ELEMENTI DI STATISTICA INFERENZIALE
(*versione preliminare*)

Barbara Torti, Mario Abundo

Indice

Capitolo 1

Richiami di Probabilità e Statistica

1.1 Spazi di Probabilità e variabili aleatorie

Uno *spazio di probabilità* è una terna $(\Omega, \mathcal{F}, P)$, dove Ω è un insieme, $\mathcal{F}$ è una σ-algebra di parti di Ω, e P è una misura di probabilità su $(\Omega, \mathcal{F})$.
Una *variabile aleatoria reale* è un' applicazione $X : \Omega \to \mathbb{R}$ tale che

$$\{\omega : \quad X(\omega) \leq t\} = X^{-1}(-\infty, t] \in \mathcal{F} \text{ per ogni } t \in \mathbb{R}$$

Osserviamo che questa relazione esprime il fatto che, per poter calcolare le probabilità di insiemi espressi come funzioni di un esperimento aleatorio, tali insiemi devono essere in $\mathcal{F}$, ovvero devono essere degli **eventi** (ad esempio, il numero di teste su n lanci di moneta...).

Le variabili aleatorie, quando utilizzabili, hanno la notevole proprietà di trasferire il calcolo delle misure di probabilità di interesse da $(\Omega, \mathcal{F}, P)$ a $\mathbb{R}$. Questo in generale rappresenta un vantaggio, poiché lo spazio $(\Omega, \mathcal{F}, P)$ potrebbe essere molto più complicato e, nella stragrande maggioranza dei casi, di dimensione strettamente maggiore di 1.
Indichiamo con $\mathcal{B}$ la σ-algebra su $\mathbb{R}$ generata dagli intervalli. Una variabile aleatoria X induce quindi una misura di probabilità P_X su $(\mathbb{R}, \mathcal{B})$ tramite l'applicazione

$$P_X : \mathcal{B} \to [0, 1]$$

tale che $P_X((a, b]) = P(\omega : \ X(\omega) \in (a, b])$ per ogni $a, b \in \mathbb{R}$.

Chiamiamo tale applicazione *legge della variabile aleatoria* X.

Ancora, la legge è una funzione di insieme, e dunque un oggetto non facile da trattare, essendo il suo dominio una σ-algebra. È tuttavia possibile caratterizzare univocamente la legge di una variabile aleatoria X tramite la sua *funzione di ripartizione* F_X definita come l'applicazione

$$F_X : \mathbb{R} \to [0, 1]$$

tale che $F_X(t) = P(\omega : \ X(\omega) \leq t)$ per ogni $t \in \mathbb{R}$.

La funzione di ripartizione di una variabile aleatoria gode delle seguenti proprietà

1 $\lim_{t \to -\infty} F_X(t) = 0$, $\lim_{t \to +\infty} F_X(t) = 1$;

2 monotonia: $F_X(s) \leq F_X(t)$ per ogni $s \leq t$;

3 continuità da destra: $\lim_{h \to +\infty} F_X(t + \frac{1}{h}) = F_X(t)$.

Le variabili aleatorie che incontreremo possono essere classificate in due tipi, a seconda della cardinalità dell'insieme di valori che possono assumere.

Una variabile aleatoria X è *discreta* se ha un codominio $Im(X)$ finito o numerabile. È invece *continua* se il suo codominio è un sottoinsieme continuo di $\mathbb{R}$.

1.2 Teoremi limite

LEGGE DEI GRANDI NUMERI
TEOREMA DEL LIMITE CENTRALE

1.3 Leggi gamma, normale, chi-quadrato, t di student, Fischer

1.3.1 Legge gamma

Definizione 1.1. *Si dice funzione gamma, l'applicazione* $\Gamma : \mathbb{R}^+ \to \mathbb{R}^+$ *definita come segue*

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

Proprietà della funzione Γ:

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

2 $\Gamma(\alpha + 1) = \int_0^\infty x^\alpha e^{-x} dx = \alpha \int_0^\infty x^{\alpha-1} e^{-x} dx = \alpha\Gamma(\alpha)$

3 Se $n \in \mathbb{N}$, dalle proprietà precedenti segue facilmente $\Gamma(n) = (n-1)!$

4 $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Una v.a. X segue una ***legge gamma di parametri*** α ***e*** λ con $\alpha, \lambda \in \mathbb{R}^+$ ($X \sim \Gamma(\alpha, \lambda)$) se ha densità pari a

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{se } \ x \geq 0 \\ 0 & \text{altrove} \end{cases}$$

Utilizzando le proprietà della funzione Γ determiniamo media e varianza di X

$$E[X] = \int_0^{+\infty} x \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} = \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} =$$
$$= \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} \underset{y=\lambda x}{=} \frac{1}{\lambda\Gamma(\alpha)} \int_0^{+\infty} y^\alpha e^{-y} dy = \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} = \frac{\alpha}{\lambda}$$

[1] Tranne in alcuni casi, come ad esempio quando $\alpha \in \mathbb{N}$, questo integrale non ha una primitiva semplice. Pertanto la funzione gamma resta espressa in questo modo.

$$E[X^2] = \int_0^{+\infty} x^2 \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} = \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\lambda x} =$$
$$= \underset{y=\lambda x}{=} \frac{1}{\lambda^2 \Gamma(\alpha)} \int_0^{+\infty} y^{\alpha+1} e^{-y} dy = \frac{\Gamma(\alpha+2)}{\lambda^2 \Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\lambda^2}$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}$$

Importante nelle applicazioni che faremo è il seguente risultato:

Teorema 1.2. *Siano date n v.a.* $X_1, ..., X_n$ *indipendenti,* $X_i \sim \Gamma(\alpha_i, \lambda),\ \alpha_i > 0$. *Allora* $\boxed{\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n \alpha_i, \lambda)}$.

Una variabile $X \sim \Gamma(1, \lambda)$ è anche detta ***esponenziale di parametro*** λ e si indica solitamente con $Exp(\lambda)$.

1.3.2 Legge normale

Una v.a. X segue una ***legge normale o gaussiana standard*** ($X \sim N(0,1)$) se ha densità pari a

$$f_X : \mathbb{R} \to \mathbb{R} \ \ t.c. \ \ f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Proprietà:

1 **Simmetria**: $-X \sim X$. Infatti

$$F_{-X}(t) = P(-X \leq t) = P(X \geq -t) = 1 - F_X(-t)$$

e derivando

$$f_{-X}(t) = -f_X(-t)(-1) = f_X(-t) = f_X(t)$$

dove l'ultima uguaglianza deriva dalla parità della densità normale standard. Si traduce nella formula operativa $\Phi(x) = 1 - \Phi(-x)$, avendo indicato, come consuetudine, con Φ la funzione di ripartizione di una variabile aleatoria gaussiana standard.

2 **Quantili**: la proprietà precedente implica $\phi_\alpha = -\phi_{1-\alpha}$, dove, per ogni $\alpha \in (0,1)$ con ${\phi_\alpha}^2$ si indica il *quantile di ordine* α della legge normale, ovvero la soluzione dell'equazione

$$P(X \leq \phi_\alpha) = \alpha$$

3 **media e varianza**: coincidono con i parametri $(0,1)$.

4 $\boxed{Y = \sigma X + \mu}$, con $\sigma \in \mathbb{R}^+$ è una trasformazione lineare di X che genera una v.a. ***normale o gaussiana di parametri*** μ, σ^2 ($Y \sim N(\mu, \sigma^2)$) di densità pari a

$$f_Y : \mathbb{R} \to \mathbb{R}^+ \ \ t.c. \ \ f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

[2] In alcuni libri (ad esempio in [8]) con ϕ_α si indica la soluzione dell'equazione $P(X > \phi_\alpha) = \alpha$, e quindi il quantile di ordine $1 - \alpha$.

Esempio 1.3. *Sia $Y \sim N(\mu, \sigma^2)$. Mostrare che $\frac{Y-\mu}{\sigma} \sim N(0,1)$*[3].

5 Date n v.a. $X_1, ..., X_n$ indipendenti, $X_i \sim N(\mu_i, \sigma_i^2)$ allora $\boxed{\sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)}$.

Esempio 1.4. *[2] p. 154 Es. 3.7*
Siano X ed Y due variabili aleatorie indipendenti e gaussiane standard. Calcolare $P(X > Y)$ e $P(X > Y + \frac{1}{2})$.

Esempio 1.5. *[8] p. 174-5.5.2 (INTRODUCE AI TEST D' IPOTESI)*
Un messaggio binario ("0" oppure "1") viene trasmesso da una sorgente A ad un ricevente B tramite un canale elettrico, inviando un segnale di 2 volt se il messaggio era "1", -2 volt se il messaggio era "0". A causa di disturbi sul canale, il ricevente B riceve un segnale pari a $R = x + N$, dove x può assumere il valore 2 o -2 ed $N \sim N(0,1)$. Il ricevente decodifica il segnale con "1" se $R \geq 0.5$, "0" se $R < 0.5$. Calcolare le probabilità di decodificare erroneamente il segnale.

1.3.3 Legge Chi-quadro con n gradi di libertà

Siano $X_1, ..., X_n$ n v.a. normali standard e indipendenti. La variabile aleatoria

$$X = X_1^2 + ... + X_n^2$$

è nota come v.a. ***chi-quadrato con n gradi di libertà*** ($X \sim \chi_n^2$) e la sua densità ha l' espressione

$$f_X : \mathbb{R} \to \mathbb{R} \;\; t.c. \;\; f_X(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \mathbb{1}(x \geq 0).$$

1 **Legame con le leggi** $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$: La legge Chi-quadro con n gradi di libertà è un caso particolare di legge Gamma, corrispondente alla scelta dei parametri indicata.

2 Date m v.a. $X_1, ..., X_m$ indipendenti, $X_i \sim \chi_{n_i}^2$ allora $\boxed{\sum_{i=1}^m X_i \sim \chi^2_{\sum_{i=1}^m n_i}}$. (Dim: caso particolare di leggi Gamma)

3 **Quantili**: con $\chi^2_{\alpha,n}$ si indica il *quantile di ordine* α della legge Chi-quadro con n gradi di libertà, ovvero la soluzione dell'equazione

$$P(X \leq \chi^2_{\alpha,n}) = \alpha$$

Esempio 1.6. *[8] p. 189 5.8.3- (OSSERVA CHE BISOGNA SCRIVERE L' EVENTO IN TERMINI DI DIS. DI V.A. TABULATE)*
Per localizzare un oggetto nello spazio tridimensionale si effettua una misurazione che porta un errore sperimentale in ciascuna delle tre direzioni che segue una legge $N(0,4)$. Supponendo i tre errori lungo le tre diverse direzioni indipendenti tra loro, calcolare la probabilità che la distanza tra la posizione reale e quella misurata sia maggiore di 3.

[3] In generale, applicando ad una qualunque variabile aleatoria la trasformazione lineare ottenuta sottraendo la media e dividendo per la deviazione standard si ottiene una variabile aleatoria con media nulla e varianza pari ad 1. Questa operazione viene chiamata ***standardizzazione***.

1.3.4 Legge t di Student con n gradi di libertà

Siano date le v.a. indipendenti $Z \sim N(0,1)$, $C_n \sim \chi^2_n$. Definiamo la v.a. T tramite l'equazione

$$T = \frac{Z}{\sqrt{\frac{C_n}{n}}}$$

La variabile aleatoria T prende il nome di v.a. ***t di Student con n gradi di libertà*** ($T \sim t_n$) e la sua densità ha l' espressione

$$f_T : \mathbb{R} \to \mathbb{R} \;\; t.c. \;\; f_T(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})} \left[1 + \frac{x^2}{n}\right]^{-\frac{(n+1)}{2}} .$$

1 Simmetria: $T \sim -T$. Si dimostra attraverso i seguenti passaggi:

(a) $(Z, C_n) \sim (-Z, C_n)$(indipendenza + simmetria della normale standard)

(b) $g : \mathbb{R}^2 \to \mathbb{R}$ allora $g(Z, C_n) \sim g(-Z, C_n)$

(c) posto $g(z, c) = \frac{z}{\sqrt{\frac{c}{n}}}$ l'osservazione precedente implica

$$P(T \leq x) = P\left(\frac{Z}{\sqrt{\frac{C_n}{n}}} \leq x\right) = P\left(-\frac{Z}{\sqrt{\frac{C_n}{n}}} \leq x\right) = P(-T \leq x)$$

2 Quantili: la proprietà precedente implica $t_{\alpha,n} = -t_{1-\alpha,n}$, dove, per ogni $\alpha \in (0,1)$ con $t_{\alpha,n}$ si indicano le quantità definite tramite l' equazione

$$P(T \leq t_{\alpha,n}) = \alpha$$

3 $\boxed{P(T \leq x) \underset{n\to\infty}{\to} \Phi(x)}$ - Spiegazione euristica: $\frac{C_n}{n} = \frac{X_1^2 + ... + X_n^2}{n}$ con $X_1, ..., X_n$ n v.a. normali standard e indipendenti. Quindi, per la legge dei grandi numeri $\frac{C_n}{n} \to 1$ e quindi, quando n diventa molto grande, T avrà circa lo stesso comportamento di Z.

1.3.5 Legge F di Fischer con n ed m gradi di libertà

Siano date le v.a. indipendenti $C_n \sim \chi^2_n$, $C_m \sim \chi^2_m$. Definiamo la v.a. F tramite l'equazione

$$F = \frac{\frac{C_n}{n}}{\frac{C_m}{m}}$$

La variabile aleatoria F prende il nome di v.a. ***F di Fischer con n ed m gradi di libertà*** ($F \sim F_{n,m}$) e la sua densità ha l' espressione

$$f_F : \mathbb{R} \to \mathbb{R} \;\; t.c. \;\; f_F(x) = \left(\frac{n}{m}\right)^{\frac{n}{2}} \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \frac{x^{\frac{n}{2}-1}}{\left(1 + \frac{n}{m}x\right)^{\frac{n+m}{2}}} \mathbb{1}(x \geq 0).$$

Questa legge rivelerà la sua utilità quando affronteremo l' analisi della varianza.

1 **Quantili**: con $F_{\alpha,n,m}$ si indicano le quantità definite tramite l' equazione

$$P(F \leq F_{\alpha,n,m}) = \alpha,$$

ovvero i quantili di ordine α.

2 **Calcolo dei quantili di ordine** $\alpha \geq 0.5$: le tavole della F Fischer, per vari valori di n ed m, sono tabulati solo per $\alpha \leq 0.5$. Quando $\alpha \geq 0.5$ si usa la relazione derivante dalle seguenti trasformazioni:

$$\begin{aligned} P\left(F \leq F_{\alpha,n,m}\right) =& P\left(\frac{\frac{C_n}{n}}{\frac{C_m}{m}} \leq F_{\alpha,n,m}\right) = P\left(\frac{\frac{C_m}{m}}{\frac{C_n}{n}} \geq \frac{1}{F_{\alpha,n,m}}\right) = \\ =& 1 - P\left(\frac{\frac{C_m}{m}}{\frac{C_n}{n}} \leq \frac{1}{F_{\alpha,n,m}}\right) = \alpha \end{aligned}$$

ovvero $P\left(\frac{\frac{C_m}{m}}{\frac{C_n}{n}} \leq \frac{1}{F_{\alpha,n,m}}\right) = 1 - \alpha$ da cui si evince la relazione operativa

$$\boxed{F_{1-\alpha,m,n} = \frac{1}{F_{\alpha,n,m}}}$$

1.4 Modelli statistici e statistiche campionarie

Fissiamo una volta per tutte uno spazio di probabilità $(\Omega, \mathcal{F}, P)$ su cui, di volta in volta, penseremo realizzate le variabili aleatorie di interesse[4].
Introduciamo il concetto di campione aleatorio. Supponiamo di voler studiare un particolare carattere (in genere numerico) di un insieme di elementi (***popolazione***). Lo scopo della statistica è quello di produrre informazioni circa il carattere in esame a partire dalla osservazione di un sottoinsieme di elementi della popolazione (***campione***). Nell' ambito della Statistica Matematica il carattere è descritto da una variabile aleatoria X la cui legge F non è completamente nota. Produrre informazioni circa il carattere significa, in questo contesto, descrivere la legge F attraverso l' osservazione di tale carattere su un campione di ampiezza n, ovvero attraverso l' osservazione di un vettore di variabili aleatorie $(X_1, ..., X_n)$, essendo X_i la variabile aleatoria che descrive il carattere in esame dell' i-simo elemento del campione. Formalmente:

Definizione 1.7. *Un* ***campione aleatorio*** *di ampiezza n è una sequenza di variabili aleatorie $(X_1, ..., X_n)$ indipendenti aventi legge comune F.*

Il problema in esame è quello di usare le osservazioni per descrivere la legge incognita F. Sono possibili 2 casi:

Caso 1 La forma funzionale della legge F è nota, dipende da un vettore θ di parametri reali incogniti.

[4]Nei casi che analizzeremo in questo corso, l' esistenza di uno spazio di probabilità su cui realizzare le variabili aleatorie in esame sarà sempre verificata.

Caso 2 La forma funzionale della legge F non è nota.

Nel primo caso produrre informazioni su F equivale a produrre informazioni sul vettore dei parametri incogniti θ, e questo è un problema di *inferenza parametrica*. Nel secondo caso si ha invece un problema (di più difficile gestione) di *inferenza non parametrica*.
In questo corso studieremo principalmente problemi di inferenza parametrica. Diamo qualche definizione di carattere generale;

Definizione 1.8. *Un **modello statistico parametrico** è una famiglia di leggi*

$$\{f_X(x;\theta), \quad \theta \in \Theta\}$$

dove θ è un parametro o un vettore di parametri che assumono valore in un sottoinsieme Θ (eventualmente infinito) di $\mathbb{R}$ o $\mathbb{R}^d$

Diamo qualche esempio di modelli statistici parametrici. Ricordiamo che la funzione indicatrice di un insieme A è definita come

$$\mathbb{1}_{\{A\}}(x) = \begin{cases} 1 & if x \in A; \\ 0 & if x \notin A \end{cases}$$

- **Modello di Bernoulli**: $f_X(x;\theta) = \theta^x \times (1-\theta)^{1-x} \quad x \in \{0,1\}, \quad \Theta = [0,1]$
- **Modello esponenziale**: $f_X(x;\theta) = \theta e^{-\theta x}\ \mathbb{1}_{[0,\infty)}(x) \quad x \in \mathbb{R} \quad \Theta = [0,\infty)$
- **Modello normale**: $f_X(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}}\ e^{\frac{(x-\theta_1)^2}{2\theta_2}} \quad x \in \mathbb{R} \quad \Theta = (-\infty,\infty) \times [0,\infty)$

La modellizzazione corretta del modello statistico è il primo fondamentale passo per affrontare il problema del computo della legge "vera" della caratteristica di interesse. Il passo successivo è l'estrazione del campione casuale. Una volta noto il modello ed estratto il campione, il problema della stima dei parametri può essere affrontato. Generalmente la stima di θ è una opportuna funzione delle osservazioni che non dipende dal parametro da stimare. Precisamente

Definizione 1.9. *Una **statistica** T è una variabile aleatoria della forma $T = f(X_1, ..., X_n)$ con $f : \mathbb{R}^n \to \Theta$.*

Le statistiche rappresentano una opportuna sintesi delle osservazioni (in genere la dimensione di Θ è molto più piccola della dimensione del campione). Due statistiche che useremo molto, sono la *media campionaria* e la *varianza campionaria*. Definiamole ed analizziamo le loro proprietà. Sia $X_1, ..., X_n$ un campione in esame estratto da una legge F di media μ e varianza σ^2. Definiamo

- Media campionaria: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$
- Varianza campionaria: $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

La media e la varianza campionaria hanno la peculiarità di avere lo stesso valore atteso della media teorica e della varianza teorica della distribuzione da cui il campione è estratto. In formule

$$E\left[\bar{X}_n\right] = \mu \qquad E\left[S_n^2\right] = \sigma^2$$

Inoltre applicando la legge dei grandi numeri si verifica che

$$\bar{X}_n \underset{n\to+\infty}{\to} \mu \text{ in probabilità}$$

$$S_n^2 \underset{n\to+\infty}{\to} \sigma^2 \text{ in probabilità}$$

La prima convergenza è proprio la tesi della LGN. Per la seconda basta riscrivere la varianza campionaria come segue

$$S_n^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right]$$

ed osservare che $\frac{n}{n-1} \underset{n\to+\infty}{\to} 1$, $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \underset{n\to+\infty}{\to} E[X]^2$, $\bar{X}_n^2 \underset{n\to+\infty}{\to} \mu^2$ così che

$$S_n^2 \underset{n\to+\infty}{\to} E[X]^2 - \mu^2 = \sigma^2$$

1.4.1 Campionamento da una distribuzione normale: proprietà di $\bar{X}_n$ e di S_n^2

Analizziamo il caso in cui $F = N(\mu, \sigma^2)$. Lo studio di questo caso particolare è molto importante, perché molte volte sarà possibile fare inferenza sui parametri incogniti di una distribuzione qualunque utilizzando l' approssimazione normale stabilita dal TLC , come ad esempio nel seguente esercizio.

Esempio 1.10. VEDI [8] ESEMPIO 6.6.1 PAG 224
È noto che un certo candidato alle elezioni nazionali gode del 45% *dei consensi. Si seleziona un campione di 200 persone da intervistare. Si trovino*

1. *valore atteso e deviazione standard del numero di intervistati che preferisce il candidato in questione;*
2. *la stima della probabilità che essi siano più della metà degli intervistati.*

Se il campione in esame è estratto dalla legge $F = N(\mu, \sigma^2)$, allora

$$\boxed{\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)} \quad \text{o, equivalentemente} \quad \boxed{\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}.$$

Osserviamo che il valore medio di $\bar{X}_n$ è la media μ della legge F, mentre la sua varianza si riduce al crescere della dimensione n del campione. Quindi, se in un problema di stima il parametro incognito è μ, sarà naturale assumere come suo valore approssimato la media campionaria.

Esempio 1.11. VEDI [8] ESEMPIO 6.3.4 PAG 216
Una popolazione formata da operai maschi, presenta pesi corporei in libbre di media 167 e deviazione standard 27.

1. *Se si seleziona un campione di 36 elementi, quanto vale la probabilità che la media campionaria dei loro pesi stia tra 163 e 171?*
2. *E se si selezionano 144 operai?*

Consideriamo ora la Varianza Campionaria. Allora

$$\boxed{(n-1)\frac{S_n^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}} \tag{1.1}$$

La dimostrazione di questo risultato passa attraverso le seguenti tappe

- $\bar{X}_n$ e S_n^2 sono variabili aleatorie indipendenti (senza dim).
- Se una variabile aleatoria con distribuzione $\mathcal{X}^2_n$ è somma di due variabili aleatorie indipendenti, di cui una con distribuzione $\mathcal{X}^2_1$, allora l' altro addendo è una variabile aleatoria con distribuzione $\mathcal{X}^2_{n-1}$ (senza dim comunque sul [7] pag 322).
- Vale la seguente decomposizione

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = n\left(\frac{\bar{X}_n-\mu}{\sigma}\right)^2 + (n-1)\frac{S_n^2}{\sigma^2}$$

  Infatti

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}_n+\bar{X}_n-\mu}{\sigma}\right)^2 = (n-1)\frac{S_n^2}{\sigma^2} + n\left(\frac{\bar{X}_n-\mu}{\sigma}\right)^2$$

- $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \mathcal{X}^2_n$ e $n\left(\frac{\bar{X}_n-\mu}{\sigma}\right)^2 \sim \mathcal{X}^2_1$

Esempio 1.12. VEDI [8] ESEMPIO 6.5.1 PAG 221
Il tempo impiegato da un microprocessore ad eseguire alcuni processi è una variabile aleatoria $N(30, 9)$. Se si osserva l'esecuzione di un campione di 15 processi, qual è la probabilità che la varianza campionaria risultante sia maggiore di 12?

Da questi risultati se ne deducono altri molto utilizzati.

Corollario 1.13. $\dfrac{\sqrt{n}(\bar{X}_n-\mu)}{S_n} \sim t(n-1)$

Dimostrazione. Poichè $\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \sim N(0,1)$ e $(n-1)\frac{S_n^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}$ si ottiene facilmente che

$$\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n} = \frac{\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma}}{\sqrt{\frac{(n-1)\frac{S_n^2}{\sigma^2}}{n-1}}} \sim t(n-1)$$

□

Corollario 1.14. *Sia $X_1, ..., X_n$ un campione estratto da una $N(\mu_1, \sigma_1^2)$ e $Y_1, ..., Y_m$ un campione estratto da una $N(\mu_2, \sigma_2^2)$ e siano tali campioni indipendenti tra loro. Allora*

$$\frac{\frac{S_{1,n}^2}{\sigma_1^2}}{\frac{S_{2,m}^2}{\sigma_2^2}} \sim F(n-1, m-1)$$

(La dimostrazione è immediata dalla definizione di F)

Esempio 1.15. [8], ESERCIZIO 20 PAG 228
Consideriamo due campioni indipendenti. Il primo ha ampiezza 10 e proviene da una popolazione normale di varianza 4, il secondo ha ampiezza 5 e proviene da una popolazione normale di varianza 2. Calcolare la probabilità che la varianza campionaria del secondo campione sia maggiore di quella del primo.

Corollario 1.16. *Sotto le stesse ipotesi del precedente corollario, allora*

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{(n-1)\frac{S_{1,n}^2}{\sigma_1^2} + (m-1)\frac{S_{2,n}^2}{\sigma_2^2}}} \sqrt{\frac{n+m-2}{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim t(n+m-2)$$

Dimostrazione. Dall' indipendenza e gaussianità dei campioni si ha

$$\bar{X}_n - \bar{Y}_m \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

e

$$(n-1)\frac{S_{1,n}^2}{\sigma_1^2} + (m-1)\frac{S_{2,n}^2}{\sigma_2^2} \sim \mathcal{X}_{(n+m-2)}$$

La definizione della t di Student permette di concludere. □

Capitolo 2

Stima puntuale e per intervalli

2.1 Stima puntuale

Definizione 2.1. *Sia* $(X_1, ..., X_n)$ *campione aleatorio estratto dalla distribuzione* $\{f_X(x;\theta),\ \theta \in \Theta\}$. *Una statistica* $d = d(X_1, ..., X_n)$ *utilizzata per stimare il parametro incognito* θ *(o una sua funzione* $h(\theta)$*) è detta* ***stimatore*** *di* θ *(di* $h(\theta)$*) .*

È evidente che tra tutti gli stimatori possibili di θ (di $h(\theta)$) ce ne saranno alcuni migliori di altri rispetto ad un qualche criterio di ottimalità prescelto. Il criterio che noi scegliamo è che sia piccolo *l' errore quadratico medio* che si commette quando si utilizza lo stimatore d come valore approssimato di θ (di $h(\theta)$), ovvero che sia piccola la quantità

$$E\left[(d(X_1, ..., X_n) - \theta)^2\right] \quad \left(E\left[(d(X_1, ..., X_n) - h(\theta))^2\right]\right)$$

2.1.1 Stimatori e proprietà di ottimalità

Determinare stimatori che minimizzino l' errore quadratico medio è quasi impossibile (vedi, ad esempio [4]), a meno che non si restringa la classe degli stimatori.
Considereremo quindi una classe di stimatori che ha come pecularietà il fatto di avere come valore medio il parametro da stimare:

Definizione 2.2. *Uno stimatore* $d = d(X_1, ..., X_n)$ *di* θ *(o di* $h(\theta)$*) è detto* ***stimatore non distorto*** *o* ***stimatore corretto*** *di* θ *(di* $h(\theta)$*) se*

$$E[d(X_1, ..., X_n)] = \theta \quad [E[d(X_1, ..., X_n)] = h(\theta)]$$

Se uno stimatore è non distorto allora l' errore quadratico medio coincide con la sua varianza. In tal caso, se ci si restringe alla classe degli stimatori non distorti, lo stimatore ottimale è, se esiste, quello di varianza minima:

Definizione 2.3. *Uno stimatore non distorto* $d^* = d^*(X_1, ..., X_n)$ *di* θ *(o di* $h(\theta)$*) di varianza minima uniformemente rispetto al parametro* θ *nella la classe degli stimatori non distorti è detto* ***ottimale*** *(UMVUE).*

Vediamo nel seguente esempio che non sempre esistono stimatori ottimali:

Esempio 2.4. [1] ESEMPIO 2.2 PAG 30
Sia $X \sim Exp(\theta)$, $\theta > 0$. Allora non esiste uno stimatore non distorto di θ basato su un campione di ampiezza 1. Infatti, qualora esista, avrebbe la forma $h(X_1)$, con h non negativa e tale che

$$E_\theta(h(X_1)) = \theta \ \ \forall\ \theta > 0, \ \textit{ovvero}$$

$$\theta = \int_0^\infty h(x)\theta e^{-\theta x}dx \ \ \forall \ \ \theta > 0 \Rightarrow 1 = \int_0^\infty h(x)e^{-\theta x}dx \ \forall \ \ \theta > 0.$$

Quindi, se consideriamo due valori $\theta_1 > \theta_2 > 0$ otteniamo le uguaglianze

$$0 = \int_0^\infty h(x)e^{-\theta_2 x}dx - \int_0^\infty h(x)e^{-\theta_1 x}dx = \int_0^\infty h(x)(e^{-\theta_2 x} - e^{-\theta_1 x})dx.$$

ma $(e^{-\theta_2 x} - e^{-\theta_1 x}) > 0$, quindi, affinché l'uguaglianza sia vera, deve essere $h(x) = 0$ per ogni $x > 0$ ovvero $E_\theta(h(X_1)) = 0$ contro l'ipotesi.

Quando gli stimatori ottimali non esistono, diventa importante poter misurare il grado di bontà *(**efficienza relativa**)* di uno stimatore corretto.
Questa misura viene fatta andando a confrontare la varianza dello stimatore in esame con un confine inferiore per la varianza che può essere calcolato grazie al seguente risultato

Teorema 2.5. *(Cramer-Rao) Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x;\theta), \ \ \theta \in \Theta\}$. Sotto alcune proprietà di regolarità (vedi [7] pag 361 o [5] pag 321) che riguardano proprietà di differenziabilità della densità $f_X(x;\theta)$ ed il fatto che il suo supporto $\{x \in \mathbb{R} \ \ tale\ che \ \ f_X(x;\theta) > 0\}$ non dipenda da θ, allora la varianza di un qualunque stimatore non distorto $d(X_1, ..., X_n)$ di una funzione derivabile del parametro $h(\theta)$ soddisfa la disuguaglianza*

$$Var\left(d(X_1, ..., X_n)\right) \geq \frac{[h'(\theta)]^2}{nE\left(\dfrac{\partial \log f_X(x;\theta)}{\partial\theta}\right)^2} \tag{2.1}$$

La quantità a secondo membro è nota come ***limite inferiore di Cramer Rao***, mentre la quantità $I_X(\theta)$ definita come

$$I_X(\theta) = E\left(\frac{\partial \log f_X(x;\theta)}{\partial\theta}\right)^2 \tag{2.2}$$

è nota come ***informazione di Fischer***.

Poniamo $\boxed{B_n(\theta) = \frac{[h'(\theta)]^2}{nI_X(\theta)}}$. A questo punto se uno stimatore non distorto ha varianza pari al limite inferiore di Cramer Rao, è sicuramente ottimale, altrimenti è possibile utilizzare come misura di efficienza $e(d)$ di uno stimatore non distorto d la quantità

$$e(d) = \frac{B_n(\theta)}{Var(d)} \leq 1 \tag{2.3}$$

Definizione 2.6. *Uno stimatore corretto $d = d(X_1, ..., X_n)$ è detto efficiente se la sua varianza coincide con il limite inferiore di Cramer Rao, o, equivalentemente, se $e(d) = 1$*

Osservazione 2.7. *Il limite inferiore di Cramer Rao* ***non è*** *l' estremo inferiore della varianza degli stimatori non distorti, quindi è possibile che uno stimatore sia il migliore possibile, nel senso che abbiamo introdotto, anche quando la sua varianza è maggiore di tale limite.*

Facciamo degli esempi di stimatori efficienti:

1 Stimatore efficiente per la media di una bernoulliana ([1] PAG 35)

2 Stimatore efficiente per il parametro θ di una legge $\Gamma(\alpha, \frac{1}{\theta})$, con α noto ([1] PAG 38)

Accanto alle proprietà di ottimalità per campioni finiti, ci sono anche delle proprietà asintotiche che elenchiamo di seguito.

Definizione 2.8. *Uno stimatore* $d = d(X_1, ..., X_n)$ *di* θ *è detto* ***asintoticamente corretto*** *se gode della proprietà*

$$\lim_{n\to\infty} E[d(X_1, ..., X_n)] = \theta$$

Definizione 2.9. *Uno stimatore* $d = d(X_1, ..., X_n)$ *di* θ *è detto* ***consistente***

$$\lim_{n\to\infty} d(X_1, ..., X_n) = \theta$$

dove il limite precedente è da intendersi in probabilità.

Due stimatori consistenti della media e della varianza sono la media campionaria e la varianza campionaria.
Ovviamente la consistenza di uno stimatore implica la sua correttezza asintotica.

Definizione 2.10. *Uno stimatore* $d = d(X_1, ..., X_n)$ *di* θ *è detto* ***asintoticamente efficiente*** *se*

$$\lim_{n\to\infty} e[d(X_1, ..., X_n)] = 1$$

ovvero se per dimensioni grandi del campione $Var[d(X_1, ..., X_n)] \approx B_n(\theta)$.

Un ultima proprietà molto importante per grandi campioni è

Definizione 2.11. *Uno stimatore* $d = d(X_1, ..., X_n)$ *di* θ *è detto* ***asintoticamente normale*** *se*

$$\lim_{n\to\infty} P\left(\frac{d(X_1, ..., X_n) - E[d(X_1, ..., X_n)]}{\sqrt{Var[d(X_1, ..., X_n)]}} \leq t\right) = \Phi(t)$$

2.1.2 Il metodo della massima verosimiglianza

Cominciamo con i metodi di costruzione di stimatori. In questo paragrafo analizziamo un metodo che produce stimatori asintoticamente efficienti e consistenti, motivo per cui sono tra i più utilizzati in statistica.
Cominciamo con qualche definizione:

Definizione 2.12. *Sia $(X_1 = x_1, ..., X_n = x_n)$ una realizzazione di un campione aleatorio $(X_1, ..., X_n)$ estratto dalla distribuzione $\{f_X(x;\theta), \quad \theta \in \Theta\}$. Sia $L : \Theta \to \mathbb{R}$ la densità congiunta del campione $(X_1, ..., X_n)$ calcolata nel punto $(x_1, ..., x_n)$, ovvero*

$$L(\theta) = \Pi_{i=1}^{n} f_X(x_i;\theta)$$

Tale densità vista come funzione del parametro incognito θ è nota con il nome di ***funzione di verosimiglianza****.*

Definizione 2.13. *Sia $\widehat{\theta}(x_1, ..., x_n)$ una statistica che massimizza la funzione di verosimiglianza $L(\theta)$ (se esiste!!!). Si chiama* ***stimatore di massima verosimiglianza*** *di θ la statistica $\widehat{\theta}(X_1, ..., X_n)$.*

Il principio euristico che motiva la scelta di tali stimatori è il seguente ([6], pag 587):
" *tra i possibili valori del parametro θ si preferisce quello che corrisponde alla massima probabilità di generare i dati osservati* "

Operativamente la costruzione di tali stimatori corrisponde alla ricerca del massimo di una funzione. Se il parametro è multidimensionale si utilizzano metodi di calcolo relativi alle funzioni di più variabili.

Osservare che il punto che rende massima una funzione non negativa (come una densità) è lo stesso che rende massimo il logaritmo di tale funzione (il logaritmo è una funzione monotona) semplifica notevolmente i calcoli. Introduciamo quindi la ***funzione di log-verosimiglianza***, cioè la funzione $\log(L(\theta))$.

Illustreremo il metodo di calcolo per la ricerca di tali stimatori nei seguenti casi:

1. $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{N(\theta, 1), \quad \theta \in \mathbb{R}\}}$ ([6], pag 588)
2. $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{N(\theta, \sigma^2), \quad (\theta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}}$ ([6], pag 589)
3. $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{N(0, \theta), \quad \theta \in \mathbb{R}^+\}}$
4. $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{U(0, \theta), \quad \theta \in \mathbb{R}^+\}}$(Da confrontare con quello trovato con il metodo dei momenti)([6], pag 590)
5. $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{Exp(\theta), \quad \theta \in \mathbb{R}^+\}}$
6. $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{P(\theta), \quad \theta \in \mathbb{R}^+\}}$([8] pag. 238)

Esempio 2.14. *[8] pag. 239 7.2.4*
Nel 1998 a Berkeley in California, il numero di incidenti stradali in 10 giornate senza pioggia scelte a caso è stato di

4 0 6 5 2 1 2 0 4 3

Si usino questi dati per stimare per quell'anno la frazione di giornate senza pioggia con non più di 2 incidenti.

7 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{B(\theta), \quad \theta \in (0,1)\}}$

8 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{U[\theta - \frac{1}{2}, \theta + \frac{1}{2}], \quad \theta \in \mathbb{R}\}}$ ([1], pag 54 - sulla non unicità dello stimatore di max verosimiglianza)

Tanti esercizi interessanti ed una sintesi delle proprietà teoriche si trovano su [1].

2.1.3 Proprietà degli stimatori di massima verosimiglianza

Gli stimatori di massima verosimiglianza hanno interessanti proprietà di ottimalità sia per campionamento finito che asintotiche.
Supponiamo che il modello statistico di riferimento verifichi le condizioni di regolarità accennate nell' enunciato del Teorema 2.5. Valgono allora le seguenti proprietà:

Proprietà 2.1. *Se esiste uno stimatore non distorto ed efficiente, tale stimatore è quello di massima verosimiglianza.*

Proprietà 2.2. *Gli stimatori di massima verosimiglianza sono asintoticamente efficienti e consistenti.*

Proprietà 2.3. *Gli stimatori di massima verosimiglianza sono asintoticamente normali, ovvero:*

$$\sqrt{nI_X(\theta)}(\widehat{\theta}(X_1, ..., X_n) - \theta) \Rightarrow N(0,1),$$

dove $I_X(\theta)$ è l'informazione di Fischer definita nell'Eq. 2.2.

La proprietà precedente vale anche quando l' informazione di Fischer $I_X(\theta)$ è approssimata per mezzo dello stimatore di massima verosimiglianza $I_X(\widehat{\theta})$, ovvero:

Proprietà 2.4.

$$\sqrt{nI_X(\widehat{\theta})}(\widehat{\theta}(X_1, ..., X_n) - \theta) \Rightarrow N(0,1).$$

SERVE UN RIFERIMENTO BIBLIOGRAFICO PER LE DIM DELLE ULTIME 2 PROPRIETÀ ELENCATE.

Queste proprietà possono essere riassunte come segue (cfr. [6] pag. 601)
" Se esiste uno stimatore efficiente per θ, lo stimatore di max. ver. coincide con esso e quindi è efficiente per per ogni n finito. D'altra parte, anche se non esiste uno stimatore efficiente per θ, lo stimatore di max. ver. è comunque asintoticamente efficiente"

Una ulteriore proprietà molto utile quando si voglia stimare una funzione del parametro incognito è la seguente:

Proprietà 2.5. ***Proprietà di invarianza*** - *Sia $g : \Theta \to \mathbb{R}$ una funzione invertibile. Allora se $\widehat{\theta}(X_1, ..., X_n)$ è uno stimatore di massima verosimiglianza per θ, $g(\widehat{\theta})$ è uno stimatore di massima verosimiglianza per $g(\theta)$.*

Come applicazione della proprietà precedente facciamo il seguente esempio:

Esempio 2.15. [6] 16.21 PAG 598
Da un campione casuale $(X_1, ..., X_n)$ generato da misurazioni sulle durate del funzionamento di componenti elettroniche, che si suppongono avere una legge $Exp(\theta)$, $\theta > 0$, si vuole stimare la probabilità che esse sopravvivano almeno 3 ore in più della durata media, ovvero si vuole stimare la quantità

$$P\left(X > 3 + \frac{1}{\theta}\right) = 1 - F_X\left(3 + \frac{1}{\theta}\right) = 1 - \left(1 - e^{-\theta(3+\frac{1}{\theta})}\right) = e^{-(3\theta+1)}$$

Si dimostra che lo stimatore di max verosimiglianza di θ è $\frac{1}{\overline{X}_n}$ (CONTROLLARE!!!!) e quindi, per la proprietà di invarianza, lo stimatore di max verosimiglianza di $P\left(X > 3 + \frac{1}{\theta}\right)$ è $e^{-\left(\frac{3}{\overline{X}_n}+1\right)}$.

2.1.4 Il metodo dei momenti

Ricordiamo la seguente definizione

Definizione 2.16. *Data una sequenza di variabili aleatorie $(X_1, ..., X_n)$ indipendenti ed identicamente distribuite, si chiama momento campionario di ordine $r \in N$ la quantità M_r così definita*

$$M_r = \frac{\sum_{i=1}^{n} X_i^r}{n} \tag{2.4}$$

L'applicabilità del metodo dei momenti è basato essenzialmente su due condizioni:

1. Il numero r dei parametri da stimare sia non maggiore del numero dei momenti teorici che possiede la distribuzione in esame;
2. i parametri da stimare siano delle funzioni **note** di tali momenti.

In tal caso, degli stimatori naturali dei momenti teorici sono i relativi momenti campionari (grazie alla legge dei grandi numeri). Si imposta quindi un sistema di r equazioni in r incognite (i parametri da stimare) e si risolve.
Le soluzioni ottenute in questo modo (se esistono- APPROFONDIRE) sono note come ***stimatori dei momenti*** dei parametri.
Il metodo dei momenti per il calcolo degli stimatori è uno dei più semplici da implementare e richiede ipotesi meno stringenti rispetto al metodo della max verosimiglianza. Ad esempio, non richiede la conoscenza della forma funzionale della distribuzione in esame. Ma, proprio perché ha meno vincoli, le stime che fornisce sono in generale "meno buone". Per le proprietà di tale metodo parafrasiamo dal [6], pag. 584:

*" Poiché sono funzioni continue dei momenti campionari, gli stimatori derivati con il metodo dei momenti sono **consistenti**, **asintoticamente non distorti** ed **asintoticamente normali**. D' altra parte, non sempre sono efficienti, neppure asintoticamente...." "...Ne deriva che le proprietà di tali stimatori sono di natura asintotica....inoltre tali stimatori non garantiscono sempre stime coerenti."*

Calcoliamo tali stimatori nei casi:

1 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{f(x, \theta) = (\theta + 1)x^\theta \mathbb{1}_{[0,1]}(x), \quad \theta \in \mathbb{R}^+\}}$ [6], pag. 582 esempio 16.3) DA OSSERVARE CHE PER ALCUNE REALIZZAZIONI DEL CAMPIONE FORNISCE STIME NON COERENTI(CORRISPONDENTI A VALORI DI θ NEGATIVI

2 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{B(N; \theta), \quad (N, \theta) \in \mathbb{N} \times (0, 1)\}}$ ([6], pag. 583) ANCHE QUI POTREBBERO ESSERCI STIME NON COERENTI

3 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{N(\theta, \sigma^2), \quad (\theta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}}$ ([6], pag. 583)

4 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{U(0, \theta), \quad \theta \in \mathbb{R}^+\}}$(Da confrontare con quello trovato con il metodo della max verosimiglianza)([6], pag. 584) PRODUCE STIMATORI NON EFFICIENTI CON EFFICIENZA ASINTOTICA TENDENTE A ZERO. COME ESEMPIO, SE $n = 3$ E SI OSSERVA $(1, 24, 2)$, ALLORA LA STIMA È PARI A 18, MA $24 > 18$!

2.2 Intervalli di confidenza

Nel capitolo precedente abbiamo fornito metodi per la costruzione di stimatori di parametri incogniti di una distribuzione la cui forma è nota. Però non sappiamo quantificare, neppure nel caso in cui gli stimatori godano di proprietà di ottimalità, " quanto" buone siano le stime ottenute. La stima intervallare, a differenza di quella puntuale, si preoccupa di fornire non un singolo valore numerico per i parametri incogniti ma un intervallo che, con un " grado di fiducia" (APPROFONDIRE IL CONCETTO) fissato a priori, contenga il parametro incognito. Studieremo intervalli di confidenza per parametri scalari, ma segnaliamo che anche il caso vettoriale può essere affrontato.

Definizione 2.17. *Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x; \theta), \ \theta \in \Theta\}$. Sia $1 - \alpha \in (0, 1)$ e siano $T_1 = t_1(X_1, ..., X_n)$ e $T_2 = t_2(X_1, ..., X_n)$ due statistiche tali che:*

- $T_1 \leq T_2$;
- $P(T_1 < \theta < T_2) = 1 - \alpha$

*allora l'intervallo casuale (T_1, T_2) si chiama **intervallo di confidenza al livello** $1 - \alpha$ per il parametro incognito θ, mentre $1 - \alpha$ è il **livello di confidenza**.*

Gli intervalli di confidenza possono essere anche unilaterali:

Definizione 2.18. *Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x; \theta), \ \theta \in \Theta\}$. Sia $1 - \alpha \in (0, 1)$ e siano $T_1 = t_1(X_1, ..., X_n)$ e $T_2 = t_2(X_1, ..., X_n)$ due statistiche tali che:*

- $P(\theta > T_1) = 1 - \alpha$
- $P(\theta < T_2) = 1 - \alpha$

*allora la statistica T_1 è l'**estremo di confidenza inferiore al livello** $1-\alpha$ per θ, mentre la statistica T_2 è l'**estremo di confidenza superiore al livello** $1-\alpha$ per θ*[1]

Osservazione 2.19. *Le precedente definizioni si estendono in modo naturale ad intervalli di confidenza per funzioni $h(\theta)$ del parametro θ.*

È evidente che, a parità di livello di confidenza, possono esistere infiniti intervalli, mentre, nel caso di campionamenti da distribuzioni continue, gli estremi di confidenza sono univocamente determinati. Vedremo nel prossimo paragrafo i criteri più utilizzati per costruire intervalli di confidenza ottimali.

2.2.1 Costruzione di intervalli di confidenza: il metodo della quantità pivotale

Definizione 2.20. *Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x;\theta),\ \theta \in \Theta\}$. Sia Q una funzione del campione $(X_1, ..., X_n)$ e del parametro incognito θ, cioè della forma $Q = q(X_1, ..., X_n; \theta)$ la cui distribuzione sia nota. Allora Q è detta **quantità pivotale**.*

Esempio 2.21. *Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{N(\theta, 4),\ \theta \in \mathbb{R}\}$. Allora $\overline{X}_n - \theta \sim N(0, \frac{4}{n})$ e $\frac{\overline{X}_n - \theta}{\frac{2}{\sqrt{n}}} \sim N(0,1)$ sono quantità pivotali, perchè dipendono funzionalmente dal campione e dal parametro ma la loro legge è nota.*

Come si determinano le quantità pivotali? E come si utilizzano nella ricerca di intervalli di confidenza? Esponiamo la procedura nei seguenti passaggi:

1. Si considera un campione casuale $(X_1, ..., X_n)$ oppure una sua statistica $T(X_1, ..., X_n)$ (in genere uno stimatore del parametro);
2. si cerca una trasformazione $Q = q(T(X_1, ..., X_n); \theta)$ la cui legge sia nota;
3. per ogni fissato livello di confidenza $1-\alpha \in (0,1)$ si determina una coppia di punti z_1, z_2 tali per cui $P(z_1 \leq Q \leq z_2) = 1-\alpha$;
4. si esprime l' evento $\{z_1 \leq Q \leq z_2\} = \{T_1(X_1, ..., X_n) \leq \theta \leq T_2(X_1, ..., X_n)\}$ laddove possibile [2].

In questo modo si è ottenuto un intervallo di confidenza $[T_1(X_1, ..., X_n), T_2(X_1, ..., X_n)]$ al livello $1-\alpha$ per il parametro incognito θ.

La procedura appena descritta lascia aperto il problema della scelta della coppia z_1, z_2, che, come già osservato, non è unica. In generale si sceglie la coppia che produce intervalli di confidenza con una delle due seguenti proprietà:

1. *Intervalli con code equiprobabili*: si sceglie la coppia z_1, z_2 tale per cui $P(Q \leq z_1) = P(Q \geq z_2) = \frac{\alpha}{2}$; ovvero, se $1-\alpha$ è l'ampiezza dell'intervallo, $z_1 = q_{\frac{\alpha}{2}}$ e $z_2 = q_{1-\frac{\alpha}{2}}$.

[1] [6], pag. 732 per esempi nel caso unilaterale

[2] Quest'ultimo passaggio è una procedura di inversione rispetto al parametro incognito θ ed è molto agevole se la variabile aleatoria pivot Q è invertibile o meglio ancora monotona come funzione del parametro.

2 *Intervalli di minima lunghezza (eventualmente media)*: si sceglie la coppia z_1, z_2 che, a parità di livello, renda minima la lunghezza $T_2(X_1, ..., X_n) - T_1(X_1, ..., X_n)$.

Un controllo sulla forma della densità dalla quale si campiona aiuta a determinare le quantità pivotali.

Definizione 2.22. *Sia dato il modello statistico* $\{f(x,\theta), \quad \theta \in \Theta \subseteq \mathbb{R}\}$. *Il parametro* θ *è un* ***parametro di posizione*** *se e solo se la densità* $f(x,\theta)$ *può essere scritta come funzione di* $(x-\theta)$, *ovvero se per una opportuna funzione* h, *si ha* $f(x,\theta) = h(x-\theta)$.
Il parametro $\theta \in \mathbb{R}^+$ *è invece un* ***parametro di scala*** *se e solo se per una opportuna funzione* h, *la densità* $f(x,\theta)$ *può essere scritta come* $f(x,\theta) = \frac{1}{\theta} h\left(\frac{x}{\theta}\right)$.

Ora se θ è un parametro di posizione, allora $\overline{X}_n - \theta$ è una quantità pivotale, mentre se θ è un parametro di scala $\frac{\overline{X}_n}{\theta}$ è una quantità pivotale.[3]

Applichiamo questo metodo per la ricerca di intervalli nei seguenti casi:

1 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{U(0,\theta), \quad \theta \in \mathbb{R}^+\}}$([6] pag 741 19.5, oppure [7] (meglio) Es.4 pag 484)
Sappiamo che $T(X_1, ..., X_n) = \max(X_1, ..., X_n) := X_{(n)}$ è lo stimatore di max verosimiglianza di θ. La densità di tale stimatore è

$$f_{X_{(n)}}(t) = \begin{cases} \frac{nt^{n-1}}{\theta^n} & \text{se } t \in (0,\theta) \\ 0 & \text{altrimenti} \end{cases}$$

Inoltre la variabile aleatoria $q(T(X_1, ..., X_n); \theta) = \frac{X_{(n)}}{\theta}$ ha densità pari a

$$f_q(t) = \begin{cases} nt^{n-1} & \text{se } t \in (0,1) \\ 0 & \text{altrimenti} \end{cases}$$

Possiamo quindi utilizzare $\frac{X_{(n)}}{\theta}$ come quantità pivotale. A questo punto, per ottenere un intervallo di confidenza a livello $1-\alpha$ basta determinare $\{z_1, z_2\}$ tali per cui

$$P\left(z_1 \leq \frac{X_{(n)}}{\theta} \leq z_2\right) = 1 - \alpha$$

Poiché $\left\{z_1 \leq \frac{X_{(n)}}{\theta} \leq z_2\right\} = \left\{\frac{X_{(n)}}{z_2} \leq \theta \leq \frac{X_{(n)}}{z_1}\right\}$ l'intervallo di confidenza cercato ha l'espressione

$$\left[\frac{X_{(n)}}{z_2}, \frac{X_{(n)}}{z_1}\right]$$

Per avere un intervallo ottimale, bisogna determinare la coppia $\{z_1, z_2\}$ in modo tale che sia minima la lunghezza dell' intervallo, ovvero dobbiamo minimizzare la funzione $L(z_1, z_2) = X_{(n)}\left(\frac{1}{z_1} - \frac{1}{z_2}\right)$ soggetta al vincolo

$$P\left(z_1 \leq \frac{X_{(n)}}{\theta} \leq z_2\right) = \int_{z_1}^{z_2} nt^{n-1} dt = {z_2}^n - {z_1}^n = 1 - \alpha$$

[3] In realtà quando il modello è di uno dei due tipi descritti, possono essere costruite molte altre quantità pivotali.

Da questa uguaglianza si deduce che ${z_1}^n = {z_2}^n - 1 + \alpha$ e che $(1-\alpha)^{\frac{1}{n}} < z_2 \leq 1$. In particolare si ottiene una espressione della lunghezza dell' intervallo come funzione della sola z_2. Applicando le regole standard per il calcolo dei minimi ed osservando che $\frac{dz_1}{dz_2} = \frac{z_2^{n-1}}{z_1^{n-1}}$ si ottiene:

$$\frac{dL}{dz_2} = X_{(n)}\left(-\frac{1}{z_1^2}\frac{z_2^{n-1}}{z_1^{n-1}} + \frac{1}{z_2^2}\right) = X_{(n)}\frac{z_1^{n+1} - z_2^{n+1}}{z_1^{n+1}z_2^2} < 0.$$

Quindi L è una funzione decrescente e dunque assume il suo minimo assoluto per $z_2 = 1$. Di conseguenza $z_1 = \alpha^{\frac{1}{n}}$ e quindi l'intervallo di confidenza ottimale è

$$\left[X_{(n)}, \frac{X_{(n)}}{\alpha^{\frac{1}{n}}}\right]$$

2 Determinare l'intervallo di confidenza a code equiprobabili per un campione di dimensione 1 estratto dalla distribuzione $\boxed{\{f_\theta(x) = \frac{2}{\theta^2}(\theta - x), \quad 0 < x < \theta\}}$
[7] PAG 479

3 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{Exp(\theta), \quad \theta \in \mathbb{R}^+\}}$ [8], ESEMPIO 7.6.1. PAG 269

Suggerimento: utilizza il fatto che, quando $Y \sim \Gamma(n, \theta)$, allora $2\theta Y \sim \chi^2(2n)$)

4 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{f_\theta(x) = \begin{cases} e^{-(x-\theta)} & \text{se } x > \theta \\ 0 & \text{altrimenti} \end{cases}}$

2.2.2 Costruzione di intervalli di confidenza: il metodo della trasformazione integrale

La scelta di una quantità pivotale per la costruzione di intervalli di confidenza dipende principalmente dal modello parametrico in esame. Le difficoltà maggiori si incontrano con i modelli discreti, poiché in tal caso il calcolo dei quantili può risultare difficoltoso.
Per i modelli continui invece esiste un metodo sempre perseguibile, basato sulla *trasformazione integrale di probabilità*:

Teorema 2.23. *Sia X una variabile aleatoria continua con funzione di ripartizione $F_X(x)$. Allora la variabile aleatoria $U = F_X(X)$ è distribuita uniformemente nell'intervallo* $(0, 1)$.

La dimostrazione del precedente risultato è lasciata per esercizio.

Sia ora $(X_1, ..., X_n)$ un campione estratto dal modello parametrico *continuo* $\{f(x, \theta), \quad \theta \in \Theta\}$. Detta $F_X(x, \theta)$ la funzione di ripartizione comune degli elementi del campione, il Teorema 2.23 garantisce che $F_X(X_i, \theta) \sim U(0, 1)$ per $i = 1, ..., n$.

Inoltre $Y_i \equiv -\ln F_X(X_i,\theta) \sim \exp(1)$ per $i = 1, ..., n$. Infatti

$$F_{Y_i}(t) = \begin{cases} 0 & \text{se } t < 0; \\ P(-\ln F_X(X_i,\theta) \le t) = P(F_X(X_i,\theta) \ge e^{-t}) = 1 - e^{-t} & \text{se } t \ge 0. \end{cases}$$

È dunque possibile considerare come quantità pivotale $Q_1(X_1, ..., X_n; \theta) = \sum_{i=1}^n -\ln F_X(X_i,\theta) \sim \Gamma(n, 1)$, oppure, alternativamente, $Q_2(X_1, ..., X_n; \theta) = 2\sum_{i=1}^n -\ln F_X(X_i,\theta) \sim \Gamma(n, \frac{1}{2}) \equiv \chi^2(2n)$.
Abbiamo appena visto un metodo standard per la determinazione di una quantità pivotale ogni qual volta il modello parametrico in esame sia continuo.
La possibilità poi di utilizzare questa quantità pivotale per la determinazione di intervalli di confidenza va valutata caso per caso ed è come noto legata alla possibilità di invertirla rispetto al parametro incognito. Questo è peraltro sempre possibile quando sia monotona in θ.

Esempio 2.24. *([5] pag 391, esempio 8.4). Sia $(X_1, ..., X_n)$ un campione estratto dal modello parametrico $\{\theta x^{\theta-1} I_{(0,1)}(x), \ \theta > 0\}$. Calcolare un intervallo di confidenza a code equiprobabili di livello $1 - \alpha$ per θ.*

Esempio 2.25. *Sia $(X_1, ..., X_n)$ un campione estratto dal modello parametrico $\{U[0, \theta], \ \theta > 0\}$. Calcolare un intervallo di confidenza a code equiprobabili di livello $1 - \alpha$ per θ basato sul metodo della trasformazione integrale e confrontarlo con quello ottenuto attraverso lo stimatore di massima verosimiglianza.*

2.2.3 Intervalli di confidenza per campioni normali

Sia $(X_1, ..., X_n)$ un campione estratto da una distribuzione $N(\mu, \sigma^2)$. Ricaveremo di seguito intervalli di confidenza per ciascuno dei parametri che caratterizza tale legge.

Intervallo di confidenza per la media Notiamo che sono possibili 2 casi

1 **σ^2 è un valore noto**. Sappiamo che lo stimatore di massima verosimiglianza per la media μ è la media campionaria $\bar{X}_n$ e che $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Di conseguenza standardizzando si ottiene $\sqrt{n}\dfrac{\bar{X}_n - \mu}{\sigma} \sim N(0, 1)$. Poiché σ^2 è noto la variabile aleatoria $\sqrt{n}\dfrac{\bar{X}_n - \mu}{\sigma}$ è una quantità pivotale. La utilizziamo quindi per determinare un intervallo di confidenza per la media a livello $1 - \alpha$. Cerchiamo l'intervallo di confidenza a code equiprobabili, cioé quello ottenuto invertendo rispetto a μ le seguenti disuguaglianze[4]

$$\left\{ \phi_{\frac{\alpha}{2}} \le \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \le \phi_{1-\frac{\alpha}{2}} \right\}$$

Una semplice procedura di inversione e la proprietà dei quantili della normale standard permettono di ricavare il seguente intervallo di confidenza per μ

$$\left\{ \bar{X}_n - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \le \mu \le \bar{X}_n + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \right\}$$

[4]Quando la legge della quantità pivotale in esame è simmetrica, l'intervallo di confidenza a code equiprobabili coincide con quello di lunghezza minima ([5] per approfondimenti).

o, equivalentemente

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

Esempio 2.26. [8], ESEMPIO 7.3.1 PAG. 246
È noto che quando un segnale elettrico di valore μ viene trasmesso da una sorgente A, il ricevente B registri effettivamente un valore X distribuito come una $N(\mu, 4)$. Per ridurre l'errore, lo stesso segnale viene inviato 9 volte e si registra la media campionaria dei segnali ricevuti, ovvero $\bar{X}_9$. Sapendo che $\bar{X}_9 = 9$, determinare un intervallo di confidenza di livello 0.95 per μ.

Esempio 2.27. [8], ESEMPIO 7.3.4 PAG. 249
Il peso dei salmoni cresciuti un un certo allevamento segue una legge normale con media μ, che varia di anno in anno, e deviazione standard $\sigma = 0.3$ libbre. Quanto grande occorre prendere il campione, per essere sicuri al 95% che la nostra stima del peso medio dei salmoni di quest'anno sia precisa entro ± 0.1 libbre?

2 σ^2 **è un valore incognito**. In tal caso la quantità $\sqrt{n}\dfrac{\bar{X}_n - \mu}{\sigma}$ non è più una quantità pivotale, perché dipende dal parametro incognito σ. Sappiamo però che la varianza campionaria S_n^2 è indipendente da $\bar{X}_n$ e che $(n-1)\frac{S_n^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}$. Ne deduciamo che $\dfrac{\frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{(n-1)\frac{S_n^2}{\sigma^2}}{n-1}}} \sim t_{n-1}$. Facendo le opportune semplificazioni si ricava $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}$ che è quindi una quantità pivotale. Anche in questo caso procedendo in modo analogo al punto precedente, determiniamo l'intervallo di confidenza a code equiprobabili

$$\left\{ t_{\frac{\alpha}{2}, n-1} \leq \sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \leq t_{1-\frac{\alpha}{2}, n-1} \right\}$$

da cui

$$\left\{ \bar{X}_n - \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \leq \mu \leq \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} \right\}$$

o, equivalentemente

$$\boxed{\left[\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1},\ \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1}\right]}$$

Intervallo di confidenza per la varianza Anche in questa situazione è possibile distinguere 2 casi

1 μ **è un valore noto**. In tal caso $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$, in quanto somma di quadrati di normali standard indipendenti, è distribuita come una chi-quadrato con n gradi di libertà ed è quindi una quantità pivotale. La utilizziamo quindi per

determinare un intervallo di confidenza a code equiprobabili[5] di livello $1-\alpha$

$$\left\{ \chi^2_{\frac{\alpha}{2},n} \leq \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \leq \chi^2_{1-\frac{\alpha}{2},n} \right\}$$

da cui, invertendo rispetto a σ^2

$$\left\{ \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2},n}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2},n}} \right\}$$

o, equivalentemente

$$\boxed{\left[\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2},n}}, \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2},n}} \right]}$$

2 μ **è un valore incognito** La quantità pivotale che si utilizza in questo caso è $\frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \sim \chi^2_{n-1}$. Si ottiene l'intervallo di confidenza a code equiprobabili di livello $1-\alpha$ attraverso gli stessi passaggi utilizzati nel caso precedente:

$$\left\{ \chi^2_{\frac{\alpha}{2},n-1} \leq \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \leq \chi^2_{1-\frac{\alpha}{2},n-1} \right\}$$

da cui, invertendo rispetto a σ^2

$$\left\{ \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}{\chi^2_{\frac{\alpha}{2},n-1}} \right\}$$

o, equivalentemente

$$\boxed{\left[\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}{\chi^2_{\frac{\alpha}{2},n-1}} \right]}$$

Esempio 2.28. [8], ESEMPIO 7.3.8 PAG. 256
Una procedura automatizzata deve produrre rondelle con una variabilità di spessore molto ridotta. Per testare questa variabilità si scelgono 10 rondelle dalla produzione e se ne misura lo spessore, che risulta, in pollici

0.123 0.133 0.124 0.125 0.126 0.128 0.120 0.124 0.130 0.126

Calcolare l'intervallo di confidenza di livello 0.9 per la deviazione standard dello spessore delle rondelle.

[5]Osserviamo che la legge chi-quadrato non è simmetrica e quindi l'intervallo di confidenza a code equiprobabili non è il migliore possibile. Un metodo per determinare l'intervallo di minima lunghezza è descritto in [5], pag 385.

Intervallo di confidenza per la differenza tra medie Consideriamo due campioni indipendenti $(X_1, ..., X_n)$ estratto da una distribuzione $N(\mu_1, \sigma_1^2)$ e $(Y_1, ..., Y_m)$ estratto da una distribuzione $N(\mu_2, \sigma_2^2)$. Dalle proprietà della normale si evince che $\bar{X}_n - \bar{Y}_m \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}\right)$,e, standardizzando

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1) \tag{2.5}$$

Per determinare un intervallo di confidenza a livello $1-\alpha$ per la differenza tra medie distinguiamo come al solito 2 casi

1 σ_1^2, σ_2^2 **sono noti**. In tal caso la quantità nell'eq. 2.5 è evidentemente una quantità pivotale e di conseguenza i soliti passaggi permettono di derivare l'intervallo

$$\boxed{\left[\bar{X}_n - \bar{Y}_m - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\frac{\alpha}{2}}, \bar{X}_n - \bar{Y}_m + \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\frac{\alpha}{2}}\right]}$$

2 σ_1^2, σ_2^2 **sono incogniti.** In questo caso la variabile aleatoria 2.5 non è più una quantità pivotale poiché dipende dai 2 parametri incogniti. D'altra parte, quando si possa assumere $\sigma_1^2 = \sigma_2^2 = \sigma_2$ si riesce a costruire una quantità pivotale osservando che $\frac{(n-1)S_{1n}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \sim \chi_{n-1}^2$ e che $\frac{(m-1)S_{2m}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{m}\left(Y_i - \bar{Y}_m\right)^2 \sim \chi_{m-1}^2$; di conseguenza, dalle proprietà della legge chi-quadro si ottiene

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

Infine, utilizzando l'indipendenza tra questa variabile aleatoria e la 2.5, si ottiene

$$\frac{\frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}}{\sqrt{\frac{\frac{(n-1)S_{1n}^2}{\sigma^2} + \frac{(m-1)S_{2m}^2}{\sigma^2}}{n+m-2}}} \sim t_{n+m-2}$$

Posto $S_p^2 = \frac{(n-1)S_{1n}^2 + (m-1)S_{2m}^2}{n+m-2}$ e semplificando la precedente espressione, si ricava finalmente la quantità pivotale

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$$

La quantità S_p^2 è uno stimatore non distorto di σ^2 ed è nota come *varianza campionaria conglobata* (perché calcolata sull'informazione ottenuta da due diversi campioni).

Siamo finalmente in grado di ricavare l'intervallo di confidenza per la differenza tra le medie

$$\left\{ t_{\frac{\alpha}{2},n+m-2} \leq \frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \leq t_{1-\frac{\alpha}{2},n+m-2} \right\}$$

da cui

$$\left\{ \bar{X}_n - \bar{Y}_m - \sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)} t_{1-\frac{\alpha}{2},n+m-2} \leq \mu_1 - \mu_2 \leq \bar{X}_n - \bar{Y}_m + \sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)} t_{1-\frac{\alpha}{2},n+m-2} \right.$$

ovvero

$$\boxed{\left[\bar{X}_n - \bar{Y}_m - \sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)} t_{1-\frac{\alpha}{2},n+m-2}, \bar{X}_n - \bar{Y}_m + \sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)} t_{1-\frac{\alpha}{2},n+m-2} \right]}$$

Esempio 2.29. [8], ESERCIZIO 44 PAG. 288
Quelli che seguono sono i tempi di combustione in secondi di due diversi tipi di candelotti fumogeni:

Tipo I	*481*	*506*	*527*	*661*	*501*	*572*	*561*	*501*	*487*	*524*
Tipo II	*526*	*511*	*556*	*542*	*491*	*537*	*582*	*605*	*558*	*578*

Assumendo che le popolazioni siano normali con stessa varianza, calcolare l'intervallo di confidenza di livello 0.99 per la differenza media dei tempi di combustione.

2.2.4 Intervalli di confidenza asintotici

Sia $(X_1, ..., X_n)$ un campione estratto dalla distribuzione $\{f(x,\theta), \quad \theta \in \Theta\}$. Supponiamo che esista uno stimatore di massima verosimiglianza $\widehat{\theta} = \widehat{\theta}(X_1, ..., X_n)$ per il parametro θ. Ricordiamo che sotto opportune ipotesi dei regolarità, le proprietà 2.3 e 2.4 assicurano la normalità asintotica degli stimatori di max verosimiglianza, ovvero

$$\sqrt{nI_X(\theta)}(\widehat{\theta} - \theta) \Rightarrow N(0,1),$$
$$\sqrt{nI_X(\widehat{\theta})}(\widehat{\theta} - \theta) \Rightarrow N(0,1),$$

dove $I_X(\theta)$ è l'informazione di Fischer.
Dunque, se la numerosità del campione lo permette, le quantità $\boxed{\sqrt{nI_X(\theta)}(\widehat{\theta} - \theta)}$ e $\boxed{\sqrt{nI_X(\widehat{\theta})}(\widehat{\theta} - \theta)}$ possono essere usate come *quantità pivotali asintotiche* per la determinazione di intervalli di confidenza per grandi campioni.
Determiniamo intervalli di confidenza asintotici a livello $1-\alpha$ per le seguenti distribuzioni:

1 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{B(\theta), \quad \theta \in (0,1)\}}$
Lo stimatore di max verosimiglianza di θ è la media campionaria $\overline{X}_n$. Ricordiamo l'

espressione della densità di Bernoulli nella forma $f_X(x,\theta) = \theta^x(1-\theta)^{1-x}$ e calcoliamo l' informazione di Fischer:

$$I_X(\theta) = E\left(\frac{\partial}{\partial\theta}\left[\log\left(\theta^X(1-\theta)^{1-X}\right)\right]\right)^2 = E\left(\frac{\partial}{\partial\theta}[X\theta + (1-X)(1-\theta)]\right)^2 =$$
$$= \frac{1}{\theta^2(1-\theta)^2}E\left(X-\theta\right)^2 = \frac{1}{\theta(1-\theta)}$$

Osservazione 2.30. *Se uno stimatore è efficiente, e molti stimatori di max verosimiglianza lo sono, la sua varianza coincide con* $\dfrac{1}{nI_X(\theta)}$*, ed i calcoli precedenti diventano superflui.*

Allora utilizzando la quantità $\sqrt{nI_X(\theta)}(\widehat{\theta}-\theta)$, dalla relazione

$$P\left(-\phi_{1-\frac{\alpha}{2}} < \sqrt{n}\frac{(\widehat{\theta}-\theta)}{\sqrt{\theta(1-\theta)}} < \phi_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

si ricava, risolvendo le disuguaglianze rispetto a θ,

$$P\left(\frac{2n\widehat{\theta} + \phi^2_{1-\frac{\alpha}{2}} - \phi_{1-\frac{\alpha}{2}}\sqrt{4n\widehat{\theta} + \phi^2_{1-\frac{\alpha}{2}} - 4n\widehat{\theta^2}}}{2(n-\phi^2_{1-\frac{\alpha}{2}})} < \theta < \frac{2n\widehat{\theta} + \phi^2_{1-\frac{\alpha}{2}} + \phi_{1-\frac{\alpha}{2}}\sqrt{4n\widehat{\theta} + \phi^2_{1-\frac{\alpha}{2}} - 4n\widehat{\theta^2}}}{2(n-\phi^2_{1-\frac{\alpha}{2}})}\right) =$$

Utilizzando invece la quantità $\sqrt{nI_X(\widehat{\theta})}(\widehat{\theta}-\theta)$, dalla relazione

$$P\left(-\phi_{1-\frac{\alpha}{2}} < \sqrt{n}\frac{(\widehat{\theta}-\theta)}{\sqrt{\widehat{\theta}(1-\widehat{\theta})}} < \phi_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

si ricava, molto più agevolmente,

$$P\left(\widehat{\theta} - \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} < \theta < \widehat{\theta} + \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right) = 1-\alpha.$$

Esempio 2.31. [8], ESEMPIO 7.5.2 PAG. 266
Il 14 ottobre del 1997 il New York Times riportò un sondaggio che indicava che il 52% *della popolazione, con un margine d'errore di* $\pm 4\%$ *era soddisfatta dell'operato del presidente Clinton. Cosa significa? Se le stime sono ottenute ad un livello di confidenza pari a* $1-\alpha = 0.95$ *è possibile stabilire quante persone furono intervistate?*

2 $\boxed{(X_1,...,X_n) \text{ campione estratto dalla distribuzione } \{f(x,\theta) = \theta e^{-\theta x}\mathbb{1}(x>0), \quad \theta \in \mathbb{R}^+\}}$
([5], esempio 8.7 pag 396)

3 $\boxed{(X_1,...,X_n) \text{ campione estratto dalla distribuzione } \{f(x,\theta) = 2\theta x e^{-\theta x^2}\mathbb{1}(x>0), \quad \theta \in \mathbb{R}^+\}}$
([1], esempio 3.7 pag 106)

4 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{f(x,\theta) = \theta x^{(\theta-1)} \mathbb{1}(0 < x < 1), \quad \theta \in \mathbb{R}^+\}}$
([1], esercizio 3.11 pag 127)

5 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{N(0,\theta), \quad \theta > 0\}}$
([6], esempio 19.6 pag 743)

6 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{f(x,\theta) = \theta(1-\theta)^{(x-1)} \quad x \in \mathbb{N}, \quad \theta \in \mathbb{R}^+\}}$
([1], esercizio 3.9 pag 123)

7 $\boxed{(X_1, ..., X_n) \text{ campione estratto dalla distribuzione } \{f(x,\theta) = \frac{\theta^x}{x!} e^{(-\theta)}, \ x \in \mathbb{N} \ \theta \in \mathbb{R}^+\}}$
([1], esercizio 3.7 pag 121)

8 Sia $(X_1, ..., X_n)$ un campione estratto dalla distribuzione $\{f(x,\theta) = \frac{1}{\theta}\mathbb{1}(0 < x < \theta), \quad \theta \in \mathbb{R}^+\}$
Posto $Z_n = n(\theta - \max_{1 \le i \le n} X_i)$, si dimostri che

- Z_n converge in legge ad una variabile aleatoria esponenziale di valor medio θ;
- utilizzando questo risultato, determinare un intervallo di confidenza asintotico per θ.

([1], esercizio 3.12 pag 128)

Osservazione 2.32. *È evidente che invertire la quantità* $\sqrt{nI_X(\widehat{\theta})}(\widehat{\theta} - \theta)$ *è estremamente semplice, perché il fattore* $\sqrt{nI_X(\widehat{\theta})}$ *non contiene il parametro incognito, ma è una statistica; infatti usando questa quantità si ottiene sempre un intervallo di confidenza asintotico della forma*

$$\left[\widehat{\theta} - \frac{\phi_{1-\frac{\alpha}{2}}}{\sqrt{nI_X(\widehat{\theta})}}, \widehat{\theta} + \frac{\phi_{1-\frac{\alpha}{2}}}{\sqrt{nI_X(\widehat{\theta})}}\right].$$

Altra cosa è l' inversione di $\sqrt{nI_X(\theta)}(\widehat{\theta} - \theta)$. *Per contro, a parità di campione e di livello, quest' ultima quantità pivotale produce intervalli migliori, perché utilizza l' esatta informazione di Fischer e non una sua approssimazione.*

Capitolo 3

Test d'ipotesi

3.1 Test parametrici

3.1.1 Descrizione e definizioni

Sia $\{f_X(x;\theta), \quad \theta \in \Theta\}$ un modello statistico parametrico.
Una **ipotesi statistica** è una asserzione circa il valore assunto dal parametro incognito θ della distribuzione $f_X(x;\theta)$.
Una ipotesi statistica può specificare completamente la legge $f_X(x;\theta)$, ed in tal caso viene detta **ipotesi statistica semplice**; in caso contrario si parla di **ipotesi statistica composta**.

Per denotare una ipotesi statistica si usa solitamente il carattere H, seguito dai 2 punti e dalla specificazione della ipotesi.

Esempio 3.1. *Sia $f_X(x;\theta) = Bin(\theta)$, potremmo formulare l'ipotesi $H : p = \frac{1}{2}$ (ipotesi semplice) oppure $H : p < \frac{1}{2}$ (ipotesi composta).*

Un **test** (che indicheremo con la lettera $\mathbf{Y}$) è una regola costruita *sulla base dell' osservazione di un campione* estratto dalla distribuzione in esame, per decidere se rifiutare o meno un'ipotesi. Facciamo un esempio per chiarire:

Esempio 3.2. *Sia $f_X(x;\theta) = N(\theta, 16)$. Formuliamo l'ipotesi $H : \theta = 3$. Si supponga di osservare il campione $(X_1, ..., X_n)$. Un possibile test $\boldsymbol{Y}$ per decidere se rifiutare o meno l'ipotesi H è :"si rifiuti H se e solo se $|\overline{X}_n - 3| > \frac{4}{\sqrt{n}}$".*

Sia n l'ampiezza del campione osservato. Indichiamo con $\mathbf{X_{(n)}}$ lo spazio di tutte le possibili realizzazioni del campione (**spazio campionario**), ovvero:

$$\mathbf{X_{(n)}} = \{(x_1, ..., x_n) \in \mathbb{R}^n \text{ tale che } (X_1, ..., X_n)(\omega) = (x_1, ..., x_n), \omega \in \Omega\}.$$

Un test individua una partizione dello spazio campionario in 2 insiemi disgiunti $\{C, C^c\}$ tale per cui la regola di decisione per il rifiuto di una ipotesi H puo' essere espressa come:

"Si rifiuti H se e solo se $(x_1, ..., x_n) \in C$ "

Il sottoinsieme C è detto **regione critica** del test. La regione critica di un test si esprime sempre tramite una relazione funzionale che coinvolge una statistica, detta **statistica test**. Per chiarire facciamo il seguente

Esempio 3.3. *Nell' esempio precedente la regione critica per l'ipotesi* $H : \theta = 3$ *è* $C = \{(x_1, ..., x_n) \in \mathbb{R}^n$ *tale che* $|\overline{x}_n - 3| > \frac{4}{\sqrt{n}}\}$ *e "una" statistica test è* $|\overline{x}_n - 3|$*. D'altro canto, se esprimiamo la regione critica come* $C = \{(x_1, ..., x_n) \in \mathbb{R}^n$ *tale che* $\overline{x}_n > 3 + \frac{4}{\sqrt{n}}\} \cup \{(x_1, ..., x_n) \in \mathbb{R}^n$ *tale che* $\overline{x}_n < 3 - \frac{4}{\sqrt{n}}\}$ *allora potremmo considerare come statistica test* $\overline{x}_n$*.*

Come vedremo, la scelta della statistica test è legata essenzialmente alla possibilità di poter calcolare i suoi quantili.

Considereremo in questo corso test in cui la regione critica individua univocamente la regola decisione (test non casualizzati).INSERIRE TEST CASUALIZZATI?

Nei problemi che incontreremo, vengono messe a confronto 2 ipotesi: l'ipotesi da verificare, indicata con H_0, detta **ipotesi nulla**, e l'ipotesi H_1, detta **ipotesi alternativa**.
Queste due ipotesi sono esaustive, nel senso che se l'ipotesi nulla è falsa si assume che quella alternativa sia vera e viceversa.
In questo modo un test produce una sola delle seguenti situazioni:

- H_0 è vera ed il test la accetta.
- H_0 è vera ed il test la rifiuta (**errore di I tipo**).
- H_0 è falsa ed il test la accetta (**errore di II tipo**).
- H_0 è falsa ed il test la rifiuta.

Nella verifica di ipotesi l'ipotesi nulla è la più importante, nel senso che si è disposti a rifiutarla solo in caso di grande evidenza empirica del contrario. Per questo motivo commettere l'errore di I tipo è considerato più grave. L'esempio tipico che si fa per spiegare questa situazione è il seguente

Esempio 3.4. *In un processo, si formulano 2 ipotesi:* H_0 *: l'imputato è innocente, ed* H_1 *: l'imputato è colpevole. In questa ottica l'errore di primo tipo produce la condanna di un innocente, mentre quello di II tipo l'assoluzione di un colpevole. Giuridicamente il primo errore è considerato più grave del secondo.*

Intuitivamente un test "buono" è un test che rende "piccola" la possibilità di commettere errori. Ma come si quantifica la grandezza degli errori? Nel resto di questo paragrafo costruiremo l'apparato teorico che ci permetterà di rispondere a questa domanda e definire test ottimali.

Formalizziamo un problema per la verifica delle ipotesi come segue:

Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x; \theta), \quad \theta \in \Theta\}$. Sia C la regione critica corrispondente ad un test per la verifica delle ipotesi $H_0 : \theta \in \Theta_0$, in alternativa ad $H_1 : \theta \in \Theta_1$, dove $\Theta_0 \cup \Theta_1 \subseteq \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$.

Definizione 3.5. *Chiamiamo* ***funzione di potenza del test Y*** *la funzione* $\pi_{\boldsymbol{Y}} : \Theta \in [0, 1]$ *tale che*

$$\pi_{\boldsymbol{Y}}(\theta) = P_\theta(C)^1,$$

[1]La misura di probabilità P_θ è la legge congiunta del campione, calcolabile grazie alle ipotesi di indipendenza

ovvero la probabilità di rifiutare H_0 quando la legge è parametrizzata da θ.

La funzione di potenza dà una misura della bontà del test; un test "buono" ha una funzione di potenza vicina a 0 se $\theta \in \Theta_0$, vicina ad 1 se $\theta \in \Theta_1$. Inoltre è evidente che ogni volta che $\theta \in \Theta_0$ la funzione di potenza descrive la probabilità di commettere un errore del I tipo posto che θ sia il valore esatto del parametro incognito. L'introduzione della funzione di potenza ci permette di controllare la probabilità di commettere un errore di I tipo:

Definizione 3.6. *Chiamiamo **livello di significatività** o **ampiezza** del test la grandezza α definita come*

$$\alpha = \sup_{\theta \in \Theta_0} \pi_{\boldsymbol{Y}}(\theta)$$

L'ampiezza del test definisce la massima probabilità di rifiutare H_0 quando è vera.

Esempio 3.7. *Calcoliamo l'ampiezza del test $\boldsymbol{Y}$ dell'esempio 3.2;*

$$\pi_{\boldsymbol{Y}}(\theta) = P_\theta\left(|\overline{X}_n - 3| > \frac{4}{\sqrt{n}}\right) = 1 - P_\theta\left(3 - \frac{4}{\sqrt{n}} \leq \overline{X}_n \leq 3 + \frac{4}{\sqrt{n}}\right) = \tag{3.1}$$

$$= 1 - P_\theta\left(-1 + \frac{3-\theta}{\frac{4}{\sqrt{n}}} \leq \frac{\overline{X}_n - \theta}{\frac{4}{\sqrt{n}}} \leq 1 + \frac{3-\theta}{\frac{4}{\sqrt{n}}}\right) = \tag{3.2}$$

$$= 1 - \Phi\left(1 + \frac{3-\theta}{\frac{4}{\sqrt{n}}}\right) + \Phi\left(-1 + \frac{3-\theta}{\frac{4}{\sqrt{n}}}\right). \tag{3.3}$$

Allora $\alpha = \pi_{\boldsymbol{Y}}(3) = 1 - \Phi(1) + \Phi(-1) = 2(1 - \Phi(1)) \approx 0.3$

Però non è possibile in generale minimizzare simultaneamente l'errore di I tipo e quello di II tipo. Per rendersene conto, sulla base di un campione $(X_1, ..., X_n)$ estratto dalla distribuzione $\{N(\mu; 1), \quad \mu \in \{2, 4\}\}$ si supponga di dover costruire un test per la verifica delle ipotesi $H_0 : \mu = 2$ in alternativa a $H_1 : \mu = 4$. Si vede allora che una regione critica del tipo $\{\overline{x}_n > c\}$ se c è grande rende piccolo l'errore di I tipo e grande quello di II tipo tanto e viceversa.

Quello che si fa è:

- fissare l'ampiezza del test, ovvero la massima probabilità di commettere un errore di I tipo;
- scegliere tra tutti i test con uguale ampiezza, quello corrispondente ad un errore di II tipo più basso.

Questo procedimento produce un test ottimale. Discuteremo la costruzione di test ottimali in 2 casi:

- verifica di ipotesi semplici in alternativa ad ipotesi semplici;
- verifica di ipotesi composte.

3.1.2 Ipotesi semplici in alternativa ad ipotesi semplici

Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x;\theta), \quad \theta \in \{\theta_1, \theta_2\}\}$. Si vuole costruire un test per la verifica dell'ipotesi

$$H_0 : \theta = \theta_0$$

in alternativa a

$$H_1 : \theta = \theta_1.$$

Vogliamo costruire un test $\mathbf{Y}^*$ che soddisfi il seguente criterio di ottimalità:

Definizione 3.8. *Un test* $\boldsymbol{Y}$ *per la la verifica dell'ipotesi* $H_0 : \theta = \theta_0$ *in alternativa a* $H_1 : \theta = \theta_1$ *è detto* ***test più potente di ampiezza*** α *se e solo se:*

1 $\pi_{\boldsymbol{Y}}(\theta_0) = \alpha$*;*

2 $\pi_{\boldsymbol{Y}}(\theta_1) \geq \pi_{\overline{\boldsymbol{Y}}}(\theta_1)$ *per ogni altro test* $\overline{\boldsymbol{Y}}$ *tale che* $\pi_{\overline{\boldsymbol{Y}}}(\theta_0) \leq \alpha$.

Descriviamo ora un metodo ottimale per la costruzione dei test di ipotesi semplici in alternativa ad ipotesi semplici.
Sia $L(\theta; x_1, ..., x_n) = \Pi_{i=1}^n f_X(x_i;\theta)$ la funzione di verosimiglianza del campione. Nel caso analizzato tale funzione assume 2 soli possibili valori. Definiamo il

Definizione 3.9. ***Test del rapporto di verosimiglianza semplice*** *di ampiezza* α*: è il test* $\boldsymbol{Y}$ *così costruito:*

1 *si determina il numero* k_α *tale che* $P_{\theta_0}\left(\dfrac{L(\theta_0; X_1, ..., X_n)}{L(\theta_1; X_1, ..., X_n)} \leq k_\alpha\right) = \alpha$ [2]

2 *si considera il test* $\boldsymbol{Y}$*: " si rifiuta se e solo se* $(x_1, ..., x_n) \in C$*", dove*

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X}_{(\mathbf{n})} \text{ tali che } \frac{L(\theta_0; x_1, ..., x_n)}{L(\theta_1; x_1, ..., x_n)} \leq k_\alpha\right\}$$

Dalla definizione precedente segue che $\pi_Y(\theta_0) = \alpha$ è l'ampiezza del test. La costruzione di questo test è basata sugli stessi principi che hanno motivato l'introduzione delle stime di massima verosimiglianza. Cioé si rifiuta H_0 solo se è più verosimile che il campione provenga dalla distribuzione parametrizzata da θ_1.

Il seguente risultato assicura che il test del rapporto di verosimiglianza semplice produce test più potenti nel senso indicato nella definizione 3.8:

Teorema 3.10. ***(Lemma di Neyman-Pearson)***
Sia $\boldsymbol{Y}$ *il test del rapporto di verosimiglianza semplice di ampiezza* α*. Allora:*

$$\pi_Y(\theta_1) \geq \pi_{\overline{Y}}(\theta_1)$$

dove $\overline{\boldsymbol{Y}}$ *è un qualunque altro test tale che* $\pi_{\overline{Y}}(\theta_0) \leq \alpha$.
In altre parole $\boldsymbol{Y}$ *è il test più potente di ampiezza* α.

[2] È evidente che si può determinare k_α solo se si conoscono i quantili della legge di $\frac{L(\theta_0; X_1, ..., X_n)}{L(\theta_1; X_1, ..., X_n)}$

Determiniamo il test più potente per la verifica dell'ipotesi $\{H_0 : \theta = \theta_0\}$ in alternativa a $\{H_1 : \theta = \theta_1\}$ nei seguenti casi:

1 $\boxed{(X_1, ..., X_n) \text{ campione aleatorio estratto dalla distribuzione } \{Exp(\theta), \quad \theta \in \{\theta_0, \theta_1\}\}.}$
([5] pag 414, es 9.7).

2 $\boxed{(X_1, ..., X_n) \text{ campione aleatorio estratto dalla distribuzione } \{B(1, \theta), \quad \theta \in \{\theta_0, \theta_1\}, \theta_0 < \theta_1\}.}$
([5] pag 415, es 9.8)NOTARE CHE IN CASO DI LEGGE DISCRETA, NON SEMPRE ESISTE UN TEST PIÙ POTENTE PER OGNI SPECIFICA AMPIEZZA.

3 $\boxed{(X_1, ..., X_n) \text{ campione aleatorio estratto dalla distribuzione } \{N(\theta, 36), \quad \theta \in \{0, 1.2\}\}.}$
([1] pag 134, es 4.2)

4 ESERCIZI SU [1]. INTERESSANTI SU [7] PAG 418

3.1.3 Ipotesi composte: test del rapporto di verosimiglianza generalizzato

Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x; \theta), \quad \theta \in \Theta\}$.
Si vuole costruire un test per la verifica dell'ipotesi

$$H_0 : \theta \in \Theta_0$$

in alternativa a

$$H_1 : \theta \in \Theta_1.$$

dove $\Theta_0 \cup \Theta_1 \subseteq \Theta$[3], $\Theta_0 \cap \Theta_1 = \emptyset$.

Anche in questo caso definiamo un criterio di ottimalità per valutare possibili test. Ovviamente in caso di ipotesi composte, sia le tecniche di costruzione che la valutazione dell' ottimalità dei test si complicano.

Definizione 3.11. *Un test* $\boldsymbol{Y}$ *per la la verifica dell'ipotesi* $H_0 : \theta \in \Theta_0$ *in alternativa a* $H_1 : \theta \in \Theta_1$ *è detto* ***test uniformemente più potente di ampiezza*** α *se e solo se:*

1 $\sup_{\theta \in \Theta_0} \pi_{\boldsymbol{Y}}(\theta) = \alpha$*;*

2 $\pi_{\boldsymbol{Y}}(\theta) \geq \pi_{\overline{\boldsymbol{Y}}}(\theta)$*per ogni* $\theta \in \Theta_1$ *e per ogni altro test* $\overline{\boldsymbol{Y}}$ *tale che* $\sup_{\theta \in \Theta_0} \pi_{\overline{\boldsymbol{Y}}}(\theta) \leq \alpha$.

Occupiamoci ora del problema della costruzione dei test nel caso di massima generalità delle ipotesi in discussione. Descriviamo un metodo sempre perseguibile, ancora una volta motivato dal principio di massima verosimiglianza.

Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x; \theta), \quad \theta \in \Theta\}$. Costruiamo un test per $H_0 : \theta \in \Theta_0$ in alternativa a $H_1 : \theta \in \Theta - \Theta_0$.

Definizione 3.12. ***Test del rapporto di verosimiglianza generalizzato*** *di ampiezza* α*: è il test* $\boldsymbol{Y}$ *così costruito:*

[3] In molte situazioni si avrà $\Theta_0 \cup \Theta_1 = \Theta$.

1 *si determina il numero* k_α *tale che* $\sup_{\theta_0 \in \Theta_0} P_{\theta_0}\left(\frac{\sup_{\theta\in\Theta_0} L(\theta; X_1, ..., X_n)}{\sup_{\theta\in\Theta} L(\theta; X_1, ..., X_n)} \leq k_\alpha\right) = \alpha$

2 *si considera il test* ***Y****: " si rifiuta se e solo se* $(x_1, ..., x_n) \in C$*", dove*

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\sup_{\theta\in\Theta_0} L(\theta; x_1, ..., x_n)}{\sup_{\theta\in\Theta} L(\theta; x_1, ..., x_n)} \leq k_\alpha\right\}$$

Il fatto di non fissare il tipo di ipotesi da verificare né alcuna proprietà della legge $f_X(x;\theta)$ si riflette sul fatto che questo test non è necessariamente un test uniformemente più potente. Tuttavia stabilisce un metodo per costruire sempre esplicitamente una statistica test

$$\Lambda_n = \frac{\sup_{\theta\in\Theta_0} L(\theta; X_1, ..., X_n)}{\sup_{\theta\in\Theta} L(\theta; X_1, ..., X_n)}$$

ed una regione di rifiuto $C = \{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \Lambda_n < k\}$.
Il problema più arduo da affrontare resta comunque il calcolo esplicito della soglia k, che, per un test di ampiezza α deve coincidere con il quantile di ordine α della legge di Λ_n. Se però il campione in esame è abbastanza numeroso e per particolari ipotesi, questo calcolo si può fare in modo approssimato usando una proprietà asintotica di una trasformazione della statistica test Λ_n. Vedremo questo approccio nel paragrafo 3.4.1..

Esempio 3.13. *([5] pag 421)*
Sia $(X_1, ..., X_n)$ *un campione aleatorio estratto dalla distribuzione* $\{\theta e^{-\theta x}\mathbb{1}_{(0,\infty)}(x), \quad \theta \in (0, \infty)\}$*. Costruiamo un test basato sul rapporto di verosimiglianza generalizzato per* $H_0 : \theta \leq \theta_0$ *in alternativa a* $H_1 : \theta > \theta_0$.

Esempio 3.14. *([6] pag 648)*
Sia $(X_1, ..., X_n)$ *un campione aleatorio estratto dalla distribuzione* $\{e^{-(x-\theta)}\mathbb{1}_{(\theta,\infty)}(x), \quad \theta \in (0, \infty)\}$*. Costruiamo un test basato sul rapporto di verosimiglianza generalizzato per* $H_0 : \theta = \theta_0$ *in alternativa a* $H_1 : \theta \neq \theta_0$.

Esempio 3.15. *([1] pag 188 esercizio 4.16)*
Sia $(X_1, ..., X_n)$ *un campione aleatorio estratto dalla distribuzione* $\{\theta x^{\theta-1}\mathbb{1}_{(0,1)}(x), \quad \theta \in (0, \infty)\}$*. Si trovi il test di rapporto di verosimiglianza generalizzato per* $H_0 : \theta = 1$ *in alternativa a* $H_1 : \theta \neq 1$.

3.1.4 Test uniformemente più potente per ipotesi unilaterali

Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x;\theta), \quad \theta \in \Theta\}$ e sia Θ **un intervallo di** $\mathbb{R}$.

Vogliamo costruire test per la verifica di ipotesi del seguente tipo

$$\boxed{H_0 : \theta \leq \theta_0 \text{ in alternativa a } H_1 : \theta > \theta_0.}$$

Caso 1 $\boxed{\text{Sia } \Theta \text{ un intervallo e sia } \{f_X(x;\theta) = a(\theta)b(x)e^{c(\theta)d(x)}\theta \in \Theta\}}$[4]

[4]Densità di questo tipo sono dette di *classe esponenziale*. Molte leggi note vi appartengono, $(exp(\theta), P(\theta), B(1,\theta), N(\theta_1, \theta_2)$, ecc... Nel caso di parametri multidimensionali come ad esempio $N(\theta_1, \theta_2)$, il prodotto nella potenza dell' esponenziale è un prodotto scalare.)

Poniamo $t(x_1, ..., x_n) = \sum_{i=1}^{n} d(x_i)$. Vale allora il seguente

Teorema 3.16. *Se*

(i) *$c(\theta)$ è una funzione monotona crescente ed esiste k_α tale che $P_{\theta_0}(t(X_1, ..., X_n) > k_\alpha) = \alpha$[5], allora il test Y corrispondente alla regione critica*

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } t(x_1, ..., x_n) > k_\alpha\right\}$$

*è il **test uniformemente più potente di ampiezza** α per la verifica delle ipotesi descritte.*

(ii) *$c(\theta)$ è una funzione monotona decrescente ed esiste k_α tale che $P_{\theta_0}(t(X_1, ..., X_n) < k_\alpha) = \alpha$, allora il test Y corrispondente alla regione critica*

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } t(x_1, ..., x_n) < k_\alpha\right\}$$

*è il **test uniformemente più potente di ampiezza** α per la verifica delle ipotesi descritte.*

Osservazione 3.17. *Il test descritto nel precedente teorema è anche il test uniformemente più potente di ampiezza α per la verifica dell' ipotesi*

$$\boxed{H_0 : \theta = \theta_0 \text{ in alternativa a } H_1 : \theta > \theta_0.}$$

Esempio 3.18. *([5] pag 424)*
Consideriamo lo stesso problema di verifica d'ipotesi affrontato nell' esempio 3.13. Per risolverlo abbiamo fatto calcoli abbastanza laboriosi. Alla luce di questo teorema invece, dobbiamo verificare che la legge $f(x;\theta)$ è di classe esponenziale e la monotonia della corrispondente funzione $c(\theta)$. Banalmente $\{\theta e^{-\theta x}\mathbb{1}_{(0,\infty)}(x), \quad \theta \in (0,\infty)\}$ è della forma indicata per $a(\theta) = \theta,\ b(x) = \mathbb{1}_{(0,\infty)}(x),\ c(\theta) = -\theta,\ d(x) = x$. Inoltre la funzione $c(\theta)$ è decrescente per $\theta \in (0,\infty) = \Theta$. Siamo quindi nella situazione (ii) descritta nel precedente teorema, e questo assicura che il test Y corrispondente alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \sum_{i=1}^{n} x_i < k_\alpha\right\}$$

dove k_α è definito dalla relazione $P_{\theta_0}(\sum_{i=1}^{n} X_i < k_\alpha) = \alpha$ è il test uniformemente più potente di ampiezza α per la verifica delle ipotesi descritte.
Per calcolare la soglia k_α dobbiamo conoscere la legge della statistica $\sum_{i=1}^{n} X_i$. Sappiamo che le $X_i, \quad i = 1, ..., n$ sono tra di loro indipendenti con legge comune $exp(\theta)$. Quindi, se $\theta = \theta_0$ $\theta_0 X_i \sim exp(1), \quad i = 1, ..., n$ e $\theta_0 \sum_{i=1}^{n} X_i \sim \Gamma(n, 1)$. Pertanto

$$P_{\theta_0}(\sum_{i=1}^{n} X_i < k_\alpha) = P_{\theta_0}(\theta_0 \sum_{i=1}^{n} X_i < \theta_0 k_\alpha)$$

[5]Con le ipotesi formulate sul modello basta un semplice calcolo per verificare che $\sup_{\theta \le \theta_0} \pi_Y(\theta) = \pi_Y(\theta_0)$

ovvero $k_\alpha = \frac{q_\alpha}{\theta_0}$, *dove* q_α *è il quantile di ordine* α *della legge* $\Gamma(n,1)$ *calcolabile risolvendo l' equazione*[6]

$$\int_0^{q_\alpha} \frac{1}{\Gamma(n)} x^{n-1} e^{-x} dx.$$

Alternativamente si può utilizzare la trasformazione $2\theta_0 \sum_{i=1}^n X_i \sim \Gamma\left(n, \frac{1}{2}\right) = \chi^2(2n)$ *che permette di calcolare* k_α *risolvendo l'equazione* $k_\alpha = \frac{\chi^2_{\alpha,2n}}{2\theta_0}$, *dove* $\chi^2_{\alpha,2n}$ *è il quantile di ordine* α *della distribuzione chi-quadrato con* $2n$ *gradi di libertà.*

Esempio 3.19. *([1] pag 143 esempio 4.5)*
Un fertilizzante è messo in vendita con la pretesa di essere più efficace di un altro correntemente in uso. Per verificare questa affermazione lo si utilizza in 25 zone agricole, tra di loro omogenee e si misura il raccolto $(X_1, ..., X_{25})$ *ottenuto in ciascuna zona. Supponendo che siano determinazioni di una variabile aleatoria normale di varianza* 4 *e media incognita* θ, *si costruisca il test UPP di ampiezza* 0.05 *per la verifica dell'ipotesi* $H_0 : \theta \leq \theta_0$ *in alternativa a* $H_1 : \theta > \theta_0$, *avendo indicato con* θ_0 *il raccolto medio ottenuto con il vecchio fertilizzante.*

Caso 2 Sia Θ un intervallo e sia $\{f_X(x;\theta)\ \theta \in \Theta\}$ tale che esiste una statistica $t(X_1, ..., X_n)$ tale per cui, per ogni coppia $\{\theta_1, \theta_2\} \in \Theta$, con $\theta_1 < \theta_2$ il rapporto di verosimiglianza

$$\frac{L(\theta_1; x_1, ..., x_n)}{L(\theta_2; x_1, ..., x_n)}$$

sia una funzione non crescente o non decrescente di $t(x_1, ..., x_n)$.

Una tale famiglia di densità si dice famiglia con ***rapporto di verosimiglianza monotono***.

Esempio 3.20. *[5] 9.14 pag 424. Scrivere il rapporto di verosimiglianza per il modello parametrico* $\{f_X(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x),\ \ \theta > 0\}$ *e determinare una statistica rispetto alla quale è un rapporto di verosimiglianza monotono.*

Esempio 3.21. *[5] 9.15 pag 424. Scrivere il rapporto di verosimiglianza per il modello parametrico* $\{f_X(x;\theta) = \frac{1}{\theta} I_{(0,\theta)}(x),\ \ \theta > 0\}$ *e determinare una statistica rispetto alla quale è un rapporto di verosimiglianza monotono.*

Nel caso di famiglie di densità con rapporto di verosimiglianza monotono vale il seguente

Teorema 3.22. *Se*

(i) *il rapporto di verosimiglianza monotono è non decrescente in* $t(x_1, ..., x_n)$ *ed esiste* k_α *tale che* $P_{\theta_0}(t(X_1, ..., X_n) < k_\alpha) = \alpha$, *allora il test* Y *corrispondente alla regione critica*

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } t(x_1, ..., x_n) < k_\alpha\right\}$$

è il ***test uniformemente più potente di ampiezza*** α *per la verifica delle ipotesi descritte.*

[6] oppure usando un pacchetto statistico

(ii) *il rapporto di verosimiglianza monotono è non crescente in $t(x_1, ..., x_n)$ ed esiste k_α tale che $P_{\theta_0}(t(X_1, ..., X_n) > k_\alpha) = \alpha$, allora il test Y corrispondente alla regione critica*

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } t(x_1, ..., x_n) > k_\alpha\right\}$$

è il ***test uniformemente più potente di ampiezza*** α *per la verifica delle ipotesi descritte.*

Esempio 3.23. *[5] 9.16 pag 425. Sia $X_1, ..., X_n$ un campione casuale estratto dal modello $\{f_X(x; \theta) = \frac{1}{\theta} I_{(0,\theta)}(x), \quad \theta > 0\}$ e costruire un test per la verifica dell'ipotesi $H_0 : \theta \leq \theta_0$ in alternativa ad $H_0 : \theta > \theta_0$. Si tratta di un test uniformemente più potente?*

Osservazione 3.24. *I teoremi descritti in questa sezione sono validi anche nel caso in cui si invertano le ipotesi H_0 ed H_1, a patto che si invertano le disuguaglianze che definiscono le regioni critiche.*

3.2 p-value

Abbiamo già osservato che il livello di significatività di un test rappresenta la massima probabilità di commettere un errore di I tipo.
In realtà quando si implementa un test utilizzando pacchetti statistici, nell'input non viene digitato il livello del test, perché comunque nell'output compare il valore di una statistica nota come ***livello di significatività*** osservato (oppure ***p-value***).
Il **p-value** è il più piccolo livello per cui si potrebbe rigettare l'ipotesi nulla con i dati ottenuti nelle osservazioni, e, per questo, fornisce una misura di quanto i dati si accordino all'ipotesi nulla.
Vediamo come calcolare il p-value. Per fare un esempio, fissiamo le idee su un test definito da una regione critica del tipo

$$C = \{(x_1, ..., x_n) \in X_{(n)} : \ T(x_1, ..., x_n) \geq t\}$$

Possiamo allora definire la funzione $\alpha : \mathbb{R}^+ \to [0, 1]$ tale che

$$\alpha(t) = \sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, ..., X_n) \geq t\right)$$

Evidentemente, se il livello α del test è fissato, allora la soglia t_α della regione critica è determinata dalla soluzione dell'equazione $\alpha = \sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, ..., X_n) \geq t_\alpha\right)$.
Invece il più piccolo livello al quale è possibile rigettare l'ipotesi nulla si ottiene andando a *calcolare la funzione $\alpha(t)$ sulla statistica test osservata*, ovvero

$$\alpha \left(T(x_1, ..., x_n)\right) = \sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, ..., X_n) \geq T(x_1, ..., x_n)\right)$$

Tale valore definisce appunto il p-value del test.

Il p-value non sostituisce il livello del test, che è stabilito a priori e che è tanto più piccolo quanto più si ritenga grave commettere un errore di primo tipo.

Il calcolo del p-value aiuta nella discussione delle ipotesi, e evita il calcolo esplicito la soglia della regione critica; infatti, se si ottiene un p-value più piccolo del livello fissato, allora si rigetta l'ipotesi nulla, altrimenti la si accetta.

Osserviamo infine che la funzione $\alpha(t)$ altro non è che la massima probabilità di commettere un errore di primo tipo in funzione della soglia t della regione critica; di conseguenza, un p- value molto basso indica che, in accordo con i dati osservati, commettere un errore di primo tipo è molto improbabile e quindi si può rigettare l'ipotesi nulla.

3.3 Verifica di ipotesi per campionamento da popolazioni normali

Accennare proprietà di ottimalità di test bilateri non distorti per giustificare le regioni di confidenza che troveremo (bene su [5])

In questo paragrafo costruiremo test per la media e per la varianza nel caso speciale di campionamento da una distribuzione normale. A tale proposito consideriamo $(X_1, ..., X_n)$ un campione estratto da una distribuzione $N(\mu, \sigma^2)$.

3.3.1 Test sulla media

Ipotesi unilaterali Siano date le seguenti ipotesi a confronto

$$H_0 : \mu \leq \mu_0 \text{ in alternativa a } H_1 : \mu > \mu_0$$

Per la costruzione del test si possono distinguere 2 casi

1 **σ^2 è un valore noto**.

Abbiamo già osservato che la densità normale è di classe esponenziale con decomposizione

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

Sono inoltre verificate le ipotesi del Teorema 3.16 e di conseguenza il test uniformemente più potente di ampiezza α corrisponde alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} : \sum_{i=1}^{n} X_i > k'_\alpha \right\} = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} : \overline{X}_n > k_\alpha \right\}$$

dove

$$\alpha = \sup_{\mu \leq \mu_0} P_\mu \left(\overline{X}_n > k_\alpha \right)$$

Per calcolare la soglia k_α ricordiamo che $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, e quindi

$$\sup_{\mu \leq \mu_0} P_\mu\left(\overline{X}_n > k_\alpha\right) = \sup_{\mu \leq \mu_0}\left(1 - \Phi\left(\frac{k_\alpha - \mu}{\sigma}\sqrt{n}\right)\right) = 1 - \Phi\left(\frac{k_\alpha - \mu_0}{\sigma}\sqrt{n}\right)$$

Da cui si ricava $\frac{k_\alpha - \mu_0}{\sigma}\sqrt{n} = \phi_{1-\alpha}$, ovvero $k_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha}$.

Riassumendo, il test uniformemente più potente di ampiezza α per le ipotesi in discussione corrisponde alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \overline{X}_n > \mu_0 + \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha}\right\}$$

o, equivalentemente

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} > \phi_{1-\alpha}\right\}$$

Osservazione 3.25. *La regione critica appena descritta corrisponde anche al test uniformemente più potente per la verifica dell'ipotesi*

$$H_0 : \mu = \mu_0 \quad \textit{in alternativa a} \quad H_1 : \mu > \mu_0$$

Osservazione 3.26. *Con tecniche analoghe, sempre grazie al Teorema 3.16, si dimostra che il test uniformemente più potente per la verifica dell'ipotesi*

$$H_0 : \mu \geq \mu_0 \quad \textit{in alternativa a} \quad H_1 : \mu < \mu_0$$

è definito dalla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \textit{ tali che } \overline{X}_n < \mu_0 + \frac{\sigma}{\sqrt{n}}\phi_\alpha\right\}$$

o, equivalentemente

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \textit{ tali che } \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} < \phi_\alpha\right\}$$

Esempio 3.27. [8], ESEMPIO 8.3.6 PAG. 304
Tutti i tipi di sigarette attualmente presenti sul mercato hanno un contenuto medio di nicotina non inferiore a 1.6 mg. Una marca di tabacchi afferma però di aver individuato un particolare trattamento delle foglie di tabacco che permette di abbassare il livello medio di nicotina al di sotto di 1.6 mg. Per verificare questa affermazione si analizza un campione di 20 sigarette di questa marca, trovando una media campionaria pari a 1.54 mg. Supponendo che la deviazione standard della popolazione sia di 0.8 mg, fissando un livello di significatività pari a 0.05, cosa decide il test?

2 σ^2 **è un valore incognito**. In tal caso i Teoremi 3.16 e 3.22 non si applicano poiché *lo spazio dei parametri* Θ *non è un intervallo della retta reale.*
Più precisamente $\Theta = \{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$.

Ciononostante, possiamo sempre fare riferimento alla statistica $\sum_{i=1}^n X_i$ per discriminare tra le ipotesi in discussione. Infatti tale statistica tenderà ad essere "piccola" sotto H_0 e "grande" sotto H_1. Di conseguenza anche in questo caso possiamo considerare il test di ampiezza α corrispondente alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \overline{X}_n > k_\alpha\right\}$$

dove

$$\alpha = \sup_{\mu \leq \mu_0} P_\mu\left(\overline{X}_n > k_\alpha\right)$$

Per calcolare la soglia k_α ricordiamo che $\dfrac{\bar{X}_n - \mu_0}{S_n}\sqrt{n} \sim t_{n-1}$

$$\sup_{\mu \leq \mu_0} P_\mu\left(\overline{X}_n > k_\alpha\right) = \sup_{\mu \leq \mu_0}\left(1 - F_{n-1}\left(\frac{k_\alpha - \mu}{S_n}\sqrt{n}\right)\right) = 1 - F_{n-1}\left(\frac{k_\alpha - \mu_0}{S_n}\sqrt{n}\right)$$

avendo indicato con F_{n-1} la funzione di ripartizione della legge t_{n-1}. Di conseguenza $\frac{k_\alpha - \mu_0}{S_n}\sqrt{n} = t_{1-\alpha,n-1}$, ovvero $k_\alpha = \mu_0 + \frac{S_n}{\sqrt{n}} t_{1-\alpha,n-1}$.

Quindi il test di ampiezza α descritto corrisponde alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \overline{X}_n > \mu_0 + \frac{S_n}{\sqrt{n}} t_{1-\alpha,n-1}\right\}$$

o, equivalentemente

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\overline{X}_n - \mu_0}{S_n}\sqrt{n} > t_{1-\alpha,n-1}\right\}$$

Anche in questo caso continuano a valere le osservazioni 3.25 e 3.26 con S_n in luogo di σ e $t_{\alpha,n-1}$ in luogo di ϕ_α.

Concludiamo osservando che il test ricavato corrisponde ad *un test del rapporto di verosimiglianza generalizzato di ampiezza* α (dimostrare!!!)

Esempio 3.28. [8], ESEMPIO 8.3.7 PAG. 309
Tra i pazienti di una clinica con un tasso di colesterolo alto si cercano dei volontari per testare un nuovo farmaco che dovrebbe aiutare a ridurre il tasso di colesterolo. Si scelgono 50 volontari e si somministra il farmaco per un mese. Alla fine si registra una riduzione media del tasso di colesterolo pari a $\overline{X}_{50} = 14.8$ *con deviazione standard campionaria pari a* $S_{50} = 6.4$. *Assumendo che la riduzione del tasso di colesterolo segua una legge normale, costruire un test di ampiezza 0.05 per verificare l'efficienza del farmaco.*

Esempio 3.29. [8], ESEMPIO 8.3.9 PAG. 311
Il produttore di un nuovo tipo di pneumatico in fibra di vetro afferma che la vita media del suo prodotto è di almeno $40 \cdot 10^3$ *miglia. Allo scopo di verificare questa affermazione si prende un campione di 12 pneumatici e si registrano i seguenti tempi di vita (in unità di* 10^3 *miglia)*

Gomma	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Vita	36.1	40.2	33.8	38.5	37	41	36.8	37.2	33	42	35.8	36

Sotto l'ipotesi di normalità, si verifichi l'affermazione del produttore con un test di ampiezza 0.05. Si calcoli anche il p-value del test.

Ipotesi bilaterali Costruiamo un test per la verifica dell'ipotesi

$$H_0 : \mu = \mu_0 \text{ in alternativa a } H_1 : \mu \neq \mu_0$$

Osserviamo che un metodo sempre praticabile quando le ipotesi in discussione siano di tipo bilatero, è il *metodo dell'intervallo di confidenza*. In pratica si costruisce un intervallo di confidenza $[T_1, T_2]$ per μ di livello $1-\alpha$ e si considera il test definito dalla regola: **" si rifiuta H_0 se $\mu_0 \notin [T_1, T_2]$".**

Si ottiene così un test di ampiezza α. Infatti

$$P_{\mu_0}(\text{rifiuto } H_0) = 1 - P_{\mu_0}(\mu_0 \in [T_1, T_2]) = 1 - (1-\alpha) = \alpha.$$

Utilizziamo questo metodo nei due casi distinti

1 σ^2 **è un valore noto**.
In tal caso l'intervallo di confidenza per μ di livello $1-\alpha$ ha l'espressione

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

Di conseguenza il test di ampiezza α basato su questo intervallo corrisponde alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \mu_0 \notin \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right]\right\} =$$
$$= \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \left|\frac{\bar{X}_n - \mu_0}{\sigma}\sqrt{n}\right| > \phi_{1-\frac{\alpha}{2}}\right\}$$

2 σ^2 **è un valore incognito**. L'espressione dell'intervallo di confidenza in questa situazione è

$$\left[\bar{X}_n - \frac{S_n}{\sqrt{n}}t_{1-\frac{\alpha}{2},n-1}, \bar{X}_n + \frac{S_n}{\sqrt{n}}t_{1-\frac{\alpha}{2},n-1}\right]$$

e quindi il test di ampiezza α basato su questo intervallo corrisponde alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \mu_0 \notin \left[\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1}, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \right] \right\} =$$
$$= \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \left| \frac{\bar{X}_n - \mu_0}{S_n} \sqrt{n} \right| > t_{1-\frac{\alpha}{2},n-1} \right\}$$

È possibile dimostrare che i test costruiti in questo paragrafo coincidono con i test del rapporto di verosimiglianza generalizzato.
Per completezza si segnala che questi test sono uniformemente più potenti su un particolare sottoinsieme della classe di tutti i possibili test bilateri[7]

Esempio 3.30. [8], ESEMPIO 8.3.1 PAG. 295

Esempio 3.31. [8], ESEMPIO 8.3.8 PAG. 309

3.3.2 Test per la varianza

Ipotesi unilaterali Siano date le seguenti ipotesi a confronto

$$H_0 : \sigma^2 \leq \sigma_0^2 \text{ in alternativa a } H_1 : \sigma^2 > \sigma_0^2$$

Distinguiamo i 2 casi

1 **μ è un valore noto**. In tal caso, essendo le ipotesi di tipo unilaterale e Θ un intervallo, applicando il Teorema 3.16 è possibile determinare il test più potente di ampiezza α; infatti, direttamente dall'espressione della densità normale $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ si ricava che un tale test corrisponde alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \sum_{i=1}^{n} (x_i - \mu)^2 > k_\alpha \right\}$$

dove

$$\alpha = \sup_{\sigma^2 \leq \sigma_0^2} P_{\sigma^2} \left(\sum_{i=1}^{n} (X_i - \mu)^2 > k_\alpha \right)$$

Ricordiamo che $\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$, e di conseguenza, indicando con F_{χ_n} la funzione di ripartizione della legge χ^2_n

$$\sup_{\sigma^2 \leq \sigma_0^2} P_{\sigma^2} \left(\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} > \frac{k_\alpha}{\sigma^2} \right) = \sup_{\sigma^2 \leq \sigma_0^2} \left(1 - F_{\chi_n} \left(\frac{k_\alpha}{\sigma^2} \right) \right) = 1 - F_{\chi_n} \left(\frac{k_\alpha}{\sigma_0^2} \right)$$

[7] Per approfondimenti si rimanda a [5]

Da cui si ricava $\frac{k_\alpha}{\sigma_0^2} = \chi^2_{1-\alpha,n}$, ovvero $k_\alpha = \sigma_0^2 \chi^2_{1-\alpha,n}$.

Pertanto il test uniformemente più potente di ampiezza α per le ipotesi in discussione corrisponde alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} > \chi^2_{1-\alpha,n} \right\}$$

2 μ **è un valore incognito**. In questo caso utilizziamo $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ come statistica test.
Poiché a meno di una costante questa statistica stima correttamente la varianza, tenderà ad essere "piccola" sotto H_0 e "grande" sotto H_1. Costruiamo quindi il test di ampiezza α corrispondente alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \sum_{i=1}^n (x_i - \bar{x}_n)^2 > k_\alpha \right\}$$

dove

$$\alpha = \sup_{\sigma^2 \leq \sigma_0^2} P_{\sigma^2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 > k_\alpha \right)$$

Poiché $\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2_{n-1}$, indicando con $F_{\chi_{n-1}}$ la funzione di ripartizione della legge χ^2_{n-1} si ottiene

$$\sup_{\sigma^2 \leq \sigma_0^2} P_{\sigma^2} \left(\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2} > \frac{k_\alpha}{\sigma^2} \right) = \sup_{\sigma^2 \leq \sigma_0^2} \left(1 - F_{\chi_{n-1}} \left(\frac{k_\alpha}{\sigma^2} \right) \right) = 1 - F_{\chi_{n-1}} \left(\frac{k_\alpha}{\sigma_0^2} \right)$$

Da cui $\frac{k_\alpha}{\sigma_0^2} = \chi^2_{1-\alpha,n-1}$, ovvero $k_\alpha = \sigma_0^2 \chi^2_{1-\alpha,n-1}$.

In definitiva il test di ampiezza α basato sulla statistica $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ corrisponde alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_0^2} > \chi^2_{1-\alpha,n-1} \right\}$$

Osserviamo che anche in questo caso il test costruito coincide col test del rapporto di verosimiglianza generalizzato.

Osservazione 3.32. *In analogia con l'osservazione 3.25 sottolineamo che le regioni critiche determinate in questo paragrafo corrispondono al test di ampiezza α per la verifica dell'ipotesi*

$$H_0 : \sigma = \sigma_0 \quad \textit{in alternativa a} \quad H_1 : \sigma > \sigma_0.$$

In particolare, nel caso in cui μ sia nota, si ottiene il test uniformemente più potente.

Osservazione 3.33. *Le stesse tecniche usate in questo paragrafo, applicate la verifica dell'ipotesi*

$$H_0 : \sigma \geq \sigma_0 \quad \textit{in alternativa a} \quad H_1 : \sigma < \sigma_0$$

producono il test definito dalla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \textit{ tali che } \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma_0^2} < \chi^2_{\alpha,n} \right\}$$

nel caso in cui μ sia nota, e il test definito dalla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \textit{ tali che } \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{\sigma_0^2} < \chi^2_{\alpha,n-1} \right\}$$

altrimenti.

In particolare, se μ è nota si ottiene il test uniformemente più potente di ampiezza α.

Esempio 3.34. [8], ESEMPIO 8.5.1 PAG. 324

Ipotesi bilaterali Analizziamo ora le seguenti ipotesi a confronto

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{in alternativa a} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

Per la costruzione del test utilizziamo il metodo dell'intervallo di confidenza. Come usuale, si possono distinguere 2 casi

1 μ **è un valore noto**. Ricordiamo che l'intervallo di confidenza di livello $1 - \alpha$ a code equiprobabili ha l'espressione

$$\left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2},n}}, \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2},n}} \right]$$

e quindi il test di ampiezza α basato su questo intervallo corrisponde alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \sigma_0^2 \notin \left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2},n}}, \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2},n}} \right] \right\}$$

2 μ **è un valore incognito** In tal caso l'intervallo di confidenza a code equiprobabili di livello $1 - \alpha$ è

$$\left[\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}{\chi^2_{\frac{\alpha}{2},n-1}} \right]$$

e quindi il test di ampiezza α basato su questo intervallo corrisponde alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \sigma_0^2 \notin \left[\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}{\chi^2_{\frac{\alpha}{2},n-1}} \right] \right\}$$

È possibile infine dimostrare che i test costruiti in questo paragrafo coincidono con i test del rapporto di verosimiglianza generalizzato.

3.3.3 Test per il confronto tra medie

Consideriamo due campioni indipendenti $(X_1, ..., X_n)$ estratto da una distribuzione $N(\mu_1, \sigma_1^2)$ e $(Y_1, ..., Y_m)$ estratto da una distribuzione $N(\mu_2, \sigma_2^2)$.
Si vogliono costruire test di ampiezza α per la verifica di ipotesi riguardanti la differenza dei valori medi delle due popolazioni.
Per semplificare utilizzeremo in tutto il paragrafo la statistica test $\bar{X}_n - \bar{Y}_m$ per discriminare le ipotesi in discussione. Sottolineamo però che, nel caso di ipotesi unilaterali con varianze note, il Teorema 3.16 garantisce che il test basato si tale statistica è il test uniformemente più potente di ampiezza α.

Ipotesi unilaterali Siano date le seguenti ipotesi a confronto

$$H_0 : \mu_1 \leq \mu_2 \text{ in alternativa a } H_1 : \mu_1 > \mu_2$$

Per la costruzione del test si possono distinguere 2 casi

1 σ_1^2, σ_1^2 **sono noti**.
Ricordiamo che $\bar{X}_n - \bar{Y}_m \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$.

In accordo con le ipotesi in discussione, costruiamo il test corrispondente alla regione critica

$$C = \left\{(x_1, ..., x_n, y_1, ..., y_m) \in \mathbf{X_{(n)}} \otimes \mathbf{X_{(m)}} \text{ tali che } \bar{X}_n - \bar{Y}_m > k_\alpha\right\}$$

dove

$$\alpha = \sup_{\mu_1 - \mu_2 \leq 0} P_{\mu_1 - \mu_2}\left(\bar{X}_n - \bar{Y}_m > k_\alpha\right)$$

Con i calcoli usuali si ottiene

$$\alpha = 1 - \Phi\left(\frac{k_\alpha}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}\right)$$

Da cui si ricava $\frac{k_\alpha}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \phi_{1-\alpha}$, ovvero $k_\alpha = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\alpha}$.

In definitiva il test descritto di ampiezza α corrisponde alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \bar{X}_n - \bar{Y}_m > \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\alpha}\right\}$$

o, equivalentemente

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > \phi_{1-\alpha}\right\}$$

Esempio 3.35. [8], ESEMPIO 8.4.1 PAG. 315

2 $\sigma_1^2 = \sigma_1^2 = \sigma^2$ **è un valore incognito**.
Usiamo sempre la statistica $\bar{X}_n - \bar{Y}_m$ come statistica test, ovvero scegliamo il test corrispondente alla regione critica

$$C = \left\{(x_1, ..., x_n, y_1, ..., y_m) \in \mathbf{X_{(n)}} \otimes \mathbf{X_{(m)}} \text{ tali che } \bar{X}_n - \bar{Y}_m > k_\alpha\right\}$$

dove

$$\alpha = \sup_{\mu_1 - \mu_2 \leq 0} P_{\mu_1 - \mu_2}\left(\bar{X}_n - \bar{Y}_m > k_\alpha\right)$$

Per calcolare la soglia k_α ricordiamo che $\dfrac{\bar{X}_n - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$, dove $S_p^2 = \frac{(n-1)S_{1n}^2 + (m-1)S_{2m}^2}{n+m-2}$ è la varianza campionaria conglobata.

Indicando con $F_{t_{n+m-2}}$ la funzione di ripartizione della legge t_{n+m-2}, si ottiene

$$\alpha = 1 - F_{t_{n+m-2}}\left(\frac{k_\alpha}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}\right)$$

Da cui si ricava $\frac{k_\alpha}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = t_{1-\alpha,n+m-2}$, ovvero $k_\alpha = S_p\sqrt{\frac{1}{n} + \frac{1}{m}}t_{1-\alpha,n+m-2}$.

In definitiva il test descritto di ampiezza α corrisponde alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \bar{X}_n - \bar{Y}_m > S_p\sqrt{\frac{1}{n} + \frac{1}{m}}t_{1-\alpha,n+m-2}\right\}$$

o, equivalentemente

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\bar{X}_n - \bar{Y}_m}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{1-\alpha,n+m-2}\right\}$$

Ribadiamo infine che con un opportuno cambio di notazioni continuano a valere le Osservazioni 3.25 e 3.26.

Esempio 3.36. [8], ESEMPIO 8.5.1 PAG. 324

Ipotesi bilaterali Si vuole costruire un test per la verifica dell'ipotesi

$$H_0 : \mu_1 = \mu_2 \text{ in alternativa a } H_1 : \mu_1 \neq \mu_2$$

A tale scopo utilizziamo il metodo dell'intervallo di confidenza. Come usuale, distinguiamo i due casi

1 σ_1^2, σ_2^2 **sono noti**. Ricordiamo che l'espressione dell'intervallo di confidenza simmetrico di livello $1-\alpha$ è la seguente

$$\left[\bar{X}_n - \bar{Y}_m - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\frac{\alpha}{2}}, \bar{X}_n - \bar{Y}_m + \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\frac{\alpha}{2}}\right]$$

E quindi il test di ampiezza α basato su tale intervallo corrisponde alla regione critica

$$\begin{aligned} C = &\left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \right. \\ &\left. 0 \notin \left[\bar{X}_n - \bar{Y}_m - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\frac{\alpha}{2}}, \bar{X}_n - \bar{Y}_m + \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\phi_{1-\frac{\alpha}{2}}\right]\right\} = \\ = &\left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \left|\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}\right| > \phi_{1-\frac{\alpha}{2}}\right\} \end{aligned}$$

2 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ **è un valore incognito.** In tal caso l'espressione dell'intervallo di confidenza simmetrico di livello $1-\alpha$ diventa

$$\left[\bar{X}_n - \bar{Y}_m - \sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}t_{1-\frac{\alpha}{2},n+m-2}, \bar{X}_n - \bar{Y}_m + \sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}t_{1-\frac{\alpha}{2},n+m-2}\right],$$

dove S_p^2 è la varianza campionaria conglobata, e quindi il test di ampiezza α basato su tale intervallo corrisponde alla regione critica

$$\begin{aligned} C = &\left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \right. \\ &\left. 0 \notin \left[\bar{X}_n - \bar{Y}_m - \sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}t_{1-\frac{\alpha}{2},n+m-2}, \bar{X}_n - \bar{Y}_m + \sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}t_{1-\frac{\alpha}{2},n+m-2}\right.\right. \\ = &\left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \left|\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}\right| > t_{1-\frac{\alpha}{2},n+m-2}\right\} \end{aligned}$$

Esempio 3.37. [8], ESEMPIO 8.4.2 PAG. 318

3.3.4 Test per il confronto tra varianze

Come nella sezione precedente, siano dati due campioni indipendenti $(X_1, ..., X_n)$ estratto da una distribuzione $N(\mu_1, \sigma_1^2)$ e $(Y_1, ..., Y_m)$ estratto da una distribuzione $N(\mu_2, \sigma_2^2)$.
Si vogliono costruire test di ampiezza α per la verifica di ipotesi riguardanti le varianze delle due popolazioni.
Come di consueto, studieremo separatamente il caso di ipotesi unilaterali e bilaterali.

Ipotesi unilaterali Siano date le seguenti ipotesi a confronto

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \text{ in alternativa a } \sigma_1^2 > \sigma_2^2$$

Dalla teoria degli stimatori sappiamo che la varianza campionaria è uno stimatore consistente della varianza teorica.[8] Quindi, indicando con S^2_{Xn} e S^2_{Ym} rispettivamente le varianze campionarie del primo e del secondo campione, il rapporto $\frac{S^2_{Xn}}{S^2_{Ym}}$ può essere usato per discriminare tra le ipotesi in discussione. Pertanto consideriamo il test di ampiezza α che corrisponde alla regione critica

$$C = \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} \geq k \right\}$$

dove k è la soluzione dell'equazione

$$\alpha = \sup_{\frac{\sigma_1^2}{\sigma_2^2} \leq 1} P_{\sigma_1^2, \sigma_2^2} \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} \geq k \right\}$$

Per determinare k ricordiamo che

$$\frac{(n-1)S^2_{Xn}}{\sigma_1^2} \sim \chi^2_{n-1} \text{ e } \frac{(m-1)S^2_{Ym}}{\sigma_2^2} \sim \chi^2_{m-1}$$

per modo tale che posto

$$F = \frac{S^2_{Xn}\sigma_2^2}{S^2_{Ym}\sigma_1^2}$$

la variabile aleatoria F è una F di Fischer con $n-1$ ed $m-1$ gradi di libertà.

Allora

$$P_{\sigma_1^2, \sigma_2^2} \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} \geq k \right\} = P_{\sigma_1^2, \sigma_2^2} \left\{ \frac{S^2_{Xn}\sigma_2^2}{S^2_{Ym}\sigma_1^2} \geq k\frac{\sigma_2^2}{\sigma_1^2} \right\} = 1 - F_{n-1,m-1}\left(k\frac{\sigma_2^2}{\sigma_1^2} \right)$$

e quindi

$$\alpha = \sup_{\frac{\sigma_1^2}{\sigma_2^2} \leq 1} \left[1 - F_{n-1,m-1}\left(k\frac{\sigma_2^2}{\sigma_1^2} \right) \right] = 1 - F_{n-1,m-1}(k)$$

per modo tale che $k = F_{1-\alpha,n-1,m-1}$. In definitiva un test di ampiezza α per le ipotesi in discussione corrisponde alla regione critica

$$C = \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} \geq F_{1-\alpha,n-1,m-1} \right\}$$

Ipotesi bilaterali Siano date le seguenti ipotesi a confronto

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ in alternativa a } \sigma_1^2 \neq \sigma_2^2$$

.

[8] Inoltre coincide con lo stimatore di massima verosimiglianza della varianza nel caso in cui la media non sia un valore noto. Per verificare di questa proprietà è sufficiente verificare che la funzione di verosimiglianza $L(\mu, \sigma^2)$ assume valore massimo nel punto $(\overline{X}_n, S^2_n)$.

Utilizzando la stessa statistica test del caso precedente si costruisce il test di ampiezza α corrispondente ad regione critica del tipo

$$C = \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} \leq k_1 \right\} \bigcup \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} \geq k_2 \right\}$$

dove k_1 e k_2 sono due quantili della F di Fischer con $n-1$ ed $m-1$ gradi di libertà scelti in modo tale che il test abbia l'ampiezza fissata.

In generale, in analogia con gli intervalli di confidenza a code equiprobabili, si scelgono i quantili in modo tale che, sotto H_0, le due regioni della regione critica abbiano stessa area pari ad $\frac{\alpha}{2}$ e questo corrisponde a scegliere $k_1 = F_{\frac{\alpha}{2},n-1,m-1}$ e $k_2 = F_{1-\frac{\alpha}{2},n-1,m-1}$, ovvero

$$C = \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} < F_{\frac{\alpha}{2},n-1,m-1} \right\} \bigcup \left\{ \frac{S^2_{Xn}}{S^2_{Ym}} > F_{1-\frac{\alpha}{2},n-1,m-1} \right\}.$$

Esempio 3.38. *[8] Esempio 8.5.2 pag.324*

3.4 Test del chi quadrato

3.4.1 Test asintotici basati sul rapporto di verosimiglianza generalizzato

Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{f_X(x;\theta),\ \theta = (\theta_1, ..., \theta_k)\}$. Ricordiamo che il test del rapporto di verosimiglianza generalizzato per la verifica dell'ipotesi $H_0 : \theta \in \Theta_0$ in alternativa a $H_1 : \theta \in \Theta - \Theta_0$ corrisponde ad una regione critica del tipo $C = \big\{(x_1, ..., x_n) \in \mathbf{X_{(n)}}$ tali che $\Lambda_n \leq k\big\}$, dove Λ_n è la statistica

$$\Lambda_n = \frac{\sup_{\theta\in\Theta_0} L(\theta; X_1, ..., X_n)}{\sup_{\theta\in\Theta} L(\theta; X_1, ..., X_n)}$$

Sebbene Λ_n sia una statistica, a volte la sua distribuzione è così complessa da rendere molto difficoltosa la ricerca dei quantili, indispensabili per calcolare esplicitamente la regione critica corrispondente ad un test di ampiezza α.

Quando però la dimensione n del campione è sufficientemente grande e $\Theta \subseteq \mathbb{R}^k$, per verificare l'ipotesi

$$H_0 : \theta \in \Theta_0$$

in alternativa a

$$H_1 : \theta \in \Theta - \Theta_0$$

dove

$$\Theta_0 = \{\theta \in \Theta : \theta_1 = \theta_1^0, ..., \theta_r = \theta_r^0, \theta_{r+1}, ...\theta_k\},$$

si utilizza il seguente risultato asintotico

Teorema 3.39. *Sotto opportune ipotesi di regolarità per la densità $f_X(x;\theta)$ ([7] pag 384 T. 8.7.4), allora, per ogni $\theta \in \Theta_0$*

$$-2\ln\Lambda_n \underset{n\to+\infty}{\to} \chi^2_r$$

.

Esempio 3.40. ([6], PAG 652) *Sia $(X_1, ..., X_n)$ campione aleatorio estratto dalla distribuzione $\{B(\theta), \quad \theta \in [0,1]\}$. Vogliamo verificare l'ipotesi $H_0 : \theta = \theta_0$ in alternativa a $H_1 : \theta \neq \theta_0$. Scriviamo il rapporto di verosimiglianza generalizzato per la densità $f_X(x;\theta) = \theta^x \times (1-\theta)^{1-x} \quad \theta \in [0,1]$:*

$$\Lambda_n = \frac{\theta_0^{\sum_{i=1}^n X_i} \times (1-\theta_0)^{n-\sum_{i=1}^n X_i}}{\overline{X}_n^{\sum_{i=1}^n X_i} \times (1-\overline{X}_n)^{n-\sum_{i=1}^n X_i}}$$

quindi

$$\begin{aligned}-2\log\Lambda_n =& -2n\left(\overline{X}_n\log\theta_0 + (1-\overline{X}_n)\log(1-\theta_0) - \overline{X}_n\log\overline{X}_n - (1-\overline{X}_n)\log(1-\overline{X}_n)\right) = \\ =& -2n\left(\overline{X}_n\log\frac{\theta_0}{\overline{X}_n} + (1-\overline{X}_n)\log\frac{(1-\theta_0)}{(1-\overline{X}_n)}\right)\end{aligned}$$

Evidentemente la legge di una tale statistica è pressoché impossibile da trattare. Però se il campione è abbastanza numeroso, allora sotto l'ipotesi H_0, $-2\log\Lambda_n \sim \chi(1)$, e quindi, la regione critica del test del rapporto di verosimiglianza generalizzato di ampiezza α si determina come segue

$$\alpha = P_{\theta_0}(\Lambda_n \leq k) = P_{\theta_0}(-2\log\Lambda_n \geq h)$$

ovvero $h = \chi^2_{1-\alpha,1}$. Il test descritto corrisponde quindi alla regione critica

$$C = \left\{(x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } -2\log\Lambda_n(x_1, ..., x_n) \geq \chi^2_{1-\alpha,1}\right\}$$

Esempio 3.41. [1] ESEMPIO 4.12 PAG 158
Sia $X \sim N(\mu, \sigma^2)$, μ noto e σ^2 incognito. Costruire un test asintotico basato sul rapporto di verosimiglianza generalizzato per la verifica dell'ipotesi $\sigma^2 = \sigma_0^2$ in alternativa a $\sigma^2 \neq \sigma_0^2$

Esempio 3.42. [1] ESEMPIO 4.13 PAG 159
Sia $X \sim N(\mu, \sigma^2)$, μ incognito e σ^2 incognito. Costruire un test asintotico basato sul rapporto di verosimiglianza generalizzato per la verifica dell'ipotesi $\sigma^2 = \sigma_0^2$ in alternativa a $\sigma^2 \neq \sigma_0^2$

VEDERE ANCHE 4.16, 4.17, 4.19, 4.20

3.4.2 Test di adattamento

In questo paragrafo descriviamo un test costruito per verificare l' adattamento dei dati osservati ad una legge fissata. Questa legge può essere completamente specificata, oppure specificata a meno di un certo numero di parametri.
Premettiamo il teorema

Teorema 3.43. *Siano date n variabili aleatorie $Y_1, ..., Y_n$, indipendenti, ed identicamente distribuite a valori in $\{1, ..., k\}$. Indichiamo con $p_j = P(Y_1 = j)$ e con $Z_j = \sum_{i=1}^{n} \mathbb{1}(Y_i = j)$, $j = 1, ..., k$. Allora la funzione di ripartizione della variabile aleatoria*

$$\sum_{j=1}^{k} \frac{(Z_j - np_j)^2}{np_j} \tag{3.4}$$

converge alla funzione di ripartizione di una variabile aleatoria $\chi^2(k-1)$.

Osservazione 3.44. *Le variabili aleatorie $Z_j \quad j = 1, .., k$ definite nel precedente teorema sono binomiali, ovvero $Z_j \sim Bin(n, p_j), \quad j = 1, .., k$.*

Questo teorema fornisce un metodo per stimare la bontà di adattamento dei dati ad una legge completamente specificata. Descriviamo di seguito la procedura.

Sia $X_1, ..., X_n$ un campione estratto da una legge F incognita. Si vuole costruire un test per la verifica dell'ipotesi

$$H_0 : F = F_0$$

in alternativa a

$$H_1 : F \neq F_0$$

dove F_0 è una legge nota.
Si considera una partizione finita (ma sufficientemente ampia) $A_1, ..., A_k$ di $\mathbb{R}$ e si considerano le k variabili aleatorie $Z_j = \sum_{i=1}^{n} \mathbb{1}(X_i \in A_j)$, $j = 1, ..., k$, dove $Z_j \sim Bin(n, p_j), \quad j = 1, .., k$.
Ora, se H_0 è vera, $p_j = P_0(X_1 \in A_j)$, avendo indicato con P_0 la misura di probabilità corrispondente alla legge F_0. Sembra ragionevole supporre allora, che quando H_0 è vera la statistica 3.4 tende ad essere piccola. Pertanto un test possibile è definito da una regione critica del tipo

$$C = \left\{ \sum_{j=1}^{k} \frac{(z_j - np_j)^2}{np_j} > k \right\}$$

e, se n è sufficientemente grande[9], per ottenere un test di ampiezza α basta porre $k = \chi^2_{1-\alpha,k-1}$.

Esempio 3.45. [8], Es. 11.2.2 pag. 468
Un produttore di lampade ad incandescenza informa i suoi clienti che la qualità dei suoi prodotti non è uniforme, e che ogni lampadina può essere indipendentemente di qualità A, B, C, D o E con probabilità del 15%, 25%, 35%, 20% e 5% rispettivamente. Tuttavia uno dei clienti acquistando grossi volumi di merce ha l'impressione di ricevere troppi pezzi di qualità E (la peggiore). Decide quindi di verificare l'affermazione del produttore testando 30 lampade. Il risultato dell'esperimento è: 3 pezzi di qualità A, 6 di qualità B, 9 di qualità C, 7 di qualità D e 5 di qualità E. I dati osservati confermano l'affermazione del produttore ad un livello di significatività del 5%?

[9]Una regola per stabilire quando n è sufficientemente grande è che l' 80% delle np_i deve essere maggiore di 5 e le restanti maggiori di 1.

Esempio 3.46. [5], Es. 9.21 pag. 446
La teoria di Mendel indica che la forma ed il colore di una certa varietà di piselli dovrebbero essere suddivisi in 4 gruppi, I°:"lisci e gialli", II°:"lisci e verdi", III°:"rugosi e gialli" e IV°:"rugosi e verdi" secondo i rapporti 9/3/3/1. Per $n = 556$ *piselli si rilevano le seguenti osservazioni:*

lisci e gialli	315
lisci e verdi	108
rugosi e gialli	101
rugosi e verdi	32

Indicata con p_i *la probabilità che un pisello appartenga all'i-simo gruppo, costruire un test di ampiezza* 0.05 *per la verifica dell'ipotesi* $H_0 : \ p_1 = \frac{9}{16}, \ p_2 = \frac{3}{16}, \ p_3 = \frac{3}{16}, \ p_4 = \frac{1}{16}$.

Nel caso in cui si voglia costruire un test per la verifica dell' adattamento dei dati ad una legge specificata a meno di un certo numero di parametri, si utilizza il seguente risultato

Teorema 3.47. *Siano date n variabili aleatorie* $Y_1, ..., Y_n$*, indipendenti ed identicamente distribuite con densità comune* $\{f_X(j;\theta), \ j \in \{1, ..., k\}, \ \theta = (\theta_1, ..., \theta_r), \quad k > r - 1\}$[10]*. Indichiamo con* $p_j(\theta) = P_\theta(Y_1 = j)$ *e con* $Z_j = \sum_{i=1}^n \mathbb{1}(Y_i = j)$*,* $j = 1, ..., k$*. Supponiamo che esistano gli stimatori di massima verosimiglianza* $(\widehat{\theta}_1, ..., \widehat{\theta}_r)$ *dei parametri* $(\theta_1, ..., \theta_r)$ *e sia* $\widehat{p}_j = p_j(\widehat{\theta})$*. Allora la funzione di ripartizione della variabile aleatoria*

$$\sum_{j=1}^{k} \frac{(Z_j - n\widehat{p}_j)^2}{n\widehat{p}_j} \tag{3.5}$$

converge alla funzione di ripartizione di una variabile aleatoria $\chi^2(k - r - 1)$.

Abbiamo ora uno strumento per stimare la bontà di adattamento dei dati ad una legge specificata a meno di un certo numero di parametri incogniti.
Più precisamente, sia $X_1, ..., X_n$ un campione estratto da una legge F incognita. Si vuole costruire un test per la verifica la verifica dell'ipotesi

$$H_0 : F = F_\theta, \quad \theta = (\theta_1, ..., \theta_r)\}$$

in alternativa a

$$H_1 : F \neq F_\theta$$

dove F_θ è una legge la cui forma funzionale è nota, ma dipende da r parametri incogniti.

Analogamente a quanto già visto, si considera una partizione finita (ma sufficientemente ampia, si ricordi che c' è il vincolo $k > r-1$) $A_1, ..., A_k$ di $\mathbb{R}$ e si considerano le k variabili aleatorie $Z_j = \sum_{i=1}^n \mathbb{1}(X_i \in A_j)$, $j = 1, ..., k$.
Se H_0 è vera, $Z_j \sim Bin(n, p_j(\theta)), \quad j = 1, .., k$ con $p_j(\theta) = P_\theta(X_1 \in A_j)$, dove P_θ è la misura di probabilità corrispondente alla legge F_θ. Sembra ragionevole supporre allora, che quando H_0 è vera la statistica 3.5 tende ad essere piccola[11]. Consideriamo quindi il test definito dalla regione critica

$$C = \left\{ \sum_{j=1}^{k} \frac{(z_j - n\widehat{p}_j)^2}{n\widehat{p}_j} > k \right\}$$

[10] ovvero variabili aleatorie a valori finiti con legge dipendente da r parametri incogniti.
[11] Ricorda la proprietà di invarianza degli stimatori di massima verosimiglianza.

Se n è abbastanza grande, per ottenere un test di ampiezza α basta porre $k = \chi^2_{1-\alpha,k-r-1}$.

Esempio 3.48. [8], Es. 11.3.1 pag. 473
Supponiamo che il numero di incidenti settimanali in un periodo di 30 settimane sia il seguente

8	0	0	1	3	4	0	2	12	5	1	8	0	2	0
1	9	3	4	5	3	3	4	7	4	0	1	2	1	2

Si verifichi l'ipotesi che la distribuzione del numero di incidenti settimanali sia di Poisson.

3.4.3 Test di indipendenza

Siano X ed Y due variabili aleatorie di legge incognita. Supponiamo di osservare un campione di dimensione n $((X_1, Y_1), ..., (X_n, Y_n))$ estratto dalla legge della coppia (X, Y). Si vuole costruire un test per la verifica dell'ipotesi

$$H_0 : X \text{ e } Y \text{ sono indipendenti;}$$

in alternativa a

$$H_1 : X \text{ e } Y \text{ non sono indipendenti.}$$

A tale proposito consideriamo due partizioni finite $A_1, ..., A_r$ e $B_1, ..., B_s$[12] di $\mathbb{R}$ e consideriamo le rs variabili aleatorie

$$Z_{i,j} = \sum_{k=1}^{n} \mathbb{1}((X_k, Y_k) \in A_i \times B_j), \quad k = 1, ..., n$$

Sia

$$p_{i,j} = P\{(X, Y) \in A_i \times B_j\} = P\{X \in A_i, Y \in B_j\}$$

Dall'osservazione 3.44 si evince che $Z_{i,j} \sim Bin(n, p_{i,j})$, $i = 1, ..., r$; $j = 1, ..., s$. Quindi il Teorema 3.43 permette di stabilire che, quando n è sufficientemente grande,

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(Z_{i,j} - np_{i,j})^2}{np_{i,j}} \approx \chi^2_{rs-1}.$$

Sia ora

$$p_{i,\cdot} = P\{X \in A_i\}, \quad p_{\cdot,j} = P\{Y \in B_j\}$$

Se H_0 è vera, allora $p_{i,j} = p_{i,\cdot}p_{\cdot,j}$, $i = 1, ..., r$; $j = 1, ..., s$. Ma le quantità $p_{i,j}, p_{i,\cdot}, p_{\cdot,j}$ non sono note, dunque le stimiamo con le stime di massima verosimiglianza, che sono:

$$\widehat{p}_{i,\cdot} = \sum_{j=1}^{s} \frac{Z_{i,j}}{n}, \quad \widehat{p}_{\cdot,j} = \sum_{i=1}^{r} \frac{Z_{i,j}}{n}.$$

In questo modo i parametri stimati sono $r + s - 2$ (infatti $1 = \sum_{i=1}^{r} = \widehat{p}_{i,\cdot} = \sum_{j=1}^{s} \widehat{p}_{\cdot,j}$).

[12] Una regola per la scelta delle partizioni è che sia $P(X \in A_i) \simeq \frac{1}{r}$ e $P(Y \in B_j) \simeq \frac{1}{s}$, $i = 1, ..., r$; $j = 1, ..., s$ ed inoltre $\frac{n}{r} \geq 5$, $\frac{n}{s} \geq 5$.

Sembra ragionevole supporre allora che, quando H_0 è vera, la statistica

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(Z_{i,j}-n\widehat{p}_{i,\cdot}\widehat{p}_{\cdot,j})^2}{n\widehat{p}_{i,\cdot}\widehat{p}_{\cdot,j}}$$

tenda ad essere piccola. Inoltre, poiché $(rs-1)-(r+s-2)=(r-1)(s-1)$, il Teorema 3.47 assicura che

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(Z_{i,j}-n\widehat{p}_{i,\cdot}\widehat{p}_{\cdot,j})^2}{n\widehat{p}_{i,\cdot}\widehat{p}_{\cdot,j}} \underset{H_0}{\approx} \chi^2_{(r-1)(s-1)}.$$

Pertanto un test di ampiezza α è definito da una regione critica del tipo

$$C=\left\{\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(Z_{i,j}-n\widehat{p}_{i,\cdot}\widehat{p}_{\cdot,j})^2}{n\widehat{p}_{i,\cdot}\widehat{p}_{\cdot,j}}>k\right\},$$

dove $k=\chi^2_{1-\alpha,(r-1)(s-1)}$.

Esempio 3.49. [5], ESERCIZIO 48 PAG. 478. *Gilby ha classificato 1725 bambini di una scuola secondo l'intelligenza e l'apparente livello economico della famiglia. Una classificazione riassuntiva è riportata di seguito:*

	Tardo	*Intelligente*	*Molto capace*
Molto ben vestito	*81*	*322*	*233*
Ben vestito	*141*	*457*	*153*
Poveramente vestito	*127*	*163*	*48*

Verificate l'ipotesi di indipendenza al livello 0.01.

3.5 Test non parametrici

Ricordiamo che in ambito non parametrico, l'inferenza statistica è volta ad ottenere informazioni su una legge di probabilità la cui forma è incognita. In tale situazione abbiamo già costruito test di ipotesi asintotici utilizzando le leggi chi-quadro; sostanzialmente ci si riduce sempre ad un problema di tipo parametrico (SPECIFICARE MEGLIO). Nei prossimi paragrafi vedremo invece come si possa stimare una funzione di ripartizione incognita e come da questa stima si possa costruire un test di adattamento ad una legge nota.

3.5.1 La funzione di ripartizione empirica

Sia $(X_1,...,X_n)$ un campione estratto da una legge con funzione di ripartizione $F(\cdot)$.

Definizione 3.50. *Si chiama* ***funzione di ripartizione empirica***[13] $F_n:(\mathbb{R},\Omega)\to[0,1]$ *la funzione definita tramite la regola*

$$F_n(x,\omega)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(-\infty,x]}(X_i(\omega)) \tag{3.6}$$

[13]La dimostrazione che la funzione di ripartizione empirica sia effettivamente una funzione di ripartizione è lasciata per esercizio.

Per ogni $x \in \mathbb{R}$, $F_n(x)$ è una statistica che indica la frequenza relativa dei valori campionari minori o uguali a x.
È inoltre evidente che per ogni $x \in \mathbb{R}$

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

e di conseguenza $F_n(x)$ è una variabile aleatoria a valori in $\left\{0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1\right\}$ con densità pari a

$$P\left(F_n(x) = \frac{k}{n}\right) = \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}$$

Pertanto

$$E[F_n(x)] = F(x),\ Var(F_n(x)) = \frac{F(x)[1 - F(x)]}{n}$$

Sono inoltre verificate le ipotesi del teorema del limite centrale e quindi per ogni $x \in \mathbb{R}$

$$P\left(\frac{F_n(x) - F(x)}{\sqrt{F(x)[1 - F(x)]}}\sqrt{n} \leq t\right) \underset{n\to+\infty}{\to} \Phi(t)$$

Riassumiamo questi risultati nel seguente

Lemma 3.51. *Per ogni $x \in \mathbb{R}$ la funzione di ripartizione empirica $F_n(x)$ è uno stimatore asintoticamente normale, corretto e consistente della funzione di ripartizione teorica $F(x)$.*

Questo primo risultato mette in luce l 'importanza della funzione di ripartizione empirica come strumento di stima non parametrica di una funzione di ripartizione teorica. Più precisamente stabilisce che $F_n(x)$ è una stima consistente della funzione di ripartizione teorica $F(x)$ *di una legge qualunque* nel punto assegnato x. Per calcolarla si segue la seguente procedura:
si ordina la sequenza delle osservazioni in senso crescente e si indica con $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ la sequenza così ottenuta.
La funzione di ripartizione empirica si scrive quindi in termini del riordinamento $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ come segue:

$$F_n(x) = \begin{cases} 0 & \text{se } x < X_{(1)}; \\ \frac{k}{n} & \text{se } X_{(k)} \leq x < X_{(k+1)}, \quad k = 1, ..., n-1; \\ 1 & \text{se } x \geq X_{(n)}; \end{cases}$$

Esempio 3.52. *Supponiamo di aver osservato un campione di dimensione $n = 10$ estratto da una distribuzione F e di aver rilevato i valori -2, 0.1, -0.7, 2, 1.5, 2, 0.1, -1.5, 1.5, 0. Calcolare la funzione di ripartizione empirica F_{10}.*

Se si è interessati alla stima di $F(\cdot)$ in ogni punto dell'asse reale, ovvero se si è interessati a valutare la distanza tra $F_n(x)$ e $F(x)$ per ogni x, si ha bisogno di un risultato più forte del precedente, risultato stabilito dal seguente teorema

Teorema 3.53. ***(Teorema di Glivenko-Cantelli)***
Sia $\{X_n\}_{n\in N}$ *una successione di variabili aleatorie indipendenti identicamente distribuite con funzione di ripartizione comune* $F(\cdot)$. *Detto* $D_n(\omega) = \sup_{x\in\mathbb{R}} |F_n(x,\omega) - F(x)|$, *si ha che*

$$P\left(\omega: \quad \lim_{n\to+\infty} D_n(\omega) = 0\right) = 1 \tag{3.7}$$

I test non parametrici per la verifica di una ipotesi nulla semplice $H_0: \ F = F_0$, sono noti con il nome di ***test di buon adattamento***.
Il Teorema di Glivenko-Cantelli stabilisce che se H_0 è vera la quantità D_n tende ad essere piccola. È inoltre possibile mostrare che *se* F_0 *è una legge continua, la distribuzione di* D_n *non dipende da* F_0 e quindi D_n può essere utilizzata come statistica test per costruire un test per la verifica dell' ipotesi

$$H_0: \ F = F_0$$

in alternativa a

$$H_1: \ F \neq F_0$$

Più precisamente si considera il test di ampiezza α definito dalla regione critica

$$C = \{(x_1, ..., x_n) \in \mathbf{X_{(n)}}: \ D_n > k_\alpha\}$$

dove $k_\alpha = D_{n,1-\alpha}$, avendo indicato con $D_{n,\beta}$[14] il quantile di ordine β della legge di D_n sotto H_0.
Resta da vedere come si calcola la statistica D_n. A tale scopo osserviamo che la funzione di ripartizione empirica $F_n(x)$ è costante a tratti, mentre la funzione di ripartizione teorica è non decrescente, e pertanto l'estremo superiore della differenza $|F_n(x,\omega) - F(x)|$ deve necessariamente essere assunto nel limite destro o sinistro di uno dei punti di salto. Questa considerazione permette di utilizzare per il calcolo di D_n la seguente formula

$$D_n(\omega) = \sup_{x\in\mathbb{R}} |F_n(x,\omega) - F(x)| = \max_{j=1,...,n} \left\{ \left| \frac{j}{n} - F\left(X_{(j)}\right) \right| \vee \left| \frac{j-1}{n} - F\left(X_{(j)}\right) \right| \right\}$$

Esempio 3.54. *Sono stati registrati* $n = 6$ *tempi di vita in ore di un componente elettrico. Queste misurazioni, ordinate in senso crescente, sono le seguenti*

$$445, \ 478, \ 587, \ 611, \ 654, \ 720$$

Verificare con un test di ampiezza 0.01 *se tali tempi di vita possano essere considerati esponenziali di media 520.*

Esempio 3.55. *Il peso espresso in chilogrammi di 6 ragazzi di una squadra di basket è*

$$68.1, \ 65.2, \ 69.7, \ 72.8, \ 74.2, \ 62.4$$

Verificare ad un livello di significatività del 0.1 *se il peso si possa ritenere distribuito con legge normale di media 70Kg e deviazione standard 4Kg. Il valore critico di interesse è* $D_{6;0.9} = 0.468$.

[14] I quantili $D_{n,\beta}$ di questa legge sono stati calcolati per valori di n piccoli, ma questi calcoli diventano laboriosi al crescere di n.

Si vuole costruire un test per la verifica dell'ipotesi

$$H_0: \ F \sim N(70, 16)$$

in alternativa a

$$H_1: \ F \neq N(70, 16)$$

A tale proposito si costruisce la seguente tabella

x_i	$F_0(x_i)$	$F_6(x_i)$	$\lvert F_0(x_i) - \frac{i}{6}\rvert \vee \lvert F_0(x_i) - \frac{i-1}{6}\rvert$
62.4	0.029	$\frac{1}{6}$	0.138
65.2	0.115	$\frac{2}{6}$	0.218
68.1	0.316	$\frac{3}{6}$	0.184
69.7	0.468	$\frac{4}{6}$	0.198
72.8	0.758	$\frac{5}{6}$	0.075
74.2	0.853	1	0.147

Quindi $D_6 = 0:218$ e poichè $D_6 = 0:218 < D_{6;0.9} = 0:468$ non posso rigettare l'ipotesi H_0.
Per campioni sufficientemente numerosi si usano risultati asintotici che stabiliscono la legge limite della statistica test oppure di una sua opportuna trasformazione.
Uno dei più noti test di buon adattamento che sfrutta la legge limite di D_n è oggetto del prossimo paragrafo.

3.5.2 Il test di adattamento di Kolmogorov e Smirnov

Costruiamo ora un test di adattamento ad una legge ***continua*** definita dalla funzione di ripartizione $F(x)$ per grandi campioni. Quest test è noto con il nome di ***test di Kolmogorov e Smirnov***.
Il risultato asintotico che sfrutteremo è il seguente

Teorema 3.56. *Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie indipendenti con stessa funzione di ripartizione continua $F(x)$. Sia inoltre $F_n(x,\omega)$ la funzione di ripartizione campionaria e $D_n(\omega) = \sup_{x\in\mathbb{R}} |F_n(x,\omega) - F(x)|$. Allora per ogni $t \in \mathbb{R}$*

$$\lim_{n\to+\infty} P\left(\sqrt{n}D_n \leq t\right) = \left[1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2t^2}\mathbb{1}_{(0,+\infty)}(t)\right]. \tag{3.8}$$

La quantità a secondo membro della 3.8 si indica con $H(t)$. I principali quantili di questa funzione di ripartizione sono talvolta tabulati, anche se, per campioni di dimensione $n \geq 35$, possono essere approssimati utilizzando solo il primo termine della serie in 3.8. Quindi, per approssimare il quantile di ordine $1-\alpha$ basta risolvere rispetto a t l' equazione:

$$1 - \alpha = P\left(\sqrt{n}D_n \leq t\right) \approx 1 - 2e^{-2t^2}$$

da cui si ricava

$$t = \sqrt{-\frac{1}{2}\log\left(\frac{\alpha}{2}\right)}$$

Inoltre la legge asintotica di $\sqrt{n}D_n$ non dipende dalla legge iniziale F se non per il fatto che tale legge è continua. Questo permette di utilizzare $\sqrt{n}D_n$ come statistica test per costruire un test di buon adattamento ad una legge continua. Vediamo come.

Sia dato un campione di dimensione n $(X_1, ..., X_n)$ estratto da una distribuzione F. Si vuole costruire un test per la verifica dell'ipotesi

$$H_0 : \ F = F_0$$

in alternativa a

$$H_1 : \ F \neq F_0$$

dove F_0 definisce una legge continua su $\mathbb{R}$.

Il Teorema 3.53 garantisce che la funzione di ripartizione empirica converge uniformemente alla funzione di ripartizione teorica che regola il campione. Di conseguenza se n è sufficientemente numeroso, la statistica $\sup_{x\in\mathbb{R}} |F_n(x,\omega) - F_0(x)|$ tenderà ad essere piccola se H_0 è vera, grande altrimenti. Ovviamente la stessa cosa vale per la trasformazione deterministica $\sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x,\omega) - F_0(x)|$ la cui legge limite, se H_0 è vera, è nota.

Di conseguenza costruiamo il test di ampiezza α corrispondente alla regione critica

$$C = \left\{ (x_1, ..., x_n) \in X_{(n)} : \quad \sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x,\omega) - F_0(x)| > k_\alpha \right\}$$

dove k_α soddisfa l'equazione

$$P_0 \left(\sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x,\omega) - F_0(x)| > k_\alpha \right) = 1 - H(k_\alpha) = \alpha$$

da cui $k_\alpha = h_{1-\alpha}$, avendo indicato $h_{1-\alpha}$ il quantile di ordine $1-\alpha$ della funzione di ripartizione $H(\cdot)$.

Esempio 3.57. *([7] Esempio 3 pag 542)*
Un algoritmo per la simulazione di una v.a. N(0,1) fornisce i seguenti valori:

0.464	0.137	2.455	−0.323	−0.068
0.906	−0.513	−0.525	0.595	0.881
−0.482	1.678	−0.057	−1.229	−0.486
−1.787	−0.261	1.237	1.046	−0.508

Verificare tramite un test di ampiezza $\alpha = 0,05$ *se la procedura di simulazione è corretta.*

Utilizziamo il test di Kolmogorov e Smirnov. Ordiniamo in una tabella i dati necessari per il calcolo della statistica test:

x_i	$F_0(x_i)$	$F_{20}(x_i)$	$\lvert F_{20}(x_i) - F_0(x_i)\rvert \vee \lvert F_{20}(x_{i-1}) - F_0(x_i)\rvert$
-1.787	0.367	$\frac{1}{20}$	0.133
-1.229	0.1093	$\frac{2}{20}$	0.0093
-0.525	0.2998	$\frac{3}{20}$	0.1498
-0.513	0.3050	$\frac{4}{20}$	0.1050
-0.508	0.3050	$\frac{5}{20}$	0.0550
-0.486	0.3121	$\frac{6}{20}$	0.121
-0.482	0.3156	$\frac{7}{20}$	0.0344
-0.323	0.3745	$\frac{8}{20}$	0.0255
-0.261	0.3974	$\frac{9}{20}$	0.0526
-0.068	0.4721	$\frac{10}{20}$	0.0279
-0.057	0.4761	$\frac{11}{20}$	0.0739
0.137	0.5557	$\frac{12}{20}$	0.0443
0.464	0.6772	$\frac{13}{20}$	0.0272
0.595	0.7257	$\frac{14}{20}$	0.0257
0.881	0.8106	$\frac{15}{20}$	0.0606
0.906	0.8186	$\frac{16}{20}$	0.136
1.046	0.8531	$\frac{17}{20}$	0.0031
1.237	0.8925	$\frac{18}{20}$	0.0075
1.678	0.9535	$\frac{19}{20}$	0.0035
2.455	0.9931	$\frac{20}{20}$	0.0069

Il valore osservato della statistica test si ottiene moltiplicando il massimo tra i valori tabulati nell'ultima colonna per $\sqrt{20}$. Il valore della statistica test è $\sqrt{n}D_n = 4.4721 \times 0.1498 = 0.67$ mentre $h_{0.95} = 1.36$, pertanto, essendo $0.67 < 1.36$ non si rifiuta l'ipotesi nulla.

Controllare se un test di adattamento del chi quadro di ampiezza 0.05 porta alla stessa decisione.

Esempio 3.58. *([8] Esercizio 25 pag 491)*
Dei dati si dicono lognormali di parametri μ e σ se i loro logaritmi naturali hanno distribuzione $N(\mu, \sigma^2)$. I valori seguenti rappresentano i giorni di vita di un campione di topi affetti da cancro e curati con una terapia sperimentale:

24 12 36 40 16 10 12 30 38 14 22 18

Verificare tramite un test di ampiezza $\alpha = 0,05$ i dati in esame possano provenire da una popolazione lognormale di parametri $\mu = 3$ e $\sigma = 4$.

Ordiniamo i dati e calcoliamo i loro logaritmi naturali

X_i	10	12	12	14	16	18	22	24	30	36	38	40
$\ln X_i$	2.30	2.48	2.48	2.64	2.77	2.89	3.09	3.18	3.40	3.58	3.64	3.69

Utilizziamo il test di Kolmogorov e Smirnov per verificare $\ln X \sim N(3, 16)$.

$\ln x_i$	$F_{12}(\ln x_i)$	$F_0(\ln x_i)$	$\lvert F_{12}(\ln x_i) - F_0(\ln x_i)\rvert \vee \lvert F_{12}(\ln x_{i-1}) - F_0(\ln x_i)\rvert$
2.30	$\frac{1}{12}$	0.36	0.36
2.48	$\frac{3}{12}$	0.40	0.32
2.64	$\frac{4}{12}$	0.43	0.18
2.77	$\frac{5}{12}$	0.46	0.13
2.89	$\frac{6}{12}$	0.5	0.043
3.09	$\frac{7}{12}$	0.52	0.06
3.18	$\frac{8}{12}$	0.54	0.13
3.40	$\frac{9}{12}$	0.58	0.17
3.58	$\frac{10}{12}$	0.61	0.22
3.64	$\frac{11}{12}$	0.63	0.29
3.69	1	0.64	0.36

Il valore osservato della statistica test si ottiene moltiplicando il massimo tra i valori tabulati nell'ultima colonna per $\sqrt{12}$. Il valore della statistica test è $\sqrt{n}D_n = 3.464 \times 0.36 = 1.247$ mentre $h_{0.95} = 1.36$, pertanto, essendo $1.247 < 1.36$ non si rifiuta l'ipotesi nulla.

Capitolo 4

Analisi della varianza

Introduciamo brevemente la tipologia di problemi che questa tecnica ci permette di affrontare.
Supponiamo di voler valutare l'incidenza di alcuni fattori sull'esito di un esperimento aleatorio.
Ad esempio potremmo essere interessati a valutare il rendimento medio scolastico degli studenti della scuola primaria, a seconda del libro (fattore 1), dell'insegnante (fattore 2) e della regione di appartenenza (fattore 3). L'ipotesi che si vuole testare è che i diversi fattori siano irrilevanti ai fini dell'apprendimento. A tale scopo si considerano dei campioni selezionati nel seguente modo:

$$Y_{jhki} = \mu + \tau_j + \iota_h + \delta_k + \gamma_{jk} + \gamma_{jh} + \gamma_{hk} + \gamma_{jhk} + e_{jhki}. \tag{4.1}$$

Il termine Y_{jhki} , $i = 1, .., n_{j,k,h}$ rappresenta l'i-esima osservazione della quantità in esame, quando si utilizzi il libro j (*primo fattore somministrato al livello* j) ci sia l'insegnante h (*secondo fattore somministrato al livello* h) e la scuola stia nella regione k (*terzo fattore somministrato al livello* k).
Si assume che Y_{jhki} sia determinata da una media generale μ, più un termine specifico per ognuno dei fattori, più i termini dovuti all'interazione dei fattori, più un residuo $e_{jhki} \sim N(0, \sigma^2)$, essendo i diversi residui indipendenti.

In tal modo l'ipotesi da verificare è che i contributi dovuti alla presenza di ciascun fattore siano nulli.
Nei prossimi paragrafi vedremo come si possa costruire un test basato sull'analisi della varianza nel caso di un fattore e due fattori senza interazioni (cioé tale che i contributi di tipo γ nella 4.1 siano nulli). Tratteremo il secondo caso in modo non rigoroso, dovendo altrimenti introdurre elementi di probabilità non adeguati al taglio di questo corso. Rimandiamo gli studenti interessati a [7] oppure [3] per una trattazione più completa.

4.1 Analisi della varianza ad un fattore

Supponiamo di dover acquistare un software di calcolo per l'implementazione di programmi numerici. A tale scopo è possibile scegliere tra k diversi prodotti che il venditore asserisce essere equivalenti, nel senso che il tempo medio impiegato ad eseguire un programma numerico è lo stesso per tutti. Per testare questa eventualità si esegue n_1 volte un

programma numerico con il primo software, n_2 volte lo stesso programma con il secondo software e così via. Si prende poi nota dei tempi impiegati a completare l'esecuzione del programma e si raccolgono in una tabella. Attraverso l'esame dei dati osservati, si vuole costruire un test per verificare l'ipotesi che i tempi medi di esecuzione relativi a ciascun software siano uguali. Questo test viene costruito andando a confrontare due stimatori della varianza (supposta incognita, ma uguale per tutti i campioni) delle variabili aleatorie in esame. Un primo stimatore, che stima sempre correttamente la varianza indipendentemente dalle ipotesi formulate, ed un secondo stimatore che stima bene la varianza solo nel caso in cui l'ipotesi di uguaglianza dei valori medi sia verificata, altrimenti produce una sovrastima. Evidentemente il rapporto tra il secondo ed il primo stimatore tende ad avvalorare l'ipotesi nulla qualora sia "piccolo", tende a confutarla altrimenti.

Il problema generale che si vuole affrontare è il seguente:
siano dati k campioni ognuno di cardinalità $n_i,\ i = 1, ..., k$, a ciascuno dei quali è stato somministrato l'unico fattore in esame al livello i, $i = 1, ..., k$.
Abbiamo dunque $n = \sum_{i=1}^{k} n_i$ osservazioni

$$X_{i,j} = \mu + \alpha_i + e_{i,j}.$$

con $e_{i,j} \sim N(0, \sigma^2), i = 1, ..., k, \quad j = 1, ..., n_i$.

In definitiva posto $\mu_i = \mu + \alpha_i$, si osservano k campioni estratti da k popolazioni normali $N(\mu_i, \sigma^2)$, tutte con stessa varianza incognita σ^2. Si vuole costruire un test per la verifica dell'ipotesi[1]

$$H_0 : \mu_1 = ... = \mu_k$$

in alternativa a

$$H_1 : \mu_i \neq \mu_j \text{ per qualche } i \neq j$$

Raccogliamo i k campioni in una tabella

$$\begin{array}{l} X_{11}, X_{12}, ..., X_{1,n_1} \\ X_{21}, X_{22}, ..., X_{2,n_2} \\ ..... \\ ..... \\ X_{k1}, X_{k2}, ..., X_{k,n_k} \end{array}$$

e definiamo le seguenti statistiche:

1 $\bar{X}_{i\ .} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$ *media campionaria relativa al campione i-simo*, $i = 1, ..., k$;

2 $\bar{X}_{.\ .} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n} \sum_{i=1}^{k} n_i X_{i\ .}$ *media campionaria generale*;

[1] In molti testi tali ipotesi sono formulate nel seguente modo

$$H_0 : \alpha_1 = ... = \alpha_k = 0$$

in alternativa a

$$H_1 : \alpha_i \neq 0 \text{ per qualche } i$$

3 $Dev(T) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{\cdot\,\cdot})^2$ *devianza totale.*

L'analisi della varianza si fonda principalmente sulla seguente scomposizione della devianza totale

$$\begin{aligned} Dev(T) &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{\cdot\,\cdot})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\,\cdot} + \bar{X}_{i\,\cdot} - \bar{X}_{\cdot\,\cdot})^2 = \\ &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\,\cdot})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{X}_{i\,\cdot} - \bar{X}_{\cdot\,\cdot})^2 + 2 \sum_{i=1}^{k} (\bar{X}_{i\,\cdot} - \bar{X}_{\cdot\,\cdot}) \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\,\cdot}) = \\ &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\,\cdot})^2 + \sum_{i=1}^{k} n_i (\bar{X}_{i\,\cdot} - \bar{X}_{\cdot\,\cdot})^2 \end{aligned}$$

essendo banalmente $\sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\,\cdot}) = 0, \quad i = 1, ..., k.$

Poniamo

- $Dev(B) = \sum_{i=1}^{k} n_i (\bar{X}_{i\,\cdot} - \bar{X}_{\cdot\,\cdot})^2$ la *devianza tra i campioni*
- $Dev(W) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\,\cdot})^2$ la *devianza entro i campioni*

per modo tale che

$$Dev(T) = Dev(B) + Dev(W) \tag{4.2}$$

Osserviamo che la devianza tra i campioni quantifica la variabilità derivante dalle differenza tra le medie che caratterizzano le k distribuzioni in esame. Infatti, per la legge dei grandi numeri, se i campioni sono sufficientemente numerosi, $\bar{X}_{i\,\cdot} \approx \mu_i, \ i = 1, ..., k.$ D'altra parte, se l'ipotesi nulla è vera, $\bar{X}_{i\,\cdot} \approx \mu_i = \mu, \ i = 1, ..., k$ ed anche $\bar{X}_{\cdot\,\cdot} \approx \mu.$ Di conseguenza $Dev(B)$ tende ad essere piccola se l'ipotesi nulla è vera, grande altrimenti. D'altra parte, detta $S_i^2 = \frac{\sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\,\cdot})^2}{n_i - 1}$ la varianza campionaria dell'i-simo campione, si può riscrivere

$$Dev(W) = \sum_{i=1}^{k} (n_i - 1) S_i^2 \tag{4.3}$$

$Dev(W)$ quindi è funzione dei dati osservati tramite le varianze campionarie che stimano sempre correttamente lo stesso parametro σ^2.
Possiamo quindi costruire un test basato sul rapporto $\frac{Dev(B)}{Dev(W)}$ o su una sua funzione, per testare le ipotesi in discussione.
A questo scopo premettiamo il seguente

Lemma 4.1. *Le variabili aleatorie $Dev(B)$ e $Dev(W)$ sono indipendenti e*

$$\frac{Dev(W)}{\sigma^2} \sim \chi^2_{n-k}.$$

D'altra parte, se l'ipotesi H_0 è vera, $\dfrac{Dev(B)}{\sigma^2} \sim \chi^2_{k-1}.$

Dimostrazione. Notiamo che $Dev(B)$ è una funzione deterministica delle medie campionarie $\bar{X}_{1\ .}, ..., \bar{X}_{k\ .}$, mentre la 4.3 indica che $Dev(W)$ è funzione delle varianze campionarie. L'indipendenza delle variabili aleatorie $Dev(B)$ e $Dev(W)$ consegue quindi dal fatto che in un campione aleatorio normale la media campionaria e la varianza campionaria sono indipendenti.
Inoltre, poiché $X_{i,j} \sim N(\mu_i, \sigma^2)\ j = 1, ..., n_i$, allora la 1.1 garantisce che

$$\frac{\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i\ .})^2}{\sigma^2} \sim \chi^2_{n_i - 1}$$

Di conseguenza dal Teorema 1.2 segue che

$$\frac{Dev(W)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i\ .})^2 \sim \chi^2_{n-k}$$

Supponiamo ora vera l'ipotesi H_0. In tal caso possiamo considerare tutte le osservazioni come determinazioni della stessa variabile aleatoria $N(\mu, \sigma^2)$. Di conseguenza

$$\frac{Dev(T)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{\cdot\ .})^2 \sim \chi^2_{n-1}$$

La decomposizione 4.2 e l'indipendenza tra $Dev(B)$ e $Dev(W)$ permette quindi di concludere che

$$\frac{Dev(B)}{\sigma^2} \sim \chi^2_{n-1-(n-k)} = \chi^2_{k-1}$$

□

Il risultato appena descritto permette di derivare una statistica test utile al nostro scopo:

$$F = \frac{\frac{Dev(B)}{k-1}}{\frac{Dev(W)}{n-k}} \tag{4.4}$$

Infatti il Lemma 4.1 garantisce che, se H_0 è vera, allora $F \sim$ F di Fischer con $k-1$ e $n-k$ gradi di libertà.

Il test di ampiezza α costruito con questo procedimento sarà quindi equivalente alla scelta della regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\frac{Dev(B)}{k-1}}{\frac{Dev(W)}{n-k}} > F_{1-\alpha,k-1,n-k} \right\} \tag{4.5}$$

Avendo indicato, come di consueto, con $F_{1-\alpha,k-1,n-k}$ il quantile di ordine $1-\alpha$ della legge di Fischer con $k-1$ e $n-k$ gradi di libertà.

Esempio 4.2. *([7], Esempio 1 pag. 515)*
Si vogliono testare i tempi di vita di tre diverse marche di batterie. A tale scopo, indicate con X, Y, Z i tempi di vita della prima, seconda, terza marca in esame, si procede

all'osservazione di una campione di 5 elementi dalla X, di 4 elementi dalla Y ed infine, di 6 elementi dalla Z. Le osservazioni sono riportate nella seguente tabella

X	Y	Z
40	60	60
30	40	50
50	55	70
50	65	65
30		75
		40

Assumendo che X, Y, Z siano normali con stessa varianza σ^2, costruire un test di ampiezza 0.05 per verificare se i tempi medi di vita delle tre marche in esame siano uguali.

La soluzione di questo esercizio si riduce al calcolo del rapporto in 4.9, avendo osservato che $\frac{Dev(B)}{\sigma^2} \overset{H_0}{\sim} \chi^2_2$, $\frac{Dev(W)}{\sigma^2} \sim \chi^2_{12}$.
Con i dati in esame si ottiene:

$$\overline{X}_1 \,.\, = 40, \quad \overline{X}_2 \,.\, = 55, \quad \overline{X}_3 \,.\, = 60 \quad \overline{X}.\,. = 52.$$

Inoltre

$$Dev(B) = \sum_{i=1}^{k} n_i (\bar{X}_i \,.\, - \bar{X}.\,.)^2 = 5 \times 12^2 + 4 \times 3^2 + 6 \times 8^2 = 1140$$

$$Dev(W) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i \,.\,)^2 = 1600$$

Infine

$$\frac{\frac{Dev(B)}{2}}{\frac{Dev(W)}{12}} \overset{H_0}{\sim} F_{2,12}, \quad \frac{\frac{Dev(B)}{2}}{\frac{Dev(W)}{12}} = 4.275.$$

Poiché $F_{2,12,0.95} = 3.83$ rifiuto H_0.

Un pacchetto statistico produce un output del seguente tipo per un problema del genere:
INSERIRE OUTPUT STATISTICO

Esempio 4.3. *([7], Esempio 2 pag. 516)*
Un corso di statistica elementare è suddiviso in tre parti, ciascuna insegnata da un diverso docente. Alla fine del corso gli studenti sostengono una prova per ciascuna parte ottenendo

i seguenti risultati

I parte	*II parte*	*III parte*
95	*88*	*68*
33	*78*	*79*
48	*91*	*91*
76	*51*	*71*
89	*85*	*87*
82	*77*	*68*
60	*31*	*79*
77	*62*	*16*
	96	*35*
	81	

Assumendo che i risultati per ciascuna prova siano normali con stessa varianza σ^2, costruire un test di ampiezza 0.1 per verificare se i risultati medi sono uguali.

4.2 Analisi della varianza a due fattori senza interazioni

Supponiamo ora che due fattori possano influenzare l'esito di un esperimento aleatorio. Come esempio, potremmo pensare alla quantità di grano prodotta per metro quadro da un terreno, quando si usino specie differenti di grano (fattore 1) e tipi diversi di fertilizzante (fattore 2).
Analizziamo il caso più semplice, quello cioé in cui non siano presenti interazioni tra fattori ed in cui sia disponibile una sola osservazione ad ogni livello di ciascuno dei due fattori.
Più precisamente, riferendoci all'esempio precedente, supponiamo che siano disponibili k differenti specie di grano e h tipi di fertilizzante. Assumiamo di avere a disposizione hk osservazioni

$$X_{i,j} = \mu + \alpha_i + \beta_j + e_{i,j}.$$

con $e_{i,j} \sim N(0, \sigma^2), i = 1, ..., h, \quad j = 1, ..., k.$

In definitiva posto $\mu_{ij} = \mu + \alpha_i + \beta_j$, si osservano hk variabili $X_{i,j} \sim N(\mu_{ij}, \sigma^2)$, tutte con stessa varianza incognita σ^2.
Notiamo che non è restrittivo[2]
assumere

$$\sum_{i=1}^{h} \alpha_i = 0, \quad \sum_{j=1}^{k} \beta_j = 0 \tag{4.6}$$

[2]Altrimenti, basta ridefinire i parametri come segue

$$X_{i,j} = \mu' + \alpha'_i + \beta'_j + e_{i,j}$$

dove $\mu' = \mu + \bar{\alpha} + \bar{\beta}$, $\alpha'_i = \alpha_i - \bar{\alpha}$, $\beta'_j = \beta_j - \bar{\beta}$.

In questa situazione si vogliono costruire test per la verifica delle seguenti coppie di ipotesi

Incidenza del fattore I

$$H_0^I : \alpha_1 = ... = \alpha_h = 0$$

in alternativa a

$$H_1^I : \alpha_i \neq 0 \text{ per qualche } i$$

Incidenza del fattore II

$$H_0^{II} : \beta_1 = ... = \beta_k = 0$$

in alternativa a

$$H_1^{II} : \beta_j \neq 0 \text{ per qualche } j$$

Raccogliamo le hk osservazioni in una tabella

$$\begin{array}{c} X_{11}, X_{12}, ..., X_{1,k} \\ X_{21}, X_{22}, ..., X_{2,k} \\ ..... \\ ..... \\ X_{h1}, X_{h2}, ..., X_{h,k} \end{array}$$

e definiamo, in analogia a quanto fatto nel precedente paragrafo, le seguenti statistiche:

1 $\bar{X}_{i\ .} = \frac{1}{k}\sum_{j=1}^{k} X_{i,j}$ *media campionaria relativa all'i-sima riga*, $i = 1, ..., h$;

2 $\bar{X}_{.\ j} = \frac{1}{h}\sum_{i=1}^{h} X_{i,j}$ *media campionaria relativa alla j-sima colonna*, $j = 1, ..., k$;

3 $\bar{X}_{.\ .} = \frac{1}{hk}\sum_{i=1}^{h}\sum_{j=1}^{k} X_{i,j} = \frac{1}{k}\sum_{j=1}^{k} X_{.\ j} = \frac{1}{h}\sum_{i=1}^{h} X_{i\ .}$ *media campionaria generale*;

4 $Dev(T) = \sum_{i=1}^{h}\sum_{j=1}^{k}(X_{i,j} - \bar{X}_{.\ .})^2$ *devianza totale.*

Osserviamo che

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(X_{ij} - \mu - \alpha_i - \beta_j)^2}{\sigma^2} \sim \chi^2_{hk} \tag{4.7}$$

Questa espressione dipende da $h + k + 1$ parametri incogniti, due dei quali in realtà possono essere determinati grazie alla 4.6. In definitiva l'espressione precedente dipende da $h + k - 1$ parametri incogniti linearmente indipendenti.

Come conseguenza di un noto Teorema di probabilità [3] si ottiene che sostituendo nella 4.7 ai parametri incogniti le rispettive stime di massima verosimiglianza, la variabile aleatoria risultante è ancora chi-quadrato con gradi di libertà che vanno diminuiti del numero dei parametri stimati.

Si può inoltre far vedere che le stime di massima verosimiglianza (non distorte) dei parametri nel caso in esame sono

[3] Il teorema di Cochran sulla proiezione di variabili multivariate gaussiane su sottospazi ortogonali di R^n (per approfondimenti si rimanda a [3])

$$\widehat{\mu} = \bar{X}_{\cdot\ \cdot}$$
$$\widehat{\alpha_i} = \bar{X}_{i\ \cdot} - \bar{X}_{\cdot\ \cdot}, \quad i = 1, ..., h-1$$
$$\widehat{\beta_j} = \bar{X}_{\cdot\ j} - \bar{X}_{\cdot\ \cdot}, \quad j = 1, ..., k-1$$

Sostituendo tali stime nella 4.7 si ottiene

$$\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(X_{ij} + \bar{X}_{\cdot\ \cdot} - \bar{X}_{i\ \cdot} - \bar{X}_{\cdot\ j})^2}{\sigma^2} \sim \chi^2_{(hk)-h-k+1} = \chi^2_{(h-1)(k-1)} \tag{4.8}$$

Indichiamo con

$$SS_e = \sum_{i=1}^{h}\sum_{j=1}^{k} (X_{ij} + \bar{X}_{\cdot\ \cdot} - \bar{X}_{i\ \cdot} - \bar{X}_{\cdot\ j})^2$$

le *somme dei quadrati totale*.

Osserviamo che la legge di $\frac{SS_e}{\sigma^2}$ non dipende dalle ipotesi in discussione e che inoltre $E\left(\frac{SS_e}{\sigma^2}\right) = (h-1)(k-1)$. Quindi, per linearità, $E\left(\frac{SS_e}{(h-1)(k-1)}\right) = \sigma^2$, ovvero, indipendentemente dalle ipotesi in discussione, $\frac{SS_e}{(h-1)(k-1)}$ è uno *stimatore non distorto per la varianza comune* σ^2.

Prendiamo ora in considerazione il primo gruppo di ipotesi

$$H_0^I : \alpha_1 = ... = \alpha_h = 0$$

in alternativa a

$$H_1^I : \alpha_i \neq 0 \text{ per qualche } i$$

e consideriamo la variabile aleatoria $\bar{X}_{i\ \cdot}$. Evidentemente $\bar{X}_{i\ \cdot} \sim N(\mu + \alpha_i, \frac{\sigma^2}{k})$ e, *se l'ipotesi* H_0^I *si suppone vera*, $\bar{X}_{i\ \cdot} \underset{H_0^I}{\sim} N(\mu, \frac{\sigma^2}{k})$. Di conseguenza

$$\sum_{i=1}^{h} k\frac{(\bar{X}_{i\ \cdot} - \mu)^2}{\sigma^2} \underset{H_0^I}{\sim} \chi^2_h$$

e, Per quanto già osservato, sostituendo a μ il suo stimatore di massima verosimiglianza, si ottiene

$$\sum_{i=1}^{h} k\frac{(\bar{X}_{i\ \cdot} - \bar{X}_{\cdot\ \cdot})^2}{\sigma^2} \underset{H_0^I}{\sim} \chi^2_{h-1}$$

Indichiamo con

$$SS_r = \sum_{i=1}^{h} k(\bar{X}_{i\ \cdot} - \bar{X}_{\cdot\ \cdot})^2$$

le *somme dei quadrati delle righe*.

Se H_0^I è vera, allora $\frac{SS_r}{h-1}$ è uno *stimatore non distorto per la varianza comune* σ^2. Vale inoltre i seguente risultato che enunciamo senza dimostrazione

Lemma 4.4. *Le variabili aleatorie SS_e e SS_r sono indipendenti e*

$$\frac{SS_e}{\sigma^2} \sim \chi^2_{(h-1)(k-1)}.$$

Inoltre, se l'ipotesi H_0^I è vera, $\dfrac{SS_r}{\sigma^2} \sim \chi^2_{h-1}$.

Le osservazioni fatte nel caso dell'analisi della varianza ad un fattore si applicano anche in questo caso; infatti, se l'ipotesi nulla è vera, $\bar{X}_{i\,.} \approx \mu_i = \mu,\ i = 1, ..., h$ ed anche $\bar{X}_{.\,.} \approx \mu$. Pertanto la legge dei grandi numeri garantisce che SS_r tende ad essere piccola.

Possiamo quindi utilizzare come statistica test il rapporto:

$$F = \frac{\frac{SS_r}{h-1}}{\frac{SS_e}{(h-1)(k-1)}} \tag{4.9}$$

Infatti dal Lemma 4.4 si evince che, se H_0^I è vera, allora $F \sim$ F di Fischer con $h-1$ e $(h-1)(k-1)$ gradi di libertà.

Il test di ampiezza α costruito con questo procedimento sarà quindi equivalente alla scelta della regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\frac{SS_r}{h-1}}{\frac{SS_e}{(h-1)(k-1)}} > F_{1-\alpha,h-1,(h-1)(k-1)} \right\} \tag{4.10}$$

con il consueto significato dei simboli.

Infine, indicato con

$$SS_c = \sum_{j=1}^{k} h(\bar{X}_{.\,j} - \bar{X}_{.\,.})^2$$

le *somme dei quadrati delle colonne*, utilizzando tecniche analoghe si costruisce il test di ampiezza α per la verifica delle ipotesi

$$H_0^{II} : \beta_1 = ... = \beta_k = 0$$

in alternativa a

$$H_1^{II} : \beta_j \neq 0 \text{ per qualche } j$$

che corrisponde alla scelta della regione critica

$$C = \left\{ (x_1, ..., x_n) \in \mathbf{X_{(n)}} \text{ tali che } \frac{\frac{SS_c}{k-1}}{\frac{SS_e}{(h-1)(k-1)}} > F_{1-\alpha,k-1,(h-1)(k-1)} \right\}. \tag{4.11}$$

Esempio 4.5. *([7], Esempio 1 pag. 522)*
la seguente tabella raccoglie i dati relativi alle quantità di grano espresse in chilogrammi per metro quadro prodotti da 3 diverse varietà di grano trattate con 4 differenti tipi di fertilizzante

	Varietà di grano		
Fertilizzante	*A*	*B*	*C*
α	*8*	*3*	*7*
β	*10*	*4*	*8*
γ	*6*	*5*	*6*
δ	*8*	*4*	*7*

Assumendo le ipotesi di questo paragrafo valide, costruire un test per la verifica dell'ipotesi che le diverse qualità di grano non influenzino la produzione media e un test per la verifica dell'ipotesi che i diversi fertilizzanti non influenzino la produzione media.

Siamo nella situazione descritta nel presente paragrafo. Abbiamo il fattore I (tipo di fertilizzante) somministrato a 4 livelli ed il fattore II (tipo di grano) somministrato a 3 livelli. Le statistiche di interesse sono:

1 $\bar{X}_{1\ .} = 6$, $\bar{X}_{2\ .} = 7.33$, $\bar{X}_{3\ .} = 5.67$, $\bar{X}_{4\ .} = 6.33$;

2 $\bar{X}_{.\ 1} = 8$, $\bar{X}_{.\ 2} = 4$, $\bar{X}_{.\ 3} = 7$;

3 $\bar{X}_{.\ .} = 6.33$.

Calcoliamo ora la somma dei quadrati totale, la somma dei quadrati delle righe e la somma dei quadrati delle colonne.

1 $SS_e = \sum_{i=1}^{4} \sum_{j=1}^{3} (X_{ij} + \bar{X}_{.\ .} - \bar{X}_{i\ .} - \bar{X}_{.\ j})^2 = 7.33$

2 $SS_r = \sum_{i=1}^{4} 3(\bar{X}_{i\ .} - \bar{X}_{.\ .})^2 = 3(0.33^2 + 1 + 0.66^2 + 0) = 4.67$

3 $SS_c = \sum_{i=1}^{3} 4(\bar{X}_{.\ j} - \bar{X}_{.\ .})^2 = 4(1.67^2 + 2.33 + 0.67^2) = 34.67$

Facciamo un test di ampiezza $\alpha = 0.05$ per valutare l'incidenza del fattore I sulla produzione media di grano. La regione critica del test è

$$C = \left\{ \frac{\frac{SS_r}{3}}{\frac{SS_e}{6}} > F_{1-\alpha,3,6} \right\}$$

La statistica test valutata sulle osservazioni vale $F = \frac{\frac{4.67}{3}}{\frac{7.33}{6}} = 1.27$; inoltre:
$F_{3,6;0.95} = 4.76$ e dunque il test non rifiuta l'ipotesi nulla, ovvero si ritiene che il fertilizzante usato non incida sulla produzione media di grano.

Valutiamo ora l'incidenza del fattore II sulla produzione media di grano con un test di ampiezza $\alpha = 0.01$. La regione critica del test è

$$C = \left\{ \frac{\frac{SS_c}{2}}{\frac{SS_e}{6}} > F_{1-\alpha,2,6} \right\}$$

La statistica test valutata sulle osservazioni vale $F = \frac{\frac{34.67}{2}}{\frac{7.33}{6}} = 14.2$; inoltre:
$F_{2,6;0.99} = 13.74$ e dunque il test rifiuta l'ipotesi nulla, ovvero si ritiene che il tipo di grano usato influenzi la produzione media di grano.

Esempio 4.6. *([7], Esercizio 3 pag. 523)*
I seguenti dati rappresentano il numero di pezzi giornalieri prodotti da 4 diverse macchinari utilizzate da 4 operai

	Operai			
Macchinari	*A*	*B*	*C*	*E*
α	*15*	*14*	*19*	*18*
β	*17*	*12*	*20*	*16*
γ	*16*	*18*	*16*	*17*
δ	*16*	*16*	*15*	*15*

Assumendo le ipotesi di questo paragrafo valide, costruire dei test di ampiezza 0.05 *per verificare l'influenza nella produzione dovuta ai macchinari e agli operai.*

Bibliografia

[1] Andreatta, G., and Runggaldier, W. *Statistica Matematica - Problemi ed esercizi risolti*. Liguori Editore, 1990.

[2] Baldi, P. *Calcolo delle Probabilità*. McGraw-Hill, 2007.

[3] Casella, G., and Berger, R. L. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[4] Hoel, P. G., Port, S. C., and Stone, C. J. *Introduction to statistical theory*. Houghton Mifflin Co., Boston, Mass., 1971. The Houghton Mifflin Series in Statistics.

[5] Mood, A., Graybill, F. A., and Boes, D. C. *Introduzione alla Statistica*. McGraw-Hill, 1988.

[6] Piccolo, D. *Statistica*. Il Mulino, 2000.

[7] Rohatgi, V. K. *An introduction to probability theory and mathematical statistics*. Wiley-Interscience [John Wiley & Sons], New York, 1976. Wiley Series in Probability and Mathematical Statistics.

[8] Ross, S. M. *Probabilità e statistica per l'ingegneria e le scienze*. Apogeo, 2008.