e S^2 sono indipendenti e che $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$. Quindi

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

e, se l'ipotesi è esatta cioè $\sigma^2 \leq \sigma_0^2$, allora

$$P\left(S^2 > \frac{\chi^2_{1-\alpha}(n-1)\sigma^2}{n-1}\right) \geq P\left(S^2 > \frac{\chi^2_{1-\alpha}(n-1)\sigma_0^2}{n-1}\right).$$

Una regione di rigetto relativa al valore di S^2 è quindi $D = \left]\frac{\chi^2_{1-\alpha}(n-1)\sigma_0^2}{n-1}, +\infty\right[$; il test si può anche effettuare calcolando $f = (n-1)\frac{S^2}{\sigma^2}$ e confrontandone il valore con i quantili di $\chi^2(n-1)$: in questo modo una regione di rigetto per f è data da $E = \left]\chi^2_{1-\alpha}(n-1), +\infty\right[$. Questo test si chiama *test di Fisher-Snedecor*.
Possiamo usarlo anche per decidere se accettare o no l'ipotesi bilaterale $\sigma^2 = \sigma_0^2$: in questo caso, se l'ipotesi è vera, allora

$$1-\alpha = P\left(\chi^2_{\frac{\alpha}{2}}(n-1) \leq (n-1)\frac{S^2}{\sigma_0^2} \leq \chi^2_{1-\frac{\alpha}{2}}(n-1)\right) =$$
$$= P\left(\frac{\chi^2_{\frac{\alpha}{2}}(n-1)\sigma_0^2}{n-1} \leq S^2 \leq \frac{\chi^2_{1-\frac{\alpha}{2}}(n-1)\sigma_0^2}{n-1}\right)$$

cioè

$$D = \left]0, \frac{\chi^2_{\frac{\alpha}{2}}(n-1)\sigma_0^2}{n-1}\right[\cup \left]\frac{\chi^2_{1-\frac{\alpha}{2}}(n-1)\sigma_0^2}{n-1}, +\infty\right[$$

è una regione di rigetto per S^2. Analogamente a quanto abbiamo fatto prima, possiamo calcolare il valore di $f = (n-1)\frac{S^2}{\sigma_0^2}$: in questo caso una regione di rigetto è $E = \left]0, \chi^2_{\frac{\alpha}{2}}(n-1)\right[\cup \left]\chi^2_{1-\frac{\alpha}{2}}(n-1), +\infty\right[$.

Confronto fra due varianze

Consideriamo due variabili aleatorie X e Y gaussiane, indipendenti e rispettivamente con leggi $N(\mu_X, \sigma_X^2)$ e$N(\mu_Y, \sigma_Y^2)$: vogliamo verificare l'ipotesi bilaterale $\sigma_X^2 = \sigma_Y^2$. Per farlo, sia $(X_1, \ldots, X_n)$ un campione di rango n di variabili indipendenti distribuite come X e $(Y_1, \ldots, Y_m)$ un campione di rango m di variabili aleatorie indipendenti e distribuite come Y; supponiamo inoltre che i due campioni siano fra loro indipendenti. Sappiamo, allora, per il corollario del teorema di

Cochran, che

$$W_X = (n-1)\frac{S_X^2}{\sigma_X^2}\chi^2(n-1) \text{ e } W_Y = (m-1)\frac{S_Y^2}{\sigma_Y^2}\chi^2(m-1)$$

e, inoltre, sono indipendenti in quanto lo sono i due campioni.

Definizione 3.5. Siano Z_n e Z_m due variabili aleatorie indipendenti con leggi $\chi^2(n)$ e $\chi^2(m)$ rispettivamente e definiamo $F(n,m) = \frac{Z_n}{Z_m}\frac{m}{n}$: allora la legge di F si chiama *legge di Fisher* con (n,m) gradi di libertà.

$F(n,m)$ è una variabile aleatoria a valori reali positivi con densità di probabilità data da

$$\gamma(x) = \frac{\Gamma\left(\frac{n}{2}+\frac{m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} (m+nx)^{-\frac{1}{2}(n+m)}$$

per $x > 0$.
Abbiamo ora

$$G = \frac{S_X^2}{\sigma_X^2}\frac{\sigma_Y^2}{S_Y^2} = \frac{W_X}{n-1}\frac{m-1}{W_Y} \sim F(n-1, m-1)$$

e, se l'ipotesi è vera cioè $\sigma_X^2 = \sigma_Y^2$, allora

$$\frac{S_X^2}{S_Y^2} \sim F(n-1, m-1).$$

Indichiamo con $F_\alpha(n-1, m-1)$ il quantile di livello α per $F(n-1, m-1)$ cioè

$$P(F(n-1, m-1)) \leq F_\alpha(n-1, m-1) = \alpha:$$

avremo quindi che G appartiene all'intervallo

$$I = \left[F_{\frac{\alpha}{2}}(n-1, m-1), F_{1-\frac{\alpha}{2}}(n-1, m-1)\right]$$

con probabilità $1-\alpha$. La regione di rifiuto dell'ipotesi è quindi il complementare di I cioè se G appartiene all'insieme $D = \mathbb{R} \setminus I$ respingeremo l'ipotesi $\sigma_X^2 = \sigma_Y^2$.

Osservazione 3.2. Nel test 3.2.1 abbiamo supposto che le varianze delle due variabili aleatorie fossero uguali per poterlo eseguire. Non avendo indicazioni sui valori delle varianze, potremmo essere portati a voler usare il test di Fisher per sapere se effettivamente sono uguali: solitamente però i test statistici vengono usati per rifiutare l'ipotesi e, quindi, non è sicuro che siano sufficientemente accurati se vogliamo non rifiutare l'ipotesi. Quindi, non possiamo sapere con certezza che le due varianze siano uguali se il test di Fisher fallisce ma possiamo comunque assumere con sufficiente certezza che, per lo meno, differiscano di poco.

3.3 Test del chi-quadrato

3.3.1 Confronto di una distribuzione con un valore assegnato

Sia X una variabile aleatoria a valori in un insieme finito $\{x_1, \ldots, x_m\}$. Supponiamo di conoscere le probabilità $\bar{p}_i = P(\{X = x_i\})$ e ripetiamo n volte un esperimento che assumiamo essere distribuito come X ovvero consideriamo $X_1, \ldots, X_n$ variabili aleatorie indipendenti e distribuite come X: vorremmo poter stimare se la variazione delle frequenze dei valori osservati dalle frequenze teoriche sono completamente casuali o se possiamo supporre che l'esperimento non fosse distribuito come X. Siano allora N_i il numero di volte in cui viene osservato il valore x_i cioè $N_i = \#\{k \leq n : X_k = x_i\}$ e $p_i = \frac{N_i}{n}$ le frequenze dei valori empirici. Consideriamo

$$T_n = \sum_{i=1}^{m} \frac{1}{np_i} (N_i - np_i)^2 = n \sum_{i=1}^{m} \frac{(p_i - \bar{p}_i)^2}{\bar{p}_i}. \tag{3.3}$$

Teorema 3.3.1. *Siano $X_1, X_2, \ldots$ variabili aleatorie indipendenti di legge $\theta = (p_1, \ldots, p_m)$: per $n \to +\infty$, T_n converge in legge verso $\chi^2(m-1)$.*

Per n sufficientemente grande, l'evento $D = \left\{T_n > \chi^2_{1-\alpha}(m-1)\right\}$ ha probabilità α ovvero D è una regione critica di livello α del test considerato.

3.3.2 Confronto fra due distribuzioni

Consideriamo due variabili aleatorie indipendenti X e Y che possono assumere i valori $\{a_1, a_2, \ldots, a_s\}$ con rispettive probabilità incognite $\left\{\theta'_1, \theta'_2, \ldots, \theta'_s\right\}$ e $\left\{\theta''_1, \theta''_2, \ldots, \theta''_s\right\}$. Ripetiamo per n volte, in condizioni di indipendenza, un esperimento che dà risultati per X con frequenze $\{n_1, n_2, \ldots, n_s\}$ e m volte, nuovamente in condizioni di indipendenza, per Y in modo da ottenere le frequenze $\{m_1, m_2, \ldots, m_s\}$: vogliamo decidere se la distribuzione di probabilità di X è la stessa di quella di Y con un livello di fiducia $\alpha \in]0, 1[$ fissato.
Il metodo che usiamo per verificare è simile a quello in cui si confronta una distribuzione sconosciuta con una teorica: definiamo una distribuzione di riferimento usando come frequenze per i vari valori $\{a_1, a_2, \ldots, a_n\}$ la media delle frequenze di$X$ e Y cioè la frequenza di a_i sarà data da $p_i = \frac{n_i + m_i}{n+m}$. Si costruisce poi una statistica T che esprime quanto le osservazioni si discostano dalla distribuzione

$(p_1, \ldots, p_S)$: sarà

$$T = \sum_{i=1}^{s} \frac{1}{np_i} (n_i - np_i)^2 + \sum_{i=1}^{s} \frac{1}{mp_i} (m_i - mp_i)^2 .$$

Si può vedere che, se n e m sono abbastanza grandi, allora $T \sim \chi^2 (s-1)$: una regione di rigetto al livello α per T sarà quindi $D = \left] \chi^2_{1-\alpha} (s-1), +\infty \right[$.

Osservazione 3.3. Calcolare il valore della statistica T tramite la sua definizione può non risultare troppo agevole. Vediamone una forma più comoda:

$$\begin{aligned}
T &= \sum_{i=1}^{s} \frac{1}{np_i} (n_i - np_i)^2 + \sum_{i=1}^{s} \frac{1}{mp_i} (m_i - mp_i)^2 && = \\
&= \sum_{i=1}^{s} \left(\frac{\left(n_i - \frac{n(n_i+m_i)}{m+m} \right)^2}{\frac{n(n_i+m_i)}{n+m}} + \frac{\left(m_i - \frac{m(n_i+m_i)}{m+m} \right)^2}{\frac{m(n_i+m_i)}{n+m}} \right) && = \\
&= \sum_{i=1}^{s} \frac{n+m}{n_i+m_i} \left(\frac{1}{n} \frac{(n_i m - m_i n)^2}{(n+m)^2} + \frac{1}{m} \frac{(m_i n - n_i m)^2}{(n+m)^2} \right) && = \\
&= \frac{1}{n+m} \sum_{i=1}^{s} \frac{(m_i n - n_i m)^2}{n_i+m_i} \left(\frac{1}{n} + \frac{1}{m} \right) && = \\
&= \frac{1}{nm} \sum_{i=1}^{s} \frac{(m_i n - n_i m)^2}{n_i+m_i} = nm \sum_{i=1}^{s} \frac{\left(\frac{n_i}{n} - \frac{m_i}{m} \right)^2}{n_i+m_i}
\end{aligned}$$

3.3.3 Indipendenza fra due distribuzioni

Siano X e Y due caratteri che vengono osservati su s individui, in modo che da ogni individuo si ottenga come risultato una coppia (X_i, Y_j) di valori aleatori. Supponiamo ora che X e Y possano assumere un numero finito di valori, in particolare $\{1, 2, \ldots, n\}$ per X e $\{1, 2, \ldots, m\}$ per Y. Possiamo allora costruire una *tabella di contingenza* con le frequenze assolute congiunte, ovvero il numero di volte $N_{i,j}$ in cui si presenta la coppia di caratteri (X_i, Y_j), e con le frequenze marginali N_i' e N_j'', cioè il numero di volte in cui si presentano $(X_i, \cdot)$ e $(\cdot, Y_j)$ rispettivamente. La tabella sarà quindi della forma

	Y_1	$\dots$	X_n	
X_1	$N_{1,1}$	$\dots$	$N_{1,n}$	$N_1^{'}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
X_n	$N_{n,1}$	$\dots$	$N_{n,n}$	N_n^2
	$N_1^{''}$	$\dots$	$N_n^{''}$	s

Osserviamo che, per ogni i, $N_i^{'} = N_{i,1} + \ldots + N_{i,n}$ e $N_i^{''} = N_{1,i} + \ldots + N_{n,i}$. Inoltre, potremmo anche voler costruire una tabella di contingenza con le frequenze relative congiunte e marginali cioè con $p_{i,j} = \frac{N_{i,j}}{s}$, $p_i^{'} = \frac{N_i^{'}}{s}$ e $p_j^{''} = \frac{N_j^{''}}{s}$.
Supponiamo allora che le osservazioni X_i siano indipendenti dalle osservazioni Y_i: il test che andremo a studiare vuole verificare se X e Y sono indipendenti.
Definiamo, per $1 \leq h \, leq \, n$, $p_h = P(X = h)$ e, per $1 \leq k \leq m$, $q_k = P(Y = k)$: questi valori sono ignoti ma possono essere stimati con le frequenze campionarie. Sappiamo che, se X e Y sono indipendenti, allora le probabilità congiunte $\pi_{h,k} = P(X = h \wedge Y = k)$ sono i prodotti delle probabilità marginali ovvero $\pi_{h,k} = p_h q_k$: chiaramente non conosciamo neanche il valore delle probabilità congiunte ma potremo, di nuovo, stimarlo con le frequenze campionarie. Gli stimatori per questi valori sono

$$\bar{p}_h = \frac{N_h^X}{n+m} = \frac{\text{numero di } i \text{ per cui } X_i = h}{n+m}$$

e

$$\bar{q}_k = \frac{N_k^Y}{n+m} = \frac{\text{numero di } i \text{ per cui } Y_i = k}{n+m}$$

cioè le frequenze relative marginali rispettivamente di X e Y e

$$\bar{\pi}_{h,k} = \frac{N_{h,k}}{n+m} = \frac{\text{numero di } i \text{ per cui } X_i = h \wedge Y_i = k}{n+m}$$

cioè le frequenze relative congiunte per (X, Y). Se X e Y sono indipendenti e il campione è abbastanza grande, vorremmo vedere che $\bar{p}_h \bar{q}_k - \bar{\pi}_{h,k}$ siano abbastanza piccole: si può infatti dimostrare che la variabile aleatoria

$$T = n \sum_{h=1}^{n} \sum_{k=1}^{m} \frac{\left(\bar{p}_h \bar{q}_k - \bar{\pi}_{h,k}\right)^2}{\bar{p}_h \bar{q}_k}$$

converge in legge a una variabile aleatoria di legge $\chi^2((n-1)(m-1))$. Possiamo così stabilire una regione critica per l'ipotesi di indipendenza di X e Y usando i quantili di questa variabile: dato che vogliamo avere valori piccoli di T per confermare l'indipendenza, la regione critica sarà $D = \left]\chi^2_{1-\alpha}((n-1)(m-1)), +\infty\right[$.

Capitolo 4

Alcune applicazioni

In questo capitolo parleremo di alcune applicazioni storiche che hanno portato allo sviluppo e al miglioramento dei test visti nei capitoli precedenti.

4.1 La controversia Mendel vs. Fisher

Fra il 1856 e il 1863, il monaco Gregor Mendel condusse una serie di esperimenti su piante di piselli odorosi che gli permisero di formulare alcune leggi sull'ereditarietà dei caratteri che segnarono l'inizio della genetica moderna. Nel 1936, lo statistico e biologo Ronald Fisher ricostruisce gli esperimenti di Mendel, ne analizza i risultati e si accorge che i risultati ottenuti dal monaco sono consistentemente troppo vicini a quello che ci si aspetta di ottenere: sospetta quindi che i dati pubblicati da Mendel siano stati falsificati per poter supportare la sua tesi.

4.1.1 Il lavoro di Mendel

Mendel osservò nei piselli odorosi sette tratti facilmente visibili nella struttura della pianta, del fiore, del baccello o del seme: si rese conto che ognuna di questi aspetti si presentava con due diverse variazioni. Riuscì a creare piante in cui queste caratteristiche rimanevano invariate da una generazione all'altra che chiamò linee pure. Secondo quello che viene riportato nel suo articolo del 1865, Mendel lavora con piante che differiscono per una sola caratteristica, ad esempio il colore del fiore, ma che sono uguali per quanto riguarda il resto: incrocia quindi due piante provenienti da diverse linee pure ed ottiene così quella che chiamò generazione filiale. Nota che in questa generazione tutte le piante hanno solo una delle due variazioni delle piante della linea pura e non, come ci si poteva aspettare, una

caratteristica che sia a metà fra quella delle due piante della linea pura: riesce così a ipotizzare che il carattere che si ripresenta nella generazione filiale sia dominante e che quello che è scomparso sia recessivo. Incrocia poi di nuovo le piante della generazione filiale e vede che, nella seconda generazione filiale, il carattere recessivo è riapparso con una proporzione abbastanza costante nei vari esperimenti.
Mendel inoltre si dedica allo studio dell'indipendenza delle caratteristiche da lui studiate: dopo aver sommariamente compreso come si comportavano i geni nel caso in cui solo una caratteristica differisce da una pianta all'altra, ha incrociato piante che differivano per due e per tre caratteristiche in modo da poter vedere se questi tratti vengono ereditate in maniera indipendente l'uno dall'altro.

4.1.2 Il contributo di Fisher

Il lavoro di Mendel viene pubblicato in tedesco nel 1866 e poi ristampato nel 1910 mentre una traduzione inglese fu pubblicata nel 1901 e ristampata con varie modifiche in più occasioni. Fisher, nel suo paper del 1936, considera il lavoro di Mendel come viene riportato da Bateson nel suo libro "Mendel's Principles of Heredity" del 1909.
Bateson può non essere considerato come un traduttore imparziale: possiamo infatti trovare nel suo lavoro l'origine di due leggende sul lavoro di Mendel. La prima riguarda il motivo per cui i contemporanei di Mendel non hanno preso in considerazione il suo lavoro: secondo Bateson, il lavoro di Darwin ha eclissato il lavoro di Mendel così tanto da nasconderlo all'opinione pubblica per più di 30 anni fino a che, nel 1900, Hugo de Vries e Carl Correns hanno verificato sperimentalmente e riscoperto i risultati di Mendel, ciascuno indipendentemente. La seconda riguarda invece il fatto che Mendel non fosse d'accordo con i risultati di Darwin e, anzi, abbia intrapreso i suoi esperimenti sui piselli odorosi proprio per cercare di confutare almeno in parte la teoria dell'evoluzione. Queste convinzioni attribuite a Mendel non si trovano nei suoi lavori, anzi è impossibile che lui avesse cominciato a lavorare coi piselli odorosi in reazione alle scoperte di Darwin semplicemente in quanto la pubblicazione dei lavori di Darwin fu successiva all'inizio degli esperimenti di Mendel e, inoltre, Mendel è a conoscenza del fatto che i suoi esperimenti e i suoi risultati potranno essere presi come base per ulteriori studi sul processo evolutivo. Bateson sottolinea però come il lavoro di Mendel, alla luce delle nuove conoscenze di genetica ottenute nei 30 anni che erano passati dalla pubblicazione del suo lavoro, sia aperto a interpretazioni sbagliate: Mendel è convinto di lavorare con coppie di piante che differiscono esclusivamente per una caratteristica e sono uguali per

quanto possibile a livello genetico per le altre sei. Come viene notato da Fisher, Mendel non poteva essere certo dell'indipendenza di queste sette caratteristiche prima di iniziare il suo lavoro e quindi è possibile che abbia usato meno coppie di piante di quante afferma nei suoi risultati: questo può però aver portato alla divisione dei caratteri ereditari delle piante in due diversi genotipi di cui non abbiamo riscontro nel lavoro di Mendel perché non siamo certi che abbia trascritto in modo assolutamente esatto i risultati dei suoi esperimenti. Possiamo dire che, in una certa misura, Fisher è prevenuto nei confronti del saggio di Mendel e, nonostante sottolinei più volte che il lavoro viene presentato ordinatamente e con molta chiarezza, è convinto che Mendel abbia falsificato in parte i risultati o omesso quelli che non aderivano completamente alle sue idee. Fisher non crede però che tutti gli esperimenti descritti da Mendel siano inventati: nell'esposizione del monaco mancano molti dettagli fra cui gli anni in cui le piante vengono coltivate e sembra che Mendel sia assolutamente sicuro dei suoi risultati in quanto non replica quasi mai gli esperimenti ma si limita ad analizzare la prima semina della seconda generazione filiale che, incidentalmente, produce risultati che confermano le sue teorie. Si potrebbe addossare la colpa di questa sua ingenuità al fatto che gli studi di Mendel all'università di Vienna non si concentrarono sulla probabilità e sulla statistica ma si era comunque occupato di studi di meteorologia che gli avrebbero fornito le conoscenze necessarie a dubitare dei risultati ottenuti e perlomeno a sottoporli a un altro giudizio.

4.1.3 Analisi degli esperimenti

Usando i dati del saggio di Mendel, possiamo effettuare un test del chi-quadro: i risultati degli esperimenti sono effettivamente delle variabili aleatorie $X_1, \ldots, X_n$ equidistribuite, indipendenti che vogliamo confrontare con delle probabilità teoriche note. In particolare, Fisher dubita di un particolare esperimento svolto da Mendel in quanto le proporzioni che Mendel ottiene sono molto vicine a proporzioni teoriche che secondo Fisher sono sbagliate. E' possibile però che Fisher non abbia interpretato correttamente l'esperimento.
Mendel vuole stabilire in quale proporzione le piante che mostrano il carattere dominante sono omozigoti, cioè con due geni dominanti uguali, o eterozigoti, cioè con un gene dominante e un gene recessivo: in questo tipo di esperimenti, Mendel ha prodotto 10 nuove piante da 100 piante che esibivano il carattere dominante, semplicemente incrociando ogni pianta con se stessa. In questo caso ci si potrebbe aspettare che $\frac{1}{3}$ delle nuove piante siano omozigoti mentre i restanti $\frac{2}{3}$ siano etero-

zigoti: in realtà, questa è la proporzione teorica che non tiene conto del fatto che una pianta eterozigote potrebbe produrre solo piante con caratteri dominanti ed essere classificata erroneamente come omozigote. Per dire che una pianta è eterozigote, cerchiamo una pianta che mostra il carattere recessivo fra quelle che essa ha prodotto: se consideriamo k piante prodotte da ciascuna di quelle che stiamo esaminando allora la proporzione piante eterozigoti su piante omozigoti sarà data da

$$\frac{\text{Eterozigoti}}{\text{Omozigoti}} = \frac{\frac{2}{3}\left[1-\left(\frac{3}{4}\right)^k\right]}{\frac{1}{3}+\frac{2}{3}\left(\frac{3}{4}\right)^k}.$$

Per $k = 10$ non otteniamo, in effetti, la proporzione che Mendel si poteva aspettare ma una proporzione di 1.7 piante eterozigoti per ogni pianta omozigote: possiamo ottenere una proporzione di circa $2 : 1$ per $k = 30$ che, secondo alcuni storici, è il numero effettivo di piante che Mendel ha piantato per essere sicuro che almeno 10 di esse potessero sopravvivere fino al momento in cui mostrano il carattere da lui studiato. Eseguiamo allora il test di chi-quadro confrontando i numeri ottenuti da Mendel sia con la proporzione $2 : 1$ che con la proporzione $1.7 : 1$. Riportiamo innanzitutto i dati ottenuti da Mendel in cinque diversi esperimenti su 100 piante:

	Omozigoti	Eterozigoti
Primo	36	64
Secondo	29	71
Terzo	40	60
Quarto	33	67
Quinto	28	72

Osserviamo che fra questi, il terzo esperimento è quello in cui la proporzione è più lontana da quella che ci aspettavamo: anche Mendel se ne accorge tanto da fare una seconda prova da cui ottiene 35 piante omozigoti e 65 eterozigoti. In tutti i casi, il numero di piante omozigoti che si aspetta di ottenere Mendel è, arrotondato all'unità, 33 mentre, il numero di piante omozigoti che Fisher ritiene sia opportuno avere è 37. Il risultato del test di chi-quadro per ogni esperimento viene calcolato con la seguente formula

$$T = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2}$$

dove o_i indica il numero di piante osservate e e_i il numero atteso di piante, arrotondato all'unità. Osserviamo che questa formula può essere ottenuta da 3.3

semplicemente considerando che $m = 2$ e $np_i = e_i$ per come è definito e_i. I risultati sono riassunti nella tabella dove con "Primo", "Secondo", ... indichiamo i cinque esperimenti che Mendel esegue con lo stesso metodo, con "Sesto" indichiamo la prova che Mendel replica per il terzo esperimento che ritiene troppo lontano dalle sue aspettative e con "Totale" il valore del test di chi-quadro che si ottiene sommando tutti i valori osservati:

	Primo	Secondo	Terzo	Quarto	Quinto	Sesto	Totale
Mendel	0.407	0.724	2.216	0	1.131	0.181	0.008
Fisher	0.043	2.746	0.386	0.686	3.475	0.172	3.302

e le rispettive probabilità, calcolate con la distribuzione $\chi^2(1)$, sono

	Primo	Secondo	Terzo	Quarto	Quinto	Sesto	Totale
Mendel	0.5235	0.3949	0.1366	1	0.2876	0.6706	0.931
Fisher	0.8359	0.0975	0.5344	0.4074	0.0623	0.6787	0.069

Osserviamo che, se consideriamo come livello di significatività $\alpha = 5\%$, allora tutti gli esperimenti passano il test: quelli che usano la proporzione teorica di Mendel, però, sono tutti a un livello più alto rispetto a quelli che usano la proporzione proposta da Fisher che, per uno solo dei test, è quasi al livello del 6%. Non possiamo quindi dire conclusivamente che Mendel abbia falsificato i suoi risultati solo analizzando questo esperimento. Fisher è comunque convinto che Mendel (o un suo assistente) abbiano modificato i dati ottenuti perché aderissero alle aspettative *sbagliate* di Mendel.
Osserviamo inoltre che in [5] viene proposta un'ulteriore proporzione per le piante eterozigoti e omozigoti che, però, richiede la conoscenza della probabilità che il seme piantato riesca a crescere fino a diventare una pianta adulta che Mendel non calcola e di cui non si riesce a ricavare una stima sufficientemente adeguata dai suoi scritti.

4.2 Lo studio dell'orzo

Verso la fine del 1800, l'azienda a conduzione familiare Guinnes viene venduta per sei milioni di sterline e diventa una società pubblica: sarà una delle prime a introdurre dei processi per il controllo qualità dei prodotti, dall'inizio alla fine del ciclo di produzione della birra.
Sotto la guida di Robert Guinness, comincia l'assunzione di chimici, biologici e matematici a cui viene insegnata la professione di birraio, fino ad allora un'arte

tramandata meccanicamente da maestro a studente: lo scopo è quello di studiare in modo scientifico il luppolo e l'orzo, dalla loro nascita, fino alla fermentazione, passando anche per i processi di essiccazione e maltatura. Fra i nuovi assunti, troviamo William S. Gosset: chimico al quale si rivolgono i colleghi con i risultati degli esperimenti sulle coltivazioni di orzo per avere indicazioni su quale piantagione produca i migliori risultati. Per non rendere noto alla concorrenza che la Guinnes aveva alle loro dipendenze uno statistico, gli fu imposto di pubblicare i suoi lavori sotto lo pseudonimo di Student.
Durante i suoi anni di lavoro alla Guinness, Student scrive molti trattati su come calcolare gli errori di media e varianza a partire da campioni sperimentali: è particolarmente interessato a osservazioni in cui il numero di ripetizioni è basso in quanto il suo studio dell'orzo non gli permette di avere un numero elevato di campioni come accadeva invece per gli altri statistici. Sviluppa in parte il test che ha il suo nome: Student studia la funzione

$$z = \frac{\overline{X} - \mu}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}}}$$

mentre è Fisher, nel 1925, a proporgli la modifica $t = z\sqrt{n-1}$ che porta alla versione moderna. Per di più, suggerisce anche l'idea di usare $n-1$ al posto di n nel calcolo di medie e varianze sperimentali in modo da compensare nel caso in cui il campione sia poco numeroso. Fisher contribuisce anche alle dimostrazioni matematiche dei risultati di Student: Gosset infatti sa di aver lasciato parti incomplete nelle sue dimostrazioni ma gli mancano alcune conoscenze per poterle finire. Inoltre i due si scriveranno lettere per tutta la loro vita dibattendo su argomenti di statistica: Gosset è a favore di un metodo in cui gli esperimenti da analizzare siano bilanciati in base alle conoscenze pregresse mentre Fisher preferisce la pura casualità nella loro preparazione e per molti anni pubblicheranno saggi per avvalorare le loro ipotesi.

4.2.1 Analisi di un esperimento

Nel suo lavoro "On the probable error of a mean", Student spiega che prima di riuscire a trovare una regola per il suo test e a generalizzare il metodo con cui svolgerlo ha passato molto tempo ad analizzare esempi pratici: uno di questi riguarda gli effetti sul sonno di due diversi isomeri ottici di una stessa molecola. I dati riportati da Student sono i seguenti:

Paziente	Isomero 1	Isomero 2	Differenza 2-1
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-0.1	-0.1	+0
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0	+4.6	+4.6
10	+2.0	+3.4	+1.4

dove i valori indicano l'aumento o la diminuzione delle ore di sonno del paziente dopo che è stato somministrato l'isomero corrispondente.
Student calcola allora il valore del test z per l'isomero 1, per l'isomero 2 e per la serie di risultati dati dalla differenza di valori fra l'isomero 1 e 2: usa quest'ultima serie di valori per giustificare il fatto che l'isomero 2 sia un sonnifero migliore dell'isomero 1. Osserviamo che calcola $z = \frac{x}{s}$ dove s è lo scarto quadratico del campione e x è la differenza fra la media dell'esperimento μ e il valore con cui la si vuole confrontare μ_0. Per tutte e tre le serie si considera $\mu_0 = 0$: per la serie data dalla differenza dei valori si usa 0 si vuole vedere se l'effetto dei due farmaci è uguale mentre per le due serie di dati dei singoli farmaci si usa 0 per confrontare la media ottenuta nell'esperimento con le notti in cui non viene assunto il farmaco che si può assumere abbiano come media di differenza di ore di sonno nulla. Per farlo calcola la media e lo scarto quadratico medio delle tre serie ottenendo

	Isomero 1	Isomero 2	Differenza 2-1
Media μ	+0.75	+2.33	+1.58
Scarto quadratico σ	1.70	1.90	1.17

Ha quindi $z_1 = \frac{0.75}{1.70} = 0.44$ che equivale a una probabilità di 0.887 di avere un aumento del numero di ore di sonno per il primo isomero, $z_2 = \frac{2.33}{1.90} = 1.23$ che equivale a una probabilità di 0.9974 per l'isomero due e $z_3 = \frac{1.58}{1.17} = 1.35$ cioè 0.9985 di probabilità per la serie della differenza. Student conclude quindi quest'esempio dicendo che questi risultati sono chiaramente in favore dell'aumento del numero di ore di sonno provocato dall'isomero numero 2.
La prima osservazione riguardo l'esperimento appena citato è che la differenza di ore di sonno del paziente 9 si discosta molto dalla media delle differenze rispetto agli altri: senza sapere come è stato ottenuto il dato sull'aumento delle ore di sonno,

sarebbe stato opportuno escludere questa osservazione. Modificando il campione in questo modo otteniamo $z_1^{'} = 1.68$ che corrisponde a circa 0.99903 come probabilità, $z_2^{'} = 1.74$ ovvero circa 0.99933 come probabilità e $z_3^{'} = 0.59$ cioè una probabilità di circa 0.936: il dato per l'isomero 1 non è quindi più piccolo del dato per l'isomero 2 ma per la loro differenza otteniamo comunque una probabilità significativamente alta anche se comunque più bassa rispetto al caso precedente. Inoltre i dati citati da Student non trovano corrispondenza esatta con quelli dell'articolo da cui li estrapola: nell'originale infatti vengono usati tre farmaci anzichè due e l'aumento delle ore di sonno viene calcolato dopo aver osservato notti in cui non vengono somministrati famaci. Nell'articolo fonte dei dati possiamo leggere la metodica con cui l'esperimento è stato svolto: ogni paziente viene osservato per un certo numero di notti e vengono riportati i valori medi delle ore di sonno per notte; inoltre non è specificato esattamente l'ordine con cui vengono somministrati i farmaci e neanche se ci sono giorni di riposo dai medicinali che permettano di smaltirne l'effetto. In [9] viene ricostruita una possibile sequenza di farmaci somministrati per giorno e si può notare come alcune notti che servono come controllo sono effettivamente fra due notti in cui viene somministrato un farmaco e quindi le misurazioni non sono completamente affidabili. Per di più, le ore di sonno riportate sono una media ottenuta dai risultati di più giorni: non è chiaro perché si analizzi la media anziché i valori dei singoli giorni nonostante questi valori siano disponibili.

4.2.2 Esperimenti bilanciati

Student è famoso per gli studi con cui ha cercato di stabilire quale varietà di luppolo o orzo fosse superiore e quale metodo di maltatura consente di lavorarli per ricavare il prodotto migliore per la fermentazione: oltre ad applicare il test che porta il suo nome, si impegnò a cercare di stabilire un metodo per gli esperimenti con cui avere risultati che non sono influenzati da fattori casuali ma solo dalla caratteristica che si vuole studiare.
Nel caso della coltivazione del orzo, lo scopo degli esperimenti svolti durante l'impiego alla Guinnes era quello di trovare una varietà che potesse dare il raccolto maggiore: chiaramente questo è influenzato sia dalla produttività della pianta che da fattori esterni come la fertilità del campo, l'attacco da parte di uccelli e conigli, il meteo imprevedibile, ... Viene così proposta sia l'idea di condurre gli esperimenti in serra che in campo aperto: in serra i problemi derivanti dalla variabilità del meteo e da animali erranti sono ridotti al minimo ma non si possono usare gli stessi strumenti per la coltivazione in campo aperto, cosa che comporta l'obbligo

di spese da parte degli sperimentatori per ottenere spazi e strumenti necessari. Student analizza quindi un esempio in cui la varietà "Plumage" coltivata in serra di fianco alla qualità "Archer" produce più di quest'ultima: il risultato è fuori dall'ordinario in quanto la varietà "Archer" all'epoca era notevolmente superiore a varietà affini a quella "Plumage" ma il fatto che i due tipi di orzo siano cresciuti troppo vicini ha fatto sì che il "Plumage" abbia tolto spazio e luce all'"Archer". Il lavoro in serra viene così definitivamente scartato come utile da Student in quanto non produce sempre risultati attendibili a causa della mancanza di spazio, è più dispendioso e va comunque validato con una prova in campo aperto.
Per evitare questo tipo di problemi, Student descrive un modello di semina che permette di coltivare a file alternate in campo aperto e individua tre modi per stimare la varianza dei dati ottenuti. Il primo modo consiste nel considerare le coppie di file alternate come osservazioni indipendenti e di stimare la varianza della differenza di raccolto delle due qualità di orzo con la quantità

$$s_1^2 = \frac{\sum \left(A_i - Bi - \overline{A-B}\right)^2}{2n\,(2n-1)}$$

dove A_i e B_i indicano rispettivamente la quantità d'orzo raccolta nella i-esima fila delle due varietà, $\overline{A-B} = \frac{\sum(A_i - B_i)}{2n}$ è una stima per la media e $2n$ è il numero totale di coppie. Nel secondo metodo, l'unità base è data da un insieme di quattro file consecutive che vengono considerate come osservazioni indipendenti l'una dall'altra: viene definita la quantità

$$\Delta_i = A_i + A_{i+1} - B_i - Bi + 1$$

e la varianza della differenza di raccolto fra le due varietà di orzo è stimata con

$$s_2^2 = \frac{\sum \left(\Delta_i - \overline{\Delta}\right)^2}{n\,(n-1)}$$

dove $\overline{\Delta} = \frac{\sum \Delta_i}{n}$ è una stima per la media. Il terzo modo prevede invece una suddivisione ulteriore delle file in m sezioni regolari e di utilizzare queste come osservazioni indipendenti per calcolare la varianza: Student afferma che quest'ultimo metodo porta ad ottenere un valore della varianza più basso anche se i dati sono gli stessi del primo caso. Inoltre dei due metodi rimanenti, il primo è considerato meno attendibile perché dà dei valori che sovrastimano il reale.
Usiamo ora un esempio tratto da [11] per vedere come, effettivamente, il secondo metodo dia risultati migliori: con questo non vogliamo dire che esso dia valori

di probabilità più alti in generale. Nell'articolo citato vengono studiati i risultati della coltivazione di due diverse varietà d'orzo con il metodo a file alterne proposto da Student; vengono poi raccolti i dati di 27 coppie di file alternate divise ognuna divisa in 10 sezioni: per applicare il secondo metodo sono stati sommati i valori di due file consecutive della stessa varietà mentre, per applicare il terzo metodo, sono state escluse dal conteggio la prima e l'ultima sezione di ogni fila. Calcolando il valore della differenza delle coppie, dei gruppi di quattro file e delle sezioni e poi media, varianza, valore della statistica t e della relativa probabilità, si ottiene

	Media	Varianza	t	P
Primo	4.02	43.91	0.093	0.9265
Secondo	8.05	51.48	0.141	0.8905
Terzo	-3.8	36.63	0.097	0.9238

Possiamo vedere che il valore della probabilità ottenuto con il secondo metodo è molto più basso degli altri due che, invece, come previsto da Student, sono quasi uguali: questo non vuol dire che usare il secondo metodo per calcolare la varianza dia risultati peggiori, semplicemente con questo metodo la differenza di raccolto fra le due varietà non viene considerata notevole e si richiedono studi più precisi. In effetti la differenza totale di raccolto delle due varietà è solo dell' 1% del raccolto totale della varietà più produttiva su un campo di più di 2000 metri quadrati che potrebbe essere attribuita a una non uniformità di agenti positivi o negativi nel campo.

Bibliografia

[1] Baldi P., *Calcolo delle probabilità e statistica*, McGraw-Hill, prima edizione, Milano, 1992

[2] Baldi P., *Introduzione alla probabilità con elementi di statistica*, McGraw-Hill, seconda edizione, Milano, 2012

[3] Box J. F., *Guinness, Gosset, Fisher, and Small Samples* in "Statistical Science" vol. 2, n. 1, 1987, pp. 45-52

[4] Fisher R. A., *Has Mendel's work been rediscovered?* in "Annals of Science" vol. 1, Londra, 1936, pp. 115-137

[5] Hartl D. L., Fairbanks D. J., *Mud sticks: on the alleged falsification of Mendel's data* in "Genetics" vol. 175, 2007, pp 975-979

[6] Mendel G., *Versuche über Plflanzenhybriden* letto a Brno, 1865, trad. Bateson W., *Experiments on plant hybridization*, 1901

[7] Neyman J. e Pearson E. S., nota in Student, *Comparison between balanced and random arrangements of field plots*, pp. 380-388

[8] Pierce B., *Genetics: a conceptual approach*, Freeman, quarta edizione, New York, 2005

[9] Preece D. A., *t is for trouble (and textbooks): a critique of some examples of the paired-samples t-test* in "The statistician" vol. 31, n.2, 1982, pp. 169-195

[10] Student, *The probable error of a mean* in "Biometrika" vol. 6, n. 1, Oxford University Press, 1908, pp. 1-25

[11] Student, *On testing varieties of cereals* in "biometrika" vol. 15, n. 3, Oxford University Press, 1923, pp. 271-293

[12] Student, *Comparison between balanced and random arrangements of field plots* in "Biometrika vol. 29, n. 3/4, 1938, pp. 363-378

[13] Zabell S. L., *On Student's 1908 article "The probable error of a mean"* in "Journal of American Statistical Association" vol. 103, n. 481, 2008