Appunti del corso di Statistica Matematica

Gadotti Andrea, Nardin Michele, Peruzzetto Marco

Indice

Capitolo 1

Prima parte del corso

1.1 Introduzione

In questa prima sezione vengono presentati i richiami di teoria della probabilità, affrontati nelle primissime lezioni del corso.

1.1.1 Funzione generatrice dei momenti

Lezione del 18/02, ultima modifica 09/04, Andrea Gadotti

Definizione 1. *Sia X una variabile casuale (discreta o assolutamente continua). Se esiste $t_0 > 0$ tale per cui $\mathbb{E}(e^{tX}) < +\infty \ \forall t \in (-t_0, t_0)$, chiameremo la funzione*

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

funzione generatrice dei Momenti di X.

Esempi

1. $X \sim b(1,p)$ con $p \in (0,1)$. Si ha:

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x=0}^{1} e^{tx}\mathbb{P}(X=x)$$

$$= \sum_{x=0}^{1} e^{tx}p^x(1-p)^{1-x} = pe^t + (1-p)$$

2. $X \sim \mathcal{P}(\lambda)$ con $\lambda > 0$. Si ha:

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x=0}^{+\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{\lambda(e^t-1)}$$

3. $X \sim \mathcal{G}(\alpha,\beta)$, ovvero

$$f_X(x;\alpha,\beta) := \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{1}{\beta}x}$$

con $\alpha > 0$, $\beta > 0$, $x > 0$ e

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

(nota: $\alpha \in \mathbb{N} \Longrightarrow \Gamma(\alpha) = (\alpha - 1)!$)

Abbiamo che:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_0^{+\infty} e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{1}{\beta}x} \mathrm{d}x$$

$$= ...[\text{sostituzione } \sigma := x(\tfrac{1}{\beta} - t)]...$$

$$= \frac{1}{(1-\beta t)^\alpha}$$

con $t < \frac{1}{\beta}$

Momenti di una variabile casuale

Definizione 2. *Se una variabile casuale ammette FGM derivabile infinite volte in un intorno di $t = 0$ e se tutti i suoi momenti sono finiti, allora definiamo il momento di ordine s non centrato:*

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

Il momento di ordine s centrato in $a \in \mathbb{R}$ è:

$$\mu_s(a) := \mathbb{E}((X-a)^s)$$

Ovvero $\mu'_s = \mu_s(0)$. E' chiaro che $\mu'_1 = \mathbb{E}(X)$. Chiameremo infine momento di ordine s centrato (senza specificare altro, intenderemo centrato in μ'_1):

$$\mu_s := \mathbb{E}((X - \mu'_1)^s)$$

Teorema 1. *Dall'ultima definizione si vede facilmente (sviluppando l'elevamento a potenza) che vale la seguente relazione tra momenti centrati e non:*

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

Osserviamo che $\mu_2 = \mathbb{E}((X-\mu'_1)^2) = Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mu'_2 - (\mu'_1)^2$

Teorema 2. *Date due (o più) v.c. X e Y indipendenti aventi f densità / f massa f_X e f_Y e fgm $M_X(t)$ e $M_Y(t)$ rispettivamente, allora si ha*

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

Teorema 3. *Siano X e Y v.c. con funzioni di ripartizione $F_X(x)$ e $F_Y(y)$ rispettivamente. Siano $M_X(t)$ e $M_Y(t)$ le fgm di X e Y. Se $M_X(t) = M_Y(t)$ per ogni t in un intorno dell'origine, allora*

$$X \stackrel{d}{=} Y$$

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

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

$$M_{S_n}(t) = \mathbb{E}\left(e^{tS_n}\right) = \mathbb{E}\left(e^{t\sum_{i=1}^{n} X_i}\right)$$

$$\stackrel{TEO2}{=} \prod_{i=1}^{n} \mathbb{E}(e^{tX_i}) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} (pe^t + (1-p)) = (pe^t + (1-p))^n$$

ovvero S_n è distribuita come $b(n,p)$ per il Teorema 2.
Esercizio Ripetere il calcolo precedente supponendo $X_i \sim P(\lambda), \forall i$.

1.1.2 Famiglia Esponenziale a k parametri

Una famiglia di f densità / f massa è detta essere una Famiglia Esponenziale a k parametri $\theta_1, ..., \theta_k$ se la corrispondente f densità / f massa (che è indicizzata da $\theta_1, ..., \theta_k$) può essere scritta come

$$f_X(x;\theta) = C^*(x)D^*(\theta)\exp\{\sum_{m=1}^{k} A_m(\theta)B_m(x)\}$$

dove $C^*(x)$ è una funzione della sola x, $D^*(\theta)$ è una funzione del solo θ, $A_m(\theta)$ è una funzione del solo θ e $B_m(x)$ è una funzione della sola x.
Esempi

1. $X \sim G(\alpha,\beta) \Longrightarrow f_X(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{1}{\beta}x} \mathbb{1}_{\mathbb{R}^+}(x)$, $\alpha > 0$, $\beta > 0$ $\mathbb{1}_{\mathbb{R}^+}$ è detto supporto della distribuzione. Quindi possiamo riscrivere $f_X(x;\alpha,\beta)$ come

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

   e quindi ponendo $D^*(\alpha,\beta) := \frac{1}{\Gamma(\alpha)\beta^\alpha}$, $C^*(x) := \mathbb{1}_{\mathbb{R}^+}(x)$, $A_1(\alpha,\beta) := (\alpha-1)$, $B_1(x) := ln(x)$, $A_2(\alpha,\beta) := -\frac{1}{\beta}$ e $B_2(x) := x$, otteniamo $G(\alpha,\beta)$ come famiglia esponenziale con $k = 2$.

2. $X \sim b(n,p) \Longrightarrow f_X(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{1}_{\{0,1,...,n\}}(x)$ con $n \in \mathbb{N}$ noto. Quindi possiamo riscrivere $f_X(x;n,p)$ come

   $$f_X(x;n,p) = \binom{n}{x} \mathbb{1}_{\{0,1,...,n\}}(x)(1-p)^n exp(x\ln(\frac{p}{1-p}))$$

con $\frac{p}{1-p}$ detto odd ratio o parametro naturale della famiglia esponenziale.
Quindi ponendo $D^*(p) := (1-p)^n$, $C^*(x) := \binom{n}{x}\mathbb{1}_{\{0,1,...,n\}}(x)$, $A_1(p) := ln(\frac{p}{1-p})$, $B_1(x) := x$, otteniamo $b(n,p)$ come famiglia esponenziale con $k=1$.

3. X vc con $f_X(x,\vartheta) = \frac{e^{1-x/\vartheta}}{\vartheta}\mathbb{1}_{(\vartheta,\infty)}(x)$: la distribuzione di X non appartiene a famiglia esponenziale. Il fatto che il supporto di f_X dipenda dal parametro ϑ NON permette a f_X di appartenere ud una famiglia esponenziale!

Osservazione 2. *Le famiglie di esponenziali hanno interessanti proprietà matematiche (proprietà di regolarità).*

Dal punto di vista statistico, ciò si traduce in un'interessante conseguenza: tutta l'informazione contenuta nei dati a disposizione $(X_1,...,X_n)$ relativa alla funzione $f_X(x;\theta)$ può essere sintetizzata attraverso k quantità (funzioni di $(X_1,...,X_n)$) che potranno essere impiegate per costruire procedure inferenziali (stima, test per la verifica di ipotesi) riguardanti il parametro θ.
Ovvero, l'appartenenza ad una famiglia esponenziale permette una riduzione dei dati (X_1, ... ,X_n) via B_m.

1.1.3 Trasformazioni di variabili casuali

Lezione del 01/03, ultima modifica 09/04, Michele Nardin

Discrete

Teorema 4. *Sia X una vc con funzione di massa $f_X(x) = P(X=x)$, e sia A_X il suo supporto. Sia W=h(X) una nuova vc. Allora*

$$P(W=w) = \sum_{\{x\in A_X : h(x)=w\}} P(X=x)$$

Esempi

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

   Considero quindi $W = n - X$. Come si distribuisce W?

$$P(W=w) = P(X=n-w) = \binom{n}{n-w}p^{n-w}(1-p)^w\mathbb{1}_{0,1,...,n}(w)$$

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

$$P(W=w) = P(X^3=w) = P(X=\sqrt[3]{w}) = \left(\frac{1}{2}\right)^{\sqrt[3]{w}} \mathbb{1}_{1,8,27,64,...}(w)$$

Assolutamente continue

Teorema 5. *Sia X una variabile casuale (ass continua) con funzione di densità $f_X(x)$ e sia $W = h(X)$, ove h è una funzione monotona. Supponiamo inoltre che $f_X(x)$ sia continua sul supporto di X e che $h^{-1}(w)$ abbia derivata continua sul supporto di W. Allora*

$$f_W(w) = f_X(h^{-1}(w)) \left| \frac{d}{dw} h^{-1}(w) \right| \mathbb{1}_{A_W}(w)$$

Esempio (Standardizzazione di una vc normale) Sia $X \sim N(m, s^2)$. Considero $W = h(X) = \frac{X-m}{s}$. Allora, dato che $h^{-1}(w) = sw + m$, che ha derivata continua su tutto $\mathbb{R}$,

$$f_W(w) = f_X(sw + m)|s| = \frac{e^{\frac{-w^2}{2}}}{\sqrt{2\pi}} = f_{N(0,1)}$$

Teorema 6. *Se $W = h(X)$ ove h è monotona a tratti (un numero di tratti finito k) e valgono le condizioni del teorema precedente (su ogni tratto), allora*

$$f_W(w) = \sum_{n=1}^{k} f_X(h_n^{-1}(w)) \left| \frac{d}{dw} h_n^{-1}(w) \right| \mathbb{1}_{A_W}(w)$$

Esempio (Chi-quadro)
Sia $X \sim N(0,1)$ e $W = h(X) = X^2$. h è monotona sui tratti $A_0 = 0$, $A_1 = (-\infty, 0)$, $A_2 = (0, +\infty)$.
Considero $h_1(x) = x^2$ per $x < 0$ mentre $h_2(x) = x^2$ per $x > 0$.
Trovo che $h_1^{-1}(w) = -\sqrt{w}$ (NB:$h_1^{-1}(w) \in A_1 \forall w \geq 0$), mentre $h_2^{-1}(w) = \sqrt{w}$ (NB:$h_2^{-1}(w) \in A_2 \forall w \geq 0$).
$\frac{d}{dw} h_1^{-1}(w) = -\frac{1}{2\sqrt{w}}$, $\frac{d}{dw} h_2^{-1}(w) = \frac{1}{2\sqrt{w}}$ sono entrambe continue su $\mathbb{R}_+$.

$$f_W(w) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(-\sqrt{w})^2}{2}} \left| \frac{1}{2\sqrt{w}} \right| + \frac{1}{\sqrt{2\pi}} e^{\frac{-(\sqrt{w})^2}{2}} \left| \frac{1}{2\sqrt{w}} \right|$$

$$= \frac{1}{\sqrt{2\pi w}} e^{\frac{-w}{2}} \mathbb{1}_{\mathbb{R}+}(w) = \frac{1}{2^{1/2}\Gamma(1/2)} w^{\frac{1}{2}-1} e^{\frac{-w}{2}}$$

Si riconosce che $W \sim \mathcal{G}(\alpha = 1/2, \beta = 2)$ e si chiama Chi quadrato con $\nu = 1$ gradi di libertà.

In generale, una vc Chi Quadro con $\nu = n$ gradi di libertà è $W = \sum_{i=1}^{n} X_i^2$, ove $X_1, X_2, ..., X_n$ sono vc iid N(0,1). Per il Teorema 2 sulla FGM di una somma di vc iid si trova immediatamente che $W \sim \mathcal{G}(\alpha = n \cdot 1/2, \beta = 2)$.

1.1.4 Convergenze

Convergenza in probabilità

Definizione 3. *Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili casuali e sia X un'altra variabile casuale, tutte definite sullo stesso spazio campionario. Diciamo che X_n converge in probabilità a X (scriviamo $X_n \xrightarrow{p} X$) se $\forall \varepsilon > 0$*

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

Osservazione 3. *Se $X_n \xrightarrow{p} X$ diciamo che la massa della differenza $|X_n - X|$ converge a 0. Inoltre, quando scriviamo $X_n \xrightarrow{p} X$, stiamo sottintendendo tutta la parte iniziale della definizione precendete, cioè il sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili casuali....*

Teorema 7. *Alcuni risultati utili:*

1. *Supponiamo che $X_n \xrightarrow{p} X$ e $Y_n \xrightarrow{p} Y$. Allora $X_n + Y_n \xrightarrow{p} X + Y$*
2. *Supponiamo che $X_n \xrightarrow{p} X$ e sia a una costante. Allora $aX_n \xrightarrow{p} aX$*
3. *Supponiamo che $X_n \xrightarrow{p} a$ costante, e sia g una funzione reale continua in a. Allora $g(X_n) \xrightarrow{p} g(a)$*
4. *(Corollario di 3.) Se $X_n \xrightarrow{p} a$, allora $X_n^2 \xrightarrow{p} a^2$, $\frac{1}{X_n} \xrightarrow{p} \frac{1}{a}$ (se $a \neq 0$), $\sqrt{X_n} \xrightarrow{p} \sqrt{a}$ ($a \geq 0$).*
5. *$X_n \xrightarrow{p} X$ e $Y_n \xrightarrow{p} Y$ allora $X_nY_n \xrightarrow{p} XY$*

Convergenza in distribuzione

Definizione 4. *Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili casuali e sia X un'altra variabile casuale, tutte definite sullo stesso spazio campionario.*

Siano F_{X_n} e F_X le relative funzioni di ripartizione (dette anche "di distribuzione"). Sia $C(F_X)$ l'insieme dei punti ove F_X è continua. Diciamo che X_n converge in distribuzione (o in legge) a X (scriviamo $X_n \xrightarrow{d} X$) se

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

Teorema 8. *Se $X_n \xrightarrow{p} X$ allora $X_n \xrightarrow{d} X$.*

Osservazione 4. *Il contrario in generale non vale, tranne nel caso in cui X è una vc degenere (cioè costante).*

Teorema 9. *Supponiamo che $X_n \xrightarrow{d} X$ e sia g una funzione continua sul supporto di X. Allora $g(X_n) \xrightarrow{d} g(X)$*

Teorema 10 (Slutsky)**.** *Supponiamo che $X_n \xrightarrow{d} X$, $A_n \xrightarrow{p} a$ costante e $B_n \xrightarrow{p} b$ costante. Allora $A_n + B_nX_n \xrightarrow{d} a + bX$*

1.1.5 Teoria asintotica

Lezione del 04/03, ultima modifica 09/04, Michele Nardin

Teorema 11. *(Δ-method) Sia $\{X_n\}_{n\in\ \mathbb{N}}$ una successione di vc tale che $\sqrt{n}(X_n - \vartheta) \xrightarrow{d} N(0, \sigma^2)$. Supponiamo che una funzione $g(X)$ sia derivabile in ϑ e che $g'(\vartheta) \neq 0$. Allora*

$$\sqrt{n}(g(X_n) - g(\vartheta)) \xrightarrow{d} N(0, \sigma^2(g'(\vartheta))^2)$$

Esempio Considero

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

ove χ^2_n è la chiquadro con n gradi di libertà. Ricordiamo che $\mathbb{E}(\chi^2_n) = n$ e che $Var(\chi^2_n) = 2n$ (discende dal fatto che $\chi^2_n \sim \mathcal{G}(\alpha = n/2, \beta = 2)$. Affermiamo che $Y_n \xrightarrow{d} N(0,1)$. Infatti:

$$Y_n = \frac{\chi^2_n - n}{\sqrt{2n}} = \frac{\sum_{i=1}^n X_i^2 - n \cdot 1}{\sqrt{n}\sqrt{2}}$$

dove $X_i \sim N(0,1)$, e quindi $X_i^2 \sim \chi^2_1$, quindi le X_i^2 hanno media $\mu = 1$ e varianza $\sigma^2 = 2$. Quindi per il Teorema centrale del Limite (vedi sotto) si ha quanto voluto.
Scrivendo ora Y_n nella forma $Y_n = \sqrt{n}\left(\frac{\chi^2_n}{\sqrt{2}n} - \frac{1}{\sqrt{2}}\right)$ riconosciamo che la prima parte delle ipotesi del Δ-method sono soddisfatte. Considero quindi $g(t) = \sqrt{t}$, che è derivabile in $\vartheta = 1/\sqrt{2}$, $g'(t) = \frac{1}{2\sqrt{t}}|_{\vartheta=1/\sqrt{2}} = 2^{-3/4}$. Allora

$$\sqrt{n}(g\left(\frac{\chi^2_n}{\sqrt{2}n}\right) - g(\vartheta)) = \sqrt{n}\left(\sqrt{\frac{\chi^2_n}{\sqrt{2}n}} - \sqrt{\frac{1}{\sqrt{2}}}\right) \xrightarrow{d} N(0, 1^2 \cdot 2^{-3/2})$$

Teorema 12. *(Teorema centrale del limite) Siano $X_1, ...X_n$ vc iid dotate di media μ e varianza finita σ^2. Allora*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n} \cdot \sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

con $\overline{X}_n$ media aritmetica delle X_i.

Esempi/Applicazioni

1. $X_n \sim b(n,p)$, $X_n \overset{a}{\sim} N(np, np(1-p))$ (ricordiamo che $X_n \sim \sum_{i=1}^n b_i$, ove $b_i \sim b(1,p)$). Quando scriviamo $\overset{a}{\sim}$ stiamo considerando un andamento asintotico, ossia sottintendiamo un'approssimazione (via via migliore con l'aumentare di n) giustificata dal TLC (il senso è che per n 'grandi' la distribuzione 'funziona circa così').

2. $X_1, ..., X_n$ vc $P(\lambda = 1)$. Considero $Y_n = \sum X_i$. Dato che $Y_n \overset{a}{\sim} N(n\lambda, n\lambda)$ e $\lambda = 1$, $\bar{Y}_n := \frac{Y_n}{n} \overset{a}{\sim} N(1, 1/n)$

   Considero quindi $W_n = \sqrt{n}(\frac{Y_n}{n} - 1) = \frac{\frac{Y_n}{n} - 1}{1/\sqrt{n}} = \frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\sqrt{Var(\bar{Y}_n)}}$, trovo che $W_n \overset{a}{\sim} N(0,1)$

Teorema 13. *Sia $\{X_n\}$ una succ di vc iid, ognuna con con FGM $M_{X_n}(t)$ definita e $< \infty$ per $t \in (-h, h)$, e sia X un'altra vc con FGM $M_X(t)$ definita e $< \infty$ per $t \in (-h_1, h_1)$, $h_1 \leq h$. Se*

$$\lim_{n \to +\infty} M_{X_n}(t) = M_X(t) \quad \forall |t| \leq h_1$$

allora $X_n \xrightarrow{d} X$.

Applicazione

1. Sia $X_n \sim b(n,p)$. Ricordiamo che $X_n = \sum X_i$ ove $X_i \sim b(1,p)$, ed inoltre $\mu = \mathbb{E}(X) = np$. Siccome $M_{X_n}(t) = \mathbb{E}(e^{tX_n}) = [(1-p) + pe^t]^n = [1 + \frac{\mu}{n}(e^t - 1)]^n$,

$$M_{X_n}(t) \overset{n\to\infty}{\longrightarrow} e^{\mu(e^t-1)}$$

che è la FGM di una Poisson di parametro μ, ovvero $X_n \overset{d}{\to} \mathcal{P}(n,p)$.

1.2 Approccio applicativo alla Statistica Matematica

Questa sezione corrisponde alla parte di corso svolta dalla seconda settimana di marzo fino a metà aprile, che riguarda gli aspetti pratici della statistica: verranno introdotte le statistiche d'ordine, gli intervalli di confidenza e i test per verifiche d'ipotesi.

Definizione 5. *(Campione Casuale) Il vettore casuale* $(X_1, ..., X_n)$ *si dice Campione Casuale relativamente ad una vc* $X \sim F_X(x, \vartheta)$ *se i suoi elementi sono vc i.i.d.*

Osservazione Il fatto che le vc siano i.i.d. implica che

$$F_{X_1,...,X_n}(X_1, ..., X_n) = \prod_{i=1}^{n} F_{X_i}(X_i)$$

e

$$f_{X_1,...,X_n}(X_1, ..., X_n) = \prod_{i=1}^{n} f_{X_i}(X_i)$$

Definizione 6. *(Statistica) Sia* $(X_1, ..., X_n)$ *un campione casuale da una distribuzione associata alla vc* X*, e sia* Ω *lo spazio campionario di* $(X_1, ..., X_n)$
Ogni funzione

$$T(X_1, ..., X_n) : \Omega \longrightarrow \mathbb{R}^k$$

che NON dipende da parametri incogniti è detta Statistica.

1.2.1 Statistiche d'ordine

Lezioni 08 e 11 Marzo, ultima modifica 21/03, Scritte da: Marco Peruzzetto

Definizione 7. *Sia* $(X_1, \ldots, X_n)$ *un campione casuale con distribuzione* $F_X(x, \theta)$*, densità* $f_X(x, \theta)$ *e supporto* $\text{supp}\{X\} \coloneqq (a, b) \subset \mathbb{R}$ *ove* $X \in \{X_1, \ldots, X_n\}$ *e* $-\infty \leq a < b \leq +\infty$*. Definiamo ricorsivamente le seguenti variabili casuali:*

- $X_{(1)} \coloneqq \min(\{X_1, \ldots, X_n\})$*;*
- $X_{(i)} \coloneqq \min\left(\{X_1, \ldots, X_n\} \setminus \{X_{(1)}, \ldots, X_{(i-1)}\}\right)$ $\forall 1 < i \leq n$*.*

Chiameremo allora $X_{(i)}$ *la i-esima Statistica d'Ordine del campione.*

Osservazione: La statistica d'ordine consiste semplicemente nel vettore per il quale le variabili casuali vengono appunto ordinate, in base al valore che assumono in un determinato punto del loro dominio comune, in ordine crescente. In particolare $X_{(i)}$ sarà l' i-esima variabile più piccola. Naturalmente, se il campione ha lunghezza n, allora $X_{(n)} = \max(\{X_1, \ldots, X_n\})$. Osserviamo che la funzione $(X_1, \ldots, X_n) \longmapsto \left(X_{(1)}, \ldots, X_{(n)}\right)$ è essa stessa una Statistica.

Teorema 14. *Sia* $(X_1, \ldots, X_n)$ *un campione casuale come sopra. Allora si ottiene* $\forall$ $1 \leq m \leq n$ *che la densità dell' m-esima statistica d'ordine è data da:*

$$f_{X_{(m)}}(x, \theta) = \frac{n!}{(m-1)!(n-m)!} f_X(x, \theta) \cdot F_X(x, \theta)^{m-1} \cdot \left(1 - F_X(x, \theta)\right)^{n-m}$$

Daremo due dimostrazioni, la seconda più bella della prima.

Dimostrazione. (a cura di Marco Perruzzetto) Innanzi tutto si ha che il supporto (a, b) può essere partizionato in n parti, per cui evidentemente si ha:

$$f_{(X_{(1)},\dots,X_{(n)})}(x_{(1)},\dots,x_{(n)},\theta) = \begin{cases} n!\prod_{i=1}^{n} f_X(x_{(i)},\theta) & \text{se } a < x_{(1)} < x_{(2)} < \ldots < x_{(n)} < b \\ 0 & \text{altrimenti.} \end{cases}$$

ove la produttoria è giustificata dal fatto che le variabili sono tutte indipendenti e che devono essere ciascuna minore dell'altra per l'ordinamento assegnato; il coefficiente fattoriale è presente poiché le n parti dell'intervallo (a, b) possono essere assegnate alle n variabili in tale numero di modi, dato che ciascuna X_i $\forall 1 \leq i \leq n$ ha la stessa distribuzione.
Adesso per trovare la distribuzione di ciascuna $X_{(m)}$ sarà dunque sufficiente integrare $f_{(X_{(1)},\dots,X_{(n)})}$ nei domini possibili di tutte le altre funzioni di distribuzione di ciascuna $X_{(i)}$ con $i \neq m$. In particolare, ciascuna $f_{X_{(i)}}$ per $i < m$ dovrà assumere a piacere valori necessariamente inferiori a $f_{X_{(m)}}$, viceversa ogni $f_{X_{(i)}}$ per $i > m$ dovrà assumere valori obbligatoriamente superiori a quelli di $f_{X_{(m)}}$ in ogni punto. Ricordando allora che possiamo scrivere la distribuzione come $\int_a^x f_X(\theta, t)dt = F_X(\theta, x)$ essendo la densità la derivata della funzione di distribuzione, otterremo quindi che $\forall a < x_{(m)} < b$ la distribuzione sarà data da:

$$\begin{aligned} f_{X(m)}(x_{(m)},\theta) = & \\ &= \int_a^{x_{(2)}} dx_{(1)} \cdots \int_a^{x_{(m)}} dx_{(m-1)} \int_{x_{(m)}}^{b} dx_{(m+1)} \cdots \int_{x_{(n-1)}}^{b} dx_{(n)} f_{(X_{(1)},\dots,X_{(n)})}(x_{(1)},\dots,x_{(n)},\theta) \\ &= \int_a^{x_{(2)}} dx_{(1)} \cdots \int_a^{x_{(m)}} dx_{(m-1)} \int_{x_{(m)}}^{b} dx_{(m+1)} \cdots \int_{x_{(n-1)}}^{b} dx_{(n)} n! \prod_{i=1}^{n} f_X(x_{(i)},\theta) \\ &= f_X(x_{(m)}) \int_a^{x_{(2)}} dx_{(1)} \cdots \int_a^{x_{(m)}} dx_{(m-1)} \int_{x_{(m)}}^{b} dx_{(m+1)} \cdots \int_{x_{(n-1)}}^{b} dx_{(n)} n! \prod_{i=1, i\neq m}^{n} f_X(x_{(i)},\theta) \\ &= \frac{n!}{(m-1)!(n-m)!} f_X(x,\theta) \cdot F_X(x,\theta)^{m-1} \cdot \big(1 - F_X(x,\theta)\big)^{n-m}, \end{aligned}$$

dove è stato usato il fatto che $\int_a^b F_X^{\alpha}(\theta, t) f_X(\theta, t)dt = \frac{F_X^{\alpha+1}}{\alpha+1}$, $\forall \alpha \neq -1$. □

Dimostrazione. Sia Ω il dominio comune del campione casuale. Definiamo per $x \in \mathbb{R}$ la nuova variabile casuale Y_x come:

$$\begin{aligned} \Omega &\longrightarrow \{0,\dots,n\} \\ Y_x(\omega) \coloneqq \sum_{i=1}^{n} \mathbb{1}_{\{X_i(\omega)\leq x\}}(\omega) &= \#\big\{i \in \{1,\dots,n\} : X_i \leq x\big\}, \end{aligned}$$

funzione che, per così dire, "conta" il numero di variabili casuali X_i che non superano x. Si vede immediatamente che $\forall 1 \leq m \leq n$, si ha la distribuzione

$$\begin{aligned} F_{X_{(m)}}(\theta, x) &= \mathbb{P}[X_{(m)} \leq x] = \mathbb{P}[\text{almeno } X_{(1)}, ..., X_{(m)} \text{ stanno sotto } x] = \\ &= \mathbb{P}[Y_x \geq m] = \sum_{k=m}^{n} \mathbb{P}[Y_x = k] = \\ &= \sum_{k=m}^{n} \binom{n}{k} F_X^k(\theta, x)\big(1 - F_X(\theta, x)\big)^{n-k}. \end{aligned}$$

Come nella prima dimostrazione usiamo il fatto che la densità si può vedere come derivata della funzione di ripartizione. Ne segue che per calcolare la densità sarà sufficiente calcolare la derivata in ciascun punto x della distribuzione appena trovata. In particolare si potrà vedere che coesisteranno il termine che vogliamo ottenere con altre due sommatorie, che tuttavia si elidono l'una con l'altra lasciando quindi la relazione espressa dal teorema. Si ha infatti che:

$$
\begin{aligned}
f_{(m)}(\theta,x) &= \frac{\partial}{\partial x}F_{X_{(m)}}(\theta,x) = \\
&= \sum_{k=m}^{n}\binom{n}{k}\cdot f_X(\theta,x)\Big\{kF_X^{k-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k} - (n-k)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1}\Big\} \\
&= m\binom{n}{m}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m} + \\
&\quad \sum_{k=m+1}^{n} k\binom{n}{k} f_X(\theta,x)F_X^{k-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k} - \\
&\quad \sum_{k=m}^{n-1}(n-k)\binom{n}{k} f_X(\theta,x)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1} \\
&= \frac{n!}{(m-1)!(n-k)!}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m} + \\
&\quad \sum_{j=m}^{n-1}(j+1)\binom{n}{j+1} f_X(\theta,x)F_X^{j}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-j-1} - \\
&\quad \sum_{k=m}^{n-1}(n-k)\binom{n}{k} f_X(\theta,x)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1} \\
&= \frac{n!}{(m-1)!(n-k)!}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m} + \\
&\quad \sum_{j=m}^{n-1}\frac{n!}{j!(n-j-1)!} f_X(\theta,x)F_X^{j}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-j-1} - \\
&\quad \sum_{k=m}^{n-1}\frac{n!}{k!(n-k-1)!} f_X(\theta,x)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1} \\
&= \frac{n!}{(m-1)!(n-k)!}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m}.
\end{aligned}
$$

□

Definizione. Sia $\big(X_{(1)},\dots,X_{(n)}\big)$ una statistica d'ordine di un campione casuale. Allora possiamo definire le nuove seguenti variabili:

- $X_{(n)} - X_{(1)}$, detta *Range* oppure *Misura di Dispersione*;
- $\frac{X_{(1)}+X_{(n)}}{2}$ detta *Mid Range* oppure *Misura di Centralità*;
- $$\left.\begin{array}{rl} \forall n \text{ pari} & \frac{X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}}{2} \\ \forall n \text{ dispari} & X_{(\frac{n+1}{2})} \end{array}\right\}$$ dette ciascuna *Mediana campionaria*;
- Sia $\frac{1}{2(n+1)} < p < 1-\frac{1}{2(n+1)}$, che possiamo in ogni caso pensare come $0<p<1$ per n molto grande. A questo punto possiamo definire l'intero $k_p \coloneqq \big\lfloor p(n+1)\big\rfloor + \big\lfloor 2\big(p(n+$

$1) - 2\lfloor p(n+1)\rfloor\big)\big\rfloor$, che risulta essere così ben definito in quanto compreso tra 1 e n e restituisce l'approssimazione all'intero più vicino al variare di p del reale $p(n+1)$.

- A questo punto, se scegliamo $\xi_p \in F_X^{-1}(p)$, chiameremo ξ_p *Quantile di popolazione* di ordine p. In seguito troveremo utile stimare tale valore. Perciò introduciamo la variabile casuale ad esso collegata $X_{(k_p)}$, detta *Quantile campionario* di ordine p. Se $p = \frac{i}{m}$, allora $X_{(k_p)}$ è detta anche i-esimo m-ile campionario. In particolare con Q_1 e Q_3 si indicano rispettivamente il primo e il terzo quartile.
  Intuitivamente, $X_{(k_p)}$ mi dà la v.c. che sta al k_p-esimo posto, ovvero al $p(n+1)$-esimo posto (se $p(n+1)$ è intero). Ad esempio, se $p = \frac{1}{3}$, $X_{(k_{\frac{1}{3}})}$ è la v.c. che nel vettore ordinato sta alla posizione $\frac{n+1}{3}$.
- Le variabili $LF := Q_1 - h$ e $UF := h + Q_3$, ove $h := \frac{3}{2}(Q_3 - Q_1)$ sono dette rispettivamente *Lower* e *Upper Fence*.

Osservazione: Osserviamo che più la misura di centralità si discosta dalla mediana, più vi è asimmetria nella funzione di densità f (*i.e.:* una funzione di distribuzione è simmetrica $:\Longleftrightarrow \exists x_0 \in \mathbb{R} : f(x_0 + x) = f(x_0 - x), \forall x \in \mathrm{Dom}(f)$.) Inoltre, ponendo che la funzione di ripartizione sia iniettiva e la funzione di densità sia simmetrica, si vede immediatamente che la media di popolazione, ovvero il quantile di popolazione di ordine $p = \frac{1}{2}$ coincide con il valore di aspettazione della variabile casuale, il quale a sua volta deve coincidere con x_0.

Dato un campione casuale di parametro $\theta \in \mathbb{R}$ fissato, sappiamo che una qualsiasi funzione di statistiche su tali variabili è, proprio per definizione, uno stimatore del parametro θ. L'esistenza di un'infinità non numerabile di stimatori è sicuramente un problema da ovviare in merito alla scelta tra essi di uno stimatore che effettivamente permetta di stimare il più correttamente possibile il parametro θ. Cercheremo dunque di individuare alcune proprietà che possano effettivamente giustificare la scelta di un determinato stimatore, affinché esso risulti il più possibile affidabile.

Definizione. Sia $(X_1 \dots, X_n)$ un campione di parametro θ e $T_n(X_1, \dots, X_n)$ uno stimatore. La funzione $B_\theta[T_n(X_1, \dots, X_n)] := \mathbb{E}_\theta[T_n(X_1, \dots, X_n)] - \theta$ si dice *distorsione* di T_n (nota: con $\mathbb{E}_\theta$ formalmente intendiamo semplicemente $\mathbb{E}$). In particolare T_n si dirà *non distorto* se e solo se la sua distorsione è nulla $\forall \theta \in \mathbb{R}$ (nel senso che uno stimatore -e.g. la media campionaria- non può stimare bene solo alcune medie, ma qualsiasi media reale, ad esempio la media di qualsiasi normale centrata in qualsiasi punto). Altrimenti si dice *distorto*. Se infine si ottiene che $\lim_{n\to+\infty} B_\theta[T_n(X_1, \dots, X_n)] = 0$, T_n si dice *asintoticamente non distorto*.

Esempio Sia $(X_{(1)}, \dots, X_{(n)})$ un campione casuale con distribuzione simmetrica (senza perdita di generalità, la assumiamo simmetrica rispetto all'origine) e scegliamo come stimatore T_n proprio la mediana campionaria. È chiaro innanzi tutto che essa in generale gode delle seguenti due proprietà:

- $\forall b \in \mathbb{R}$, $T_n(X_1 + b, \dots, X_n + b) = T_n(X_1, \dots, X_n) + b$;
- $T_n(-X_1, \dots, -X_n) = -T_n(X_1, \dots, X_n)$.

Abbiamo inoltre che la distribuzione di $(X_1, \ldots, X_n)$ e del vettore $(-X_1, \ldots, -X_n)$ coincidono (ricordando che l'origine è il centro di simmetria). Si avrà dunque:

$$\begin{aligned}\mathbb{E}[T_n] &= \mathbb{E}[T_n(X_1, \ldots, X_n)] = \mathbb{E}[T_n(-X_1 \ldots, -X_n)] \\ &= \mathbb{E}[-T_n(X_1, \ldots, X_n)] = -\mathbb{E}[T_n(X_1 \ldots, X_n)] \\ &= -\mathbb{E}[T_n]\end{aligned}$$

perciò, in definitiva, $2\mathbb{E}[T_n] = 0$ ovvero $\mathbb{E}[T_n] = 0$. Quindi, nel caso di una distribuzione simmetrica, la media campionaria è uno stimatore non distorto del valore di aspettazione, del punto di simmetria e della media della popolazione (dato che tutti loro nel nostro caso coincidono).

Esempio Sia $(X_1, \ldots, X_n)$ un campione casuale di parametro $\theta \in \mathbb{R}$ fissato e con $\mu \coloneqq \mathbb{E}[X]$, $\sigma^2 \coloneqq \mathrm{Var}[X]$. Vogliamo provare a calcolare la distorsione di due stimatori "classici":

1. Scegliamo come stimatore la *media campionaria* $\overline{X}_n \coloneqq \frac{1}{n}\sum_{i=1}^n X_i$. Allora $B_\theta[\overline{X}_n] = \mathbb{E}_\theta[\overline{X}_n] - \theta = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] - \theta = \frac{1}{n} \cdot n\mathbb{E}_\theta[X] = \mu - \theta$. Dunque la distorsione è costante $\forall n \in \mathbb{N}$. In particolare è uno stimatore non distorto per il valore di aspettazione μ. Possiamo calcolare facilmente anche la varianza di $\overline{X}_n$ che risulta essere $\frac{\sigma^2}{n}$. La media campionaria si rivela essere quindi un buon stimatore. (nota: nel calcolo della varianza stiamo trattando lo stimatore come una variabile casuale essa stessa, quindi più la varianza è piccola migliore è lo stimatore).

2. Prendiamo ora come stimatore la *varianza campionaria*, data dalla variabile $S_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Allora:
   $\mathbb{E}[S_n^2] = \frac{1}{n-1}\sum_{i=1}^n \mathbb{E}[(X_i - \overline{X}_n)^2] = \frac{1}{n-1}\sum_{i=1}^n \mathbb{E}\left[\left((X_i - \mu) - (\overline{X}_n - \mu)\right)^2\right] =$
   $= \frac{1}{n-1}\left(\sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2] - \sum_{i=1}^n \mathbb{E}[(\overline{X}_n - \mu)^2]\right) = \frac{1}{n-1} \cdot (n-1)\sigma^2 = \sigma^2$. Perciò S_n^2 è uno stimatore non distorto di σ^2. Notiamo che lo stimatore $S_n^* \coloneqq \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ avrebbe distorsione $-\frac{\sigma^2}{n}$, e dunque è peggiore della varianza campionaria, anche se è asintoticamente non distorto.

Calcoleremo adesso la varianza della varianza campionaria. Assumiamo per il momento che il campione provenga da una distribuzione normale $N(\mu, \sigma^2)$. In tal caso mostriamo che $\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2_{n-1}$ e dunque si avrà subito $\mathrm{Var}[S_n^2] = \frac{2\sigma^4}{n-1}$. Infatti, $\frac{n-1}{\sigma^2}S_n^2 = \sum_{i=1}^n \left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2 =$
$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} - \frac{\overline{X}_n - \mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 - n\left(\frac{\overline{X}_n - \mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \Rightarrow$
$\sum_{i=1}^n \left(\sim \chi^2_1\right) - \left(\sim \chi^2_1\right) \Rightarrow \frac{n-1}{\sigma^2}S_n^2 \sim \chi^2_{n-1}$, ove abbiamo usato il seguente teorema:

Teorema 15. *Sia $(X_1, \ldots, X_n)$ un campione casuale ove la funzione generatrice di ciascuna X_i, $1 \le i \le n$ è $M_X(t)$. Allora $M_{\overline{X}_n}(t) = (M_X(\frac{t}{n}))^n$.*

per mostrare che $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$. Infatti:

$M_{\overline{X}_n}(t) = (M_X(\frac{t}{n}))^n = \left(e^{\mu\frac{t}{n} + \frac{\sigma^2 t^2}{2n^2}}\right)^n = e^{\mu t + \frac{\sigma^2}{n}\cdot\frac{t}{2}}$, da cui la tesi.

Teorema 16. *Sia $(X_1, \dots, X_n)$ un campione casuale da una popolazione con distribuzione discreta o assolutamente continua dove la densità associata sia della forma $f(x, \theta) = C(x)D(\theta)\exp\{\sum_{m=1}^{k} A_m(\theta)B_m(x)\}$ con k naturale positivo. Siano $T_1, \dots, T_k$ statistiche definite $\forall 1 \leq m \leq k$ da $T_m(X_1, \dots, X_n) \coloneqq \sum_{i=1}^{n} B_m(X_i)$. Allora la distribuzione di $(T_1, \dots, T_k)$ sarà ancora della forma esponenziale:*

$$f_{(T_1,\dots,T_k)}(\theta, t_1, \dots, t_k) = C(t_1, \dots, t_k)D(\theta)^n exp\left\{ \sum_{i=1}^{k} A_m(\theta)t_m \right\}$$

.

Esempio. Sia $(X_1, \dots, X_n)$ un campione casuale. Supponiamo che $X \sim \text{Bin}(1, p)$. Allora la densità sarà discreta, ossia sarà $f(p, x) = \mathbb{P}[X = x] = \mathbb{1}_{\{0,1\}}(x)(1-p)\exp\left\{x \log\left(\frac{p}{1-p}\right)\right\}$. Applicando il teorema otteniamo $T_1(X_1, \dots, X_n) = \sum_{i=1}^{n} B_1(X_i) = \sum_{i=1}^{n} X_i$ da cui si deduce subito che $T_1 \sim \text{Bin}(n, p)$. In particolare possiamo scriverne la densità: $f_{T_1}(p, t_1) = \mathbb{1}_{\{0,\dots,n\}}(t_1)\binom{n}{t_1}(1-p)^n \exp\{t_1 \log\left(\frac{p}{1-p}\right)\}$.
Esempio. Sia $(X_1, \dots, X_n)$ un campione casuale e supponiamo che il nostro campione casuale abbia distribuzione uniforme $\text{Unif}([0, \theta])$. Vogliamo stimare θ. Supponiamo che, essendo θ il massimo valore che ciascuna variabile può assumere, un plausibile buon stimatore possa essere proprio il massimo della statistica ordinata, ovvero $T_n(X_1, \dots, X_n) \coloneqq X_{(n)}$. Per il teorema di pagina 8, ne conosciamo già la distribuzione:

$$f_{T_n}(\theta, x) = \frac{n!}{(n-1)!(n-n)!}\frac{1}{\theta}\left(\frac{x}{\theta}\right)^{n-1}\left(1 - \frac{x}{\theta}\right)^{n-n} = \frac{n}{\theta^n}x^{n-1}$$

Allora $\mathbb{E}[T_n] = \int_0^\theta \frac{n}{\theta^n}x^n dx = \frac{n}{n+1}\theta \neq \theta \implies B_\theta[T_n] = \frac{-\theta}{n+1}$. Ne segue che è distorto, ma asintoticamente non distorto per θ. Possiamo anche calcolarne la varianza: $\text{Var}[T_n] = \mathbb{E}[T_n^2] - \mathbb{E}[T_n]^2 = \int_0^\theta \frac{n}{\theta^n}x^{n+1}dx - \frac{n}{n+1} = \frac{n\theta^2}{(n+1)^2(n+2)} \xrightarrow[n\to\infty]{} 0$. Perciò il massimo $X_{(n)}$ rimane in ogni caso uno stimatore affidabile. Osserviamo che possiamo tuttavia introdurre un nuovo stimatore che ci assicura la non distorsione, ovvero $T_n^* \coloneqq \frac{n+1}{n}T$, che possiede le proprietà cercate.

Definizione. Sia $(X_1, \dots, X_n)$ un campione casuale con distribuzione $F(\theta, x)$, ove $\theta \in \Theta \subset \mathbb{R}$. Sia poi T_n una statistica $\forall n \in \mathbb{N}$. Diremo T_n essere uno stimatore *consistente* di $\theta :\iff T_n(X_1, \dots, X_n) \xrightarrow[n\to\infty]{\mathbb{P}} \theta$.

Esempio. Sia $(X_1, \dots, X_n)$ un campione casuale, ove $X \in \mathcal{L}^2(\mathbb{R})$. Indichiamo come al solito media e varianza rispettivamente con μ e σ^2. Allora abbiamo:

1. La media campionaria $\overline{X}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P}} \mu$, grazie alla legge debole dei grandi numeri poiché $\lim_{n\to+\infty} \mathbb{P}[(\overline{X}_n - \mu) > \varepsilon] = 0$, $\forall \varepsilon > 0$.
2. Consideriamo adesso la varianza campionaria

$$S_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right)$$

. Abbiamo ora i seguenti tre termini:

- $\lim_{n\to+\infty} \frac{n}{n-1} = 1$, un semplice limite;
- $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[X^2]$, ancora grazie alla legge debole dei grandi numeri e al fatto che X^2 rimane ancora sommabile;
- $\overline{X}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mu^2 = \mathbb{E}[X]^2$ grazie al Teorema 4 sulla convergenza.

Ne segue quindi che $S_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2$, sempre per i teoremi sulla convergenza di somma, prodotto e prodotto per costanti di variabili casuali.

3. Consideriamo ancora il campione casuale distribuito uniformemente $\mathrm{Unif}([0,\theta])$ con stimatore $T_n(X_1 \dots, X_n) \coloneqq X_{(n)}$. Troviamo che anch'esso è consistente per la stima del massimo. Infatti, $\mathbb{P}[|T_n - \theta| > \varepsilon] = \mathbb{P}[\theta - T_n > \varepsilon] = \mathbb{P}[X_{(n)} \leq \theta - \varepsilon] = F_{X_{(n)}}(\theta - \varepsilon) = \left(1 - \frac{\varepsilon}{\theta}\right)^n \xrightarrow[n\to\infty]{} 0$. Allo stesso modo si può verificare che anche T_n^* è consistente per θ.

Definizione. Sia $(X_1, \dots, X_n)$ un campione casuale e $T_n : \mathfrak{X} \longrightarrow \mathcal{Y}_{T_n}$ una statistica (stimatore). Vi sia inoltre una funzione di parametri $a : \Theta \longrightarrow \mathcal{Y}_\Theta$. Allora la funzione non negativa $\mathrm{Loss} : (\mathcal{Y}_{T_n} \cup \mathcal{Y}_\Theta) \times \mathcal{Y}_\Theta \longrightarrow \mathbb{R}_{\geq 0}$ viene detta *Funzione di Perdita* se soddisfa alle seguenti condizioni:

1. $\mathrm{Loss}\big(a(\theta), a(\theta)\big) = 0$, $\forall \theta \in \Theta$;
2. Per ogni $T_n \in \mathcal{T}$, esiste una funzione $\mathrm{Risk} : Y_{T_n} \times Y_\Theta \longrightarrow \mathbb{R}$, detta *Funzione di Rischio*, tale che $\mathrm{Risk}\big(T_n, a(\theta)\big) = \mathbb{E}_\theta[\mathrm{Loss}\big(T_n, a(\theta)\big)]$, $\forall \theta \in \Theta$.

Osservazione. La funzione di perdita può essere pensata come una misura della discrepanza tra l'azione T_n e lo stato della natura $a(\theta)$.

Definizione. Possiamo già definire due tipologie di funzioni di perdita che spesso vengono utilizzate in statistica:

1. $\mathrm{Loss}_1\big(T_n, a(\theta)\big) \coloneqq |T_n - a(\theta)|$, chiamata *Errore assoluto*;
2. $\mathrm{Loss}_2\big(T_n, a(\theta)\big) \coloneqq \big(T_n - a(\theta)\big)^2$. Essa ammette anche come possibile funzione di rischio $\mathrm{Risk}_2\big(T_n, a(\theta)\big) \coloneqq \mathbb{E}_\theta[T_n - a(\theta)]^2$; se tuttavia $a = \mathrm{id}_\Theta$, allora la funzione $\mathrm{MSE}_\theta(T_n) \coloneqq \mathrm{Risk}_2\big(T_n, \theta\big)$ prende il nome di *Mean Square Error* (oppure *Errore Quadratico Medio*).

Osservazione. Semplicemente aggiungendo e sottraendo il valore $\mathbb{E}[T_n]$ si ottiene subito la seguente uguaglianza: $\mathrm{MSE}_\theta(T_n) = \mathrm{Var}_\theta[T_n] + B_\theta[T_n]^2$.

Teorema 17. *Sia T_n uno stimatore di θ (non necessariamente non distorto). Allora si ha che $\lim_{n\to+\infty} \mathrm{MSE}_\theta(T_n) = 0$ è condizione sufficiente (ma non necessaria) per la consistenza di T_n.*

Dimostrazione. Si ha infatti la seguente semplice catena di diseguaglianze:

$$\begin{aligned} \mathbb{P}[|T_n - \theta| > \varepsilon] &= \int_{|T_n-\theta|>\varepsilon} f_{T_n}(\theta, t_n)dt_n \\ &< \int_{|T_n-\theta|>\varepsilon} \frac{(t_n-\theta)^2}{\varepsilon^2} f_{T_n}(\theta, t_n)dt_n < \frac{1}{\varepsilon^2}\,\mathrm{MSE}_\theta(T_n). \end{aligned}$$

□

1.2.2 Intervalli di confidenza

Lezione del 15/03, ultima modifica 26/03, Michele Nardin

Sia $(X_1, ..., X_n)$ un campione casuale definito da una variabile casuale avente funzione di ripartizione $F_X(x, \vartheta)$. Vogliamo stimare l'incognita ϑ, e per farlo ci serviamo di uno stimatore T_n. Una volta estratto il campione casuale, e quindi in possesso di una n-upla di valori reali $(x_1, ..., x_n)$ che ne rappresenta una determinazione, possiamo effettivamente calcolare valore della nostra stima: è impensabile però che la stima *coincida esattamente* con il valore incognito (se X ha distribuzione continua $\mathbb{P}(T_n = \vartheta) = 0$!). Dobbiamo quindi associare a T_n un *margine di errore*.

Introduciamo innanzitutto il concetto di Statistica Pivot:

Definizione 8. *Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione con funzione di ripartizione $F_X(x, \vartheta)$, $\vartheta \in \Theta$. Definiamo Statistica Pivot una funzione $Q((X_1, ..., X_n), \vartheta)$ tale che*

1. *Q è funzione del campione casuale e del parametro ϑ (parametro su cui si vuol fare inferenza)*
2. *Q non contiene parametri incogniti oltre a ϑ*
3. *la distribuzione di Q, F_Q, è completamente nota (ossia non dipende da ϑ)*
4. *Q è invertibile rispetto a ϑ*

Esempi: Campione casuale da $N(\mu, \sigma^2)$:

1. Supponiamo di conoscere la varianza: allora un esempio di statistica pivot è
$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$
la quale, grazie all'ipotesi di campionamento da vc normale, ha distribuzione N(0,1).
2. Supponiamo di non conoscere la varianza: in tal caso, al posto della varianza usiamo lo stimatore varianza campionaria S_n^2, il quale è non distorto (già dimostrato) e consistente (infatti $\text{MSE}_{\sigma^2}(S_n^2) = Var(S_n^2) + B^2(S_n^2) = \frac{2\sigma^4}{n-1} \to 0$) e quindi la statistica pivot in questione sarà
$$Q = \frac{\bar{X}_n - \mu}{\S_n/\sqrt{n}}$$
la quale (dimostreremo che) ha distribuzione t-student con n-1 gradi di libertà.

Esempio introduttivo (exit poll)

Vogliamo stimare la proporzione p_i dei voti ricevuti dall'iesimo partito sul totale. Il nostro problema sarà quello di trovare un intervallo centrato nella stima $\hat{p}_i$, ed un margine d'errore, ME, tale per cui, ad una fissata soglia di probabilità α si abbia

$$P(p_i \in (\hat{p}_i - ME, \hat{p}_i + ME)) = 1 - \alpha$$

Costruzione generale

In generale, sia ϑ_0 il valore vero del parametro ϑ che vogliamo stimare, e per semplicità assumiamo che T_n sia un suo stimatore tale che

$$\sqrt{n}(T_n - \vartheta_0) \xrightarrow{d} N(0, \sigma^2_{T_n})$$

Per il momento assumiamo di conoscere $\sigma^2_{T_n}$, sicché

$$Z_n = \frac{\sqrt{n}(T_n - \vartheta_0)}{\sigma_{T_n}} \overset{a}{\sim} N(0,1)$$

Bisogna notare che Z_n è una statistica pivot. Fissato $\alpha \in (0,1)$, consideriamo i quantili della distribuzione N(0,1), $\pm z_{\alpha/2}$ (ossia quei valori tali per cui, se $X \sim N(0,1)$, $P(-z_{\alpha/2} \leq X \leq z_{\alpha/2}) = 1-\alpha$). Possiamo affermare che, per n sufficientemente grande, (il simbolo $\doteq$ indica un'uguaglianza approssimata)

$$P(-z_{\alpha/2} \leq Z_n \leq z_{\alpha/2}) \doteq 1 - \alpha$$

da cui

$$P(-z_{\alpha/2} \leq \frac{\sqrt{n}(T_n - \vartheta_0)}{\sigma_{T_n}} \leq z_{\alpha/2}) \doteq 1 - \alpha$$

e ancora

$$P(T_n - z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}} \leq \vartheta_0 \leq T_n + z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}) \doteq 1 - \alpha$$

Possiamo quindi definire un intervallo casuale,

$$IC = \left[T_n - z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}, T_n + z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}\right]$$

(è casuale perchè per T_n è una vc). IC è uno Stimatore Intervallare. Si può affermare che $P(\vartheta \in IC) \doteq 1 - \alpha$.

Nomenclatura : $z_{\alpha/2}$ si dice Fattore di Affidabilità, $\frac{\sigma_{T_n}}{\sqrt{n}}$ si dice Standard Error dello stimatore T_n.

Sia ora $(x_1, ..., x_n)$ una determinazione campionaria (ossia i dati effettivamente osservati da un campione casuale) (cioè una n-upla) e sia $T_n(x_1, ..., x_n) = t_n$ l'effettivo valore assunto dallo stimatore.

Definiamo di seguito *l'intervallo di confidenza con probabilità di copertura* $1 - \alpha$

$$IC_\vartheta(1-\alpha) := \left[t_n - z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}, t_n + z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}\right]$$

La probabilità di copertura viene anche detta livello di confidenza.

Nella pratica, $\sigma^2_{T_n}$ non è noto a priori. Possiamo però usare lo stimatore varianza campionaria di T_n, $S^2_{T_n}$, il quale sappiamo che converge in probabilità a $\sigma^2_{T_n}$. Allora, per il teorema 10 (di Slutsky), troviamo che

$$Z_n = \frac{\sqrt{n}(T_n - \vartheta_0)}{S_{T_n}} = \frac{\sqrt{n}}{S_{T_n}} T_n - \frac{\sqrt{n}}{S_{T_n}} \vartheta_0 \xrightarrow{d} N(0,1)$$

Possiamo quindi ripetere il ragionamento fatto poco sopra usando la varianza campionaria al posto di $S^2_{T_n}$, e quindi costruire l'intervallo di confidenza con probabilità di copertura pari a $1-\alpha$ come

$$IC_\vartheta(1-\alpha) := \left[t_n - z_{\alpha/2} \frac{S_{T_n}}{\sqrt{n}}, t_n + z_{\alpha/2} \frac{S_{T_n}}{\sqrt{n}} \right]$$

Intervallo di confidenza per la media μ

Sia $(X_1, ..., X_n)$ un campione casuale, media e varianza incognite. Siano $\bar{X}_n$ e S^2_n gli stimatori di media e varianza della popolazione. Allora per il TLC e per il teorema di Slutsky si ha che

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1)$$

che è una statistica pivot. Quindi l'intervallo di confidenza con probabilità di copertura $1-\alpha$ (sempre approssimato) sarà

$$IC_\mu(1-\alpha) = \left[\bar{X}_n - z_{\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

Intervallo di confidenza per una proporzione p

Sia $(X_1, ..., X_n)$ un campione casuale da $b(1,p)$ e sia $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$ lo stimatore (corretto e consistente) di p. Troviamo che per il TLC e per la WLLN (legge dei grandi numeri)

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{d} N(0,1)$$

e quindi l'intervallo di confidenza con probabilità di copertura $1-\alpha$ approssimato sarà

$$IC_p(1-\alpha) = \left[\hat{p}_n - z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \frac{S_n}{\sqrt{n}} \right]$$

Distribuzione esatta della statistica pivot: distribuzione t di Student

Lezione del 18/03, ultima modifica 26/03, Michele Nardin

La distribuzione t di Student con ν gradi di libertà è definita come $T = \frac{Z}{\sqrt{S^2/\nu}}$ ove $Z \sim N(0,1)$ mentre $S^2 \sim \chi^2_\nu$ (chiquadro con ν gradi di libertà). La funzione di densità è

$$f_{t_\nu}(t,\nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\pi\nu}} \frac{1}{[1+t^2/\nu]^{\frac{v+1}{2}}} \mathbb{1}_{\mathbb{R}}(t)$$

tale funzione è simmetrica, ha la classica forma a campana come la normale, ma a differenza di quest'ultima ha le code più pesanti. Risulta che la statistica pivot per la media in campioni poco numerosi [1] (in caso di campionamento da normale) ha distribuzione esatta t di Student. Infatti

$$Q = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S_n^2}{\sigma^2}}}$$

troviamo al numeratore $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, (grazie al fatto che le X_i sono equi distribuite normalmente) mentre al denominatore abbiamo che

$$\sqrt{\frac{S_n^2}{\sigma^2}} = \sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}} = \sqrt{\frac{H}{(n-1)}}$$

Abbiamo già dimostrato che $H = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$, quindi in definitiva al denominatore abbiamo la radice di una chiquadro diviso i suoi gradi di libertà, ovvero siamo proprio in presenza di una distribuzione t di Student.

Osservazione importante: Quindi, quando il campione casuale è poco numeroso, è conveniente usare i quantili della distribuzione t di student per costruire gli intervalli di confidenza. Per numerosità campionarie $n > 30$, approssimare la distribuzione t di student con la distribuzione normale offre risultati soddisfacenti. Ricordiamo che per il tlc $Q \to N(0,1)$)

Intervallo di confidenza esatto

Fissato un livello di confidenza $1-\alpha$, consideriamo i quantili della distribuzione t di student (con n-1 gradi di libertà, ove n è la dimensione campionaria) $\pm t_{(\alpha/2;n-1)}$, troviamo

$$P\left(-t_{(\alpha/2;n-1)} \leq \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \leq t_{(\alpha/2;n-1)}\right) = 1-\alpha$$

Notiamo che questa volta vale l'uguaglianza 'vera', poiché non stiamo considerando approssimazioni asintotiche. In presenza del campione effettivamente estratto, $(x_1, ..., x_n)$, scriviamo $\overline{x}_n$ e s_n^2 i valori assunti da media e varianza campionaria, l'intervallo di confidenza è

$$IC_\mu(1-\alpha) = \left[\overline{x}_n - t_{(\alpha/2;n-1)}\sqrt{\frac{s_n^2}{n}}, \overline{x}_n + t_{(\alpha/2;n-1)}\sqrt{\frac{s_n^2}{n}}\right]$$

Osservazione 5. *Alcune osservazioni che, pur sembrando banali, è bene tenere a mente:*

1. *Al crescere del livello di confidenza $(1-\alpha)$ e/o della varianza campionaria S_n^2 cresce anche l'ampiezza di IC*
2. *Al crescere dell'ampiezza campionaria n, (fermo restando il livello di confidenza) l'ampiezza di IC diminuisce*

[1] In realtà vale per tutti i campioni, è solo che da un certo punto in poi la differenza con la normale è davvero trascurabile! Sulle tavole si riporta solo per $\nu < 120$

Intervalli di confidenza per la varianza

Sia $(X_1, ..., X_n)$ un campione casuale da $N(\mu, \sigma^2)$. Consideriamo la statistica pivot

$$W = \frac{n-1}{\sigma^2} S_n^2$$

Abbiamo già mostrato che $W \sim \chi^2_{n-1}$. Ma allora, dato che noi cerchiamo q_1, q_2 t.c.

$$P\left(q_1 \leq \frac{n-1}{\sigma^2} S_n^2 \leq q_2\right) = 1 - \alpha$$

troviamo che essi sono i quantili di ordine $\alpha/2$ e $1 - \alpha/2$ della chiquadro con n-1 gradi di libertà, che indicheremo $q_1 = \chi^2_{(n-1,\alpha/2)}$ e $q_2 = \chi^2_{(n-1,1-\alpha/2)}$. Con qualche passaggio otteniamo:

$$P\left(\frac{1}{q_2} \leq \frac{\sigma^2}{(n-1)S_n^2} \leq \frac{1}{q_1}\right) = 1 - \alpha$$

$$P\left(\frac{(n-1)S_n^2}{q_2} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{q_1}\right) = 1 - \alpha$$

Troviamo così l'intervallo casuale (e di conseguenza il relativo intervallo di confidenza, una volta estratto il campione e trovato un valore a S_n^2)

$$IC = \left[\frac{(n-1)S_n^2}{q_2}; \frac{(n-1)S_n^2}{q_1}\right]$$

Intervalli di confidenza per la differenza di medie

Vogliamo confrontare due distribuzioni: *sintetizziamo* la differenza tra due popolazioni tramite la differenza delle loro media.

Supponiamo inizialmente di avere due campioni casuali tra loro indipendenti:
$(X_1, ..., X_{n_1})$ da una distribuzione D1, con media μ_1 (ignota) e varianza σ_1^2 (nota)
$(Y_1, ..., Y_{n_2})$ da una distribuzione D2, con media μ_2 (ignota) e varianza σ_2^2 (nota)
NB: non necessariamente n_1 dev'essere uguale a n_2

Consideriamo gli stimatori media campionaria per le due medie, che indicheremo con $\overline{X}$ e $\overline{Y}$. La statistica pivot che ci interessa per $\Delta = \mu_1 - \mu_2$ sarà

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]^{\frac{1}{2}}}$$

Notiamo che $var(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ dato che $cov(\overline{X}, \overline{Y}) = 0$ per l'indipendenza. Ma allora $Z \overset{a}{\sim} N(0,1)$, quindi possiamo trovare un intervallo di confidenza [2]

$$IC_\Delta(1-\alpha) = [(\overline{X} - \overline{Y}) - ME; (\overline{X} - \overline{Y}) + ME]$$

[2] Approssimato, visto che conosciamo solo l'andamento asintotico di Z! D1 e D2 non è detto che siano mormali!

ove $ME = z_{\alpha/2}\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]^{\frac{1}{2}}$. Al posto delle varianze possiamo usare anche gli stimatori corretti e consistenti varianza campionaria, e giungere allo stesso risultato per il teorema di Slutsky.

In generale non conosciamo la varianza delle distribuzioni: in base al problema che dobbiamo affrontare, può essere plausibile supporre di conoscere la distribuzione delle due popolazioni a meno di uno o più parametri.

Location Model: Supponiamo di avere $(X_1, ..., X_{n_1})$ da distribuzione normale con media μ_1 e varianza σ_1^2 (ignote), i loro stimatori $\overline{X}$ e S_1^2 e $(Y_1, ..., Y_{n_2})$ da distribuzione normale con media μ_2 e varianza σ_2^2 (ignote) e i loro stimatori $\overline{Y}$ e S_2^2. Supponiamo che i due campioni siano tra loro indipendenti ed inoltre che $\sigma_1 = \sigma_2 = \sigma$. Possiamo 'fondere' le informazioni contenute in S_1^2 e S_2^2:

$$(PooledVariance)\ S_p^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

che risulta essere uno stimatore corretto e consistente di σ^2 (esercizio). La statistica Pivot che prendiamo in considerazione sarà

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}}$$

la quale risulta essere distribuita come $t_{n_1+n_2-2}$. Ricalcando i passaggi delle applicazioni precedenti, fissato α troviamo l'intervallo casuale per Δ

$$IC = \left[(\overline{X} - \overline{Y}) - t_{(n_1+n_2-2;\alpha/2)}S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}; (\overline{X} - \overline{Y}) + t_{(n_1+n_2-2;\alpha/2)}S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}\right]$$

Intervalli di confidenza per la differenza di proporzioni

Supponiamo di avere $(X_1, ..., X_{n_1})$ da distribuzione $b(1, p_1)$, con stimatore $\hat{p_1}$ e $(Y_1, ..., Y_{n_2})$ da distribuzione $b(1, p_2)$, con stimatore $\hat{p_2}$. Supponiamo che i due campioni siano tra loro indipendenti. Allora

$$\Delta = \hat{p_1} - \hat{p_2} \overset{a}{\sim} N\left(p_1 - p_2, \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}\right)$$

quindi usando la statistica Pivot

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)}} \overset{a}{\sim} N(0, 1)$$

trovo l'intervallo di confidenza

$$IC_\Delta(1 - \alpha) = \left[(\hat{p_1} - \hat{p_2}) - z_{\alpha/2}\sqrt{A(p_1, p_2)}; (\hat{p_1} - \hat{p_2}) + z_{\alpha/2}\sqrt{A(p_1, p_2)}\right]$$

ove $A(p_1, p_2) = \left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$. Ovviamente al posto di p_1 e p_2 uso gli stimatori corretti e consistenti $\hat{p_1}$ e $\hat{p_2}$.

Lezione del 20 marzo, ultima modifica 26 marzo, Michele Nardin

Intervalli di confidenza per rapporti di varianze

Introduciamo per prima cosa la *Distribuzione F di Snedecor-Fisher*.

Siano $W_1 \sim \chi^2_{\nu_1}$ e $W_2 \sim \chi^2_{\nu_2}$ indipendenti. La distribuzione F_{ν_1,ν_2} è definita come il rapporto tra due chiquadrato divise per i rispettivi gradi di libertà, in formule

$$W = \frac{W_1/\nu_1}{W_2/\nu_2}$$

ed ha funzione di densità

$$f_W(w;\nu_1,\nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2}) + \Gamma(\frac{\nu_2}{2})}(\nu_1/\nu_2)^{\nu_1/2} w^{\nu_1/2-1}\left[1 - \frac{\nu_1}{\nu_2}w\right]^{\frac{\nu_1+\nu_2}{2}} \mathbb{1}_{\mathbb{R}^+}(w)$$

Vale la seguente proprietà, utile per calcolare i quantili non tabulati:

$$Se\ W \sim F_{\nu_1,\nu_2} \Rightarrow \frac{1}{W} \sim F_{\nu_2,\nu_1}$$

Le tavole (comunemente) forniscono i valori dei quantili per $(1-\alpha) \in \{0.80, 0.90, 0.95, 0.975, 0.99, 0.999\}$. Quindi possiamo sfruttare la proprietà sopra scritta per trovare che

$$w_{\alpha;\nu_1,\nu_2} = \frac{1}{w_{1-\alpha;\nu_1,\nu_2}}$$

Intervallo di confidenza per rapporti di varianze

Supponiamo di avere $(X_1, ..., X_{n_1})$ da distribuzione normale con media μ_1 e varianza σ_1^2 (ignote), i loro stimatori $\overline{X}$ e S_1^2 e $(Y_1, ..., Y_{n_2})$ da distribuzione normale con media μ_2 e varianza σ_2^2 (ignote) e i loro stimatori $\overline{Y}$ e S_2^2. Ricordiamo che $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ (idem per S_2^2).

Fissato $1-\alpha$ consideriamo una statistica pivot e w_1, w_2 tc $P(w_1 \leq W \leq w_2) = 1-\alpha$. La statistica pivot in questione sarà

$$W = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/n_1 - 1}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/n_2 - 1} \sim F_{(n_1-1),(n_2-1)}$$

poichè rapporto di due chiquadro. Risulta inoltre, semplificando:

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

w_1 e w_2 saranno i quantili di ordine $\alpha/2$ e $1-\alpha/2$ della distribuzione $F_{(n_1-1),(n_2-1)}$, esplicitamente $w_1 = w_{(n_1-1,n_2-1;\alpha/2)} = \frac{1}{w_{(n_2-1,n_1-1;1-\alpha/2)}}$ e $w_2 = w_{(n_1-1,n_2-1;1-\alpha/2)}$ e quindi troviamo che

$$1-\alpha = P(w_1 \leq \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \leq w_2) = P\left(\frac{S_1^2/S_2^2}{w_2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2/S_2^2}{w_1}\right)$$

da cui troviamo anche il relativo intervallo casuale. Quando invece abbiamo un campione effettivamente estratto, posti s_1^2 e s_2^2 i valori assunti dalla varianza campionaria, l'intervallo di confidenza sarà

$$IC_{\frac{\sigma_1^2}{\sigma_2^2}}(1-\alpha) = \left[\frac{s_1^2/s_2^2}{w_2}; \frac{s_1^2/s_2^2}{w_1}\right] = \left[\frac{s_1^2/s_2^2}{w_{(n_1-1,n_2-1;1-\alpha/2)}}; \frac{s_1^2/s_2^2}{\frac{1}{w_{(n_2-1,n_1-1;1-\alpha/2)}}}\right]$$

1.2.3 Test per la verifica di ipotesi

Lezione del 25/03, ultima modifica 20/05, Andrea Gadotti

La procedura di test per la verifica di ipotesi che descriveremo a breve cerca di fornire una soluzione ai seguenti problemi:

1. Determinare quanto un'ipotesi è realistica, verosimile, compatibile con l'informazione empirica a disposizione.
2. Trovare un ragionamento oggettivo (matematico) per inferire dall'informazione disponibile (ovvero il contenuto di un campione) circa la veridicità dell'ipotesi formulata.
3. Misurare in qualche modo questa "vicinanza" tra ipotesi e realtà.

Useremo statistiche pivot in ambito parametrico: la distribuzione da cui proviene il campione casuale $(X_1, ..., X_n)$ è nota a meno di uno o più parametri.
Di seguito si può vedere una descrizione generale della situazione che si riproporrà per tutta questa sezione:

$X \sim F_X(\underline{x}, \theta)$, $\theta \in \Theta$, $(X_1, ..., X_n)$ iid. Le nostre due ipotesi saranno della forma:

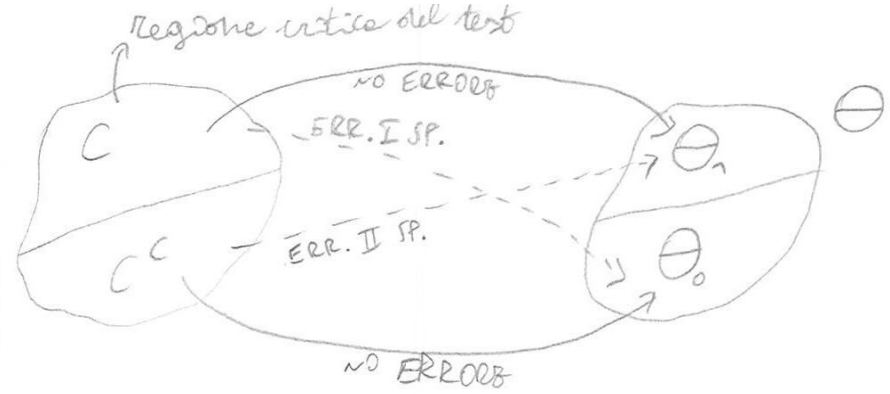
$$\begin{cases} H_0: & \theta \in \Theta_0 \\ H_1: & \theta \in \Theta_1 \end{cases}$$

con $\Theta_0 \cup \Theta_1 = \Theta$.

Chiameremo H_0 *ipotesi nulla* e H_1 *ipotesi alternativa*. Di solito H_0 rappresenta la conoscenza pregressa, la supposizione vera fino a prova contraria. Al contrario, H_1 è l'ipotesi di lavoro, quella su cui ripieghiamo nel momento in cui il nostro test risulta in contraddizione con H_0.
Il test si riduce a una *regola di decisione* in merito a H_0 e H_1 sulla base del campione

casuale $(X_1, ..., X_n)$ da $X \sim F_X(x;\theta)$. Dividiamo lo spazio dei campioni in due regioni disgiunte: C (regione critica del test) e C^c. La decisione può chiaramente essere corretta, ma anche errata, poiché il campione costituisce un'informazione non completa. Risulta quindi necessario formulare delle *conslusioni in probabilità*, ovvero associare alla nostra conclusione la probabilità che questa sia corretta, cercando ovviamente di massimizzarla. Possiamo riassumere le varie possibilità nella tabella e nel disegno sottostanti:

	H_0 è vera	H_0 è falsa
Rifiuto H_0	errore di I specie	nessun errore
Non rifiuto H_0	nessun errore	errore di II specie



Esempio Lancio di una moneta onesta. Consideriamo il campione casuale $(X_1, ..., X_n)$ e il numero di teste $S_n = \sum_{i=1}^n X_i$. Vorremmo stimare la probabilità che esca testa con la media campionaria $\overline{X}_n$. In questo caso potremmo avere:

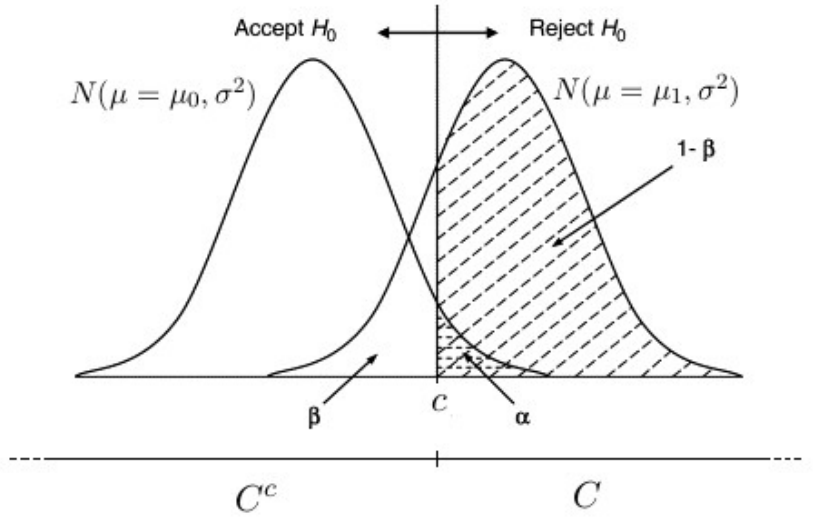
$$\begin{cases} H_0: & p = 1/2 \\ H_1: & p \neq 1/2 \end{cases}$$

La regola di decisione consiste quindi nel rifiutare H_0 se $(X_1, ..., X_n) \in C$ e invece rifiutare H_0 se $(X_1, ..., X_n) \in C^c$. Ci piacerebbe trovare una regola di decisione che permetta di minimizzare la probabilità di commettere errori di I o II tipo. Purtroppo questo non è possibile, per la natura stessa della relazione che corre tra gli errori di I e II tipo. Di seguito un esempio che ci dà un'idea del perché:

Esempio Consideriamo un campione casuale $(X_1, ..., X_n)$ da $N(\mu, \sigma^2)$ con σ^2 noto. Supponiamo che le nostre due ipotesi siano:

$$\begin{cases} H_0: & \mu = \mu_0 \quad \text{ovvero } N(\mu = \mu_0, \sigma^2) \\ H_1: & \mu = \mu_1 \quad \text{ovvero } N(\mu = \mu_1, \sigma^2) \end{cases}$$

con $\mu_1 > \mu_0$.



Consideriamo $\alpha := P(\text{rifiutare } H_0 \mid H_0 \text{ vera}) = P(\text{il nostro campione appartiene a C} \mid H_0 \text{ vera}) = P(\text{il nostro campione è } \leq c) = P(\text{commettere un errore di I tipo})$ e $\beta := P(\text{non rifiutare } H_0 \mid H_0 \text{ falsa}) = P(\text{il nostro campione appartiene a } C^c \mid H_0 \text{ falsa}) = P(\text{il nostro campione è } \geq c) = P(\text{commettere un errore di II tipo})$. (Nota: α è detto *livello di significatività del test*)
È evidente che non è possibile annullare contemporaneamente sia α che β.
La procedura si divide quindi in due passi: il primo consiste nel **fissare** α, il secondo nell'individuare la regola di decisione che minimizza β, in modo da trovare un test *ottimo*.

Lezione del 05/04, ultima modifica 20/05, Andrea Gadotti
In generale una statistica test si può descrivere come di seguito:

$$\begin{cases} H_0: & \theta \in \Theta_0 \\ H_1: & \theta \in \Theta_1 \end{cases}$$

dove Θ è lo spazio dei possibili parametri della distribuzione e $\Theta = \Theta_0 \cup \Theta_1$.
Quello che vogliamo trovare è la regola di partizionamento che divida lo spazio dei campioni C in C_0 e C_1 in funzione di α (deciso da noi). Per farlo imponiamo la condizione $\alpha = P(\underline{x} \in C \mid \theta \in \Theta_0)$. Vediamo ora un esempio con il lancio di una moneta:

Esempio Consideriamo il campione casuale $(X_1, ..., X_n)$ dove $X_i = 0$ con probabilità p e $X_i = 1$ con probabilità $1 - p$. Facciamo le nostre ipotesi:

$$\begin{cases} H_0: & p \leq 1/2 \\ H_1: & p > 1/2 \end{cases}$$

Prediamo ora come regola di decisione $\frac{S}{n} = \frac{\sum X_i}{n}$. Deciso un α a nostra discrezione, imponiamo l'equazione in k:

$$\alpha = P(S > k \mid p \leq 1/2) \; (= P(S > k \mid H_0 \text{ vera}))$$

A questo punto risolvendo l'equazione troveremo il k per il quale rifiuteremo H_0 se $S > k$.

Esempio Sia $(X_1, ..., X_n)$ un campione casuale con $X_i \sim b(1, p)$ e sia

$$\begin{cases} H_0 : & p = p_0 \\ H_1 : & p < p_0 \end{cases}$$

Sulla base dell'informazione circa p contenuta in $(X_1, ..., X_n)$ vogliamo sottoporre a verifica il sistema di ipotesi in questione. Procediamo in questo ordine:

(a) Prendiamo $S := \sum X_i \sim b(n, p)$, che è di fatto il numero di successi. Sotto H_0 abbiamo che $S \sim b(n, p_0)$.

(b) Scegliamo una regola di decisione (usando anche la distribuzione -nota- di S sotto H_0). Ovvero, individuiamo la *regione di rifiuto del test.* A questo punto vorremo rifiutare H_0 a favore di H_1 quando $S \leq k$, dove k è l'incognita che troveremo nel punto (c).

(c) Scelto il nostro α, si ha che il valore di k deve essere tale per cui

$$\alpha = P(S \leq k \mid p = p_0) = \sum_{s=0}^{k} \binom{n}{s} p_0^s (1 - p_0)^{n-s}$$

A questo punto, essendo α fissato, abbiamo un'equazione in k che risolta ci restituisce il suo valore.

Esempio particolare Nella situazione generale sopra descritta prendiamo un caso particolare con $n = 20$ e $p_0 = 0,7$. Decidiamo $\alpha = 0.15$. L'equazione diventa: $0,15 = \sum_{s=0}^{k} \binom{20}{s} 0,7^s 0,3^{20-s}$. Osserviamo che il valore di $P(S \leq k \mid p = 0,7)$ per $k = 11$ risulta $0,1133$, mentre per $k = 12$ è $0,2277$. Quindi il nostro k è compreso tra 11 e 12. In conclusione, se il nostro test dovesse presentare 12 (o più) successi, allora non rifiuteremmo H_0. In caso contrario scarteremmo H_0 a favore di H_1.

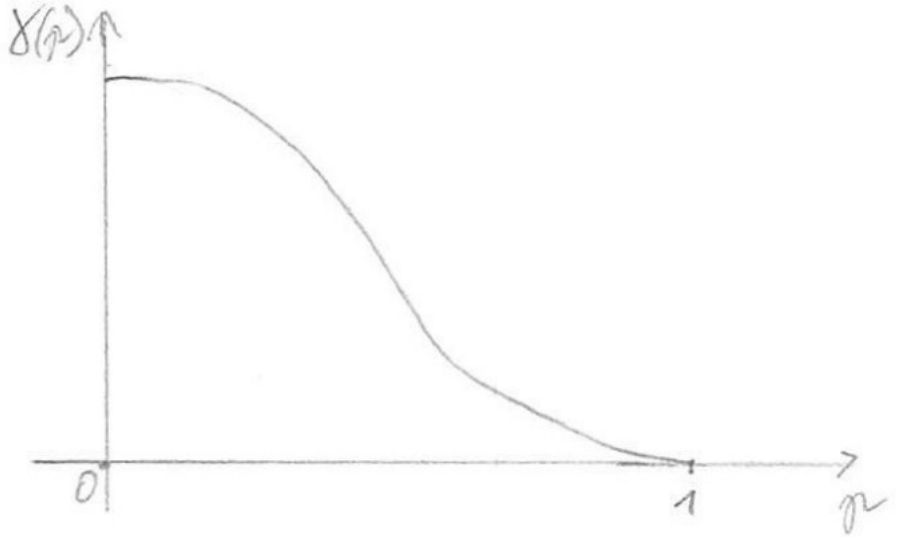
Definizione 9. *Sia $\beta := P(\underline{x} \in C_0 \mid \theta \in \Theta_1)$, ovvero β è la probabilità (fissato α) di commettere un errore di II specie. Chiamiamo potenza del test il valore $\gamma := 1 - \beta$. Un test risulta ottimale quando la sua potenza è massima. Notiamo che possiamo definire una funzione di potenza $\gamma(t) := 1 - \beta(t)$*

Osservazione 6. *Prendendo di nuovo in considerazione l'esempio precedente sul campione casuale normale, è chiaro che una volta fissato α, ovvero c, minore è μ_1, più piccola è l'area sottesa dalla coda della relativa normale, ovvero β.*

Esempio Prendendo di nuovo in considerazione l'esempio precedente sulla bernoulliana, abbiamo che

$$\gamma(p) = P(S \leq k \mid p < p_0) = \sum_{s=0}^{k} \binom{n}{s} p^s (1 - p)^{n-s}$$

Di seguito possiamo osservare il grafico della funzione:



Osservazione Il test in merito al precedente sistema di ipotesi relative a p è un *test esatto*, perché poggia sulla distribuzione *esatta* di S ($S \sim b(n,p)$). Questo però non accade sempre, e quindi talvolta è necessario ricorrere alla teoria asintotica e dei test approssimati.

1.2.4 Esempi di statistiche test (generali e particolari)

Test per la media di una popolazione qualsiasi Supponiamo di avere un campione casuale $(X_1, ..., X_n)$ proveniente da una distribuzione non nota di media μ e varianza σ^2 (finita) non note.
Le nostre ipotesi sono

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$$

Decidiamo di "condensare" l'informazione presente nel campione circa μ e σ^2 tramite $\overline{X}_n$ e S_n^2 (che ricordiamo essere stimatori non distorti e consistenti), sapendo che $\overline{X}_n \xrightarrow{P} \mu$ e $S_n^2 \xrightarrow{P} \sigma^2$.
A questo punto, la nostra regola di decisione consisterà nel rifiutare H_0 a favore di H_1 se $\overline{X}_n$ è molto più grande di μ_0.
Noi sappiamo che $\overline{X}_n \overset{a}{\sim} N\left(\mu, \frac{S_n^2}{n}\right)$, ovvero $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{D} N(0,1) =: Z$
Usando questo risultato possiamo individuare la regione critica del test di livello α fissato.
Imponiamo la seguente uguaglianza:

$$\begin{aligned} \alpha &= P(\overline{x} \in C \mid \mu = \mu_0) = P(\overline{X}_n \geq k \mid \mu = \mu_0) \\ &= P\left(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \geq \frac{\overline{k} - \mu}{S_n/\sqrt{n}} \mid \mu = \mu_0\right) \\ &= P\left(\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \geq \frac{\overline{k} - \mu_0}{S_n/\sqrt{n}}\right) = P(Z \geq z_\alpha) \end{aligned}$$

Dove z_α è il valore da cercare sulle tavole relative alla distribuzione normale in funzione dell'α scelto. Nota: l'ultima uguaglianza è in realtà un'approssimazione che è tanto più corretta quanto più grande è n.
In definitiva, abbiamo che $C = \{\underline{x} \in \varkappa \ : \ \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \geq z_\alpha\} = \{\underline{x} \in \varkappa \ : \ \overline{X}_n \geq \mu_0 + z_\alpha \frac{S}{\sqrt{n}}\}$

Possiamo anche considerare la *funzione di potenza approssimata*:

$$\begin{aligned}
\gamma(\mu) = 1 - \beta(\mu) &= P(\overline{X}_n \geq \mu_0 + z_\alpha \sigma/\sqrt{n} \mid \mu > \mu_0) \\
&= P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 + z_\alpha \sigma/\sqrt{n} - \mu}{\sigma/\sqrt{n}}\right) \\
&= 1 - P\left(Z \geq z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) \\
&= 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right)
\end{aligned}$$

dove Φ è la funzione di ripartizione di $N(0,1)$.
Notiamo che il valore di $\gamma(\mu)$ tende a 1 per $n \to \infty$. Intuitivamente, questo è esattamente ciò che ci aspettiamo, in quanto più è grande μ, più esso è distante dal nostro μ_0, e di conseguenza è lecito aspettarsi che la probabilità che un campione abbia media vicina a μ sarà bassa, ovvero la potenza del test è elevata.
È chiaro quindi che una funzione di potenza è tanto migliore quanto più il suo grafico sta vicino alla retta $y = 1$.

Esempio In riferimento al caso generale appena trattato, supponiamo di avere $\mu_0 = 12$, $\overline{X}_n = 14,3$, $S_n^2 = 22,5$, $n = 50$. Se fissiamo $\alpha = 0,05$, usando le tavole per la distribuzione normale $N(0,1)$ troviamo $z_\alpha = 1,645$. Ne segue che $k = 12 + 1,645\sqrt{22,5/50}$, che è minore di $14,3$. Concludiamo quindi rifiutando H_0.

Esempio di test esatto con t di Student Abbiamo $(X_1, ..., X_n)$ campione casuale da $N(\mu, \sigma^2)$ con μ e σ^2 non noti. Le nostre ipotesi sono:

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu > \mu_0 \end{cases}$$

Sappiamo che $\overline{X}_n \overset{H_0}{\sim} N(\mu_0, \sigma^2/n)$ e quindi:

$$T := \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \frac{1}{\sqrt{S_n^2/\sigma^2}} \sim \frac{Z}{S_n/\sqrt{n}} \sim t_{n-1}$$

(vedi pag. 17 e pag. 12)
A questo punto, possiamo trovare il nostro valore critico k usando le tavole della distribuzione t_{n-1}.
Notiamo che quello appena mostrato è un *test esatto*, in quanto non si basa su un'approssimazione dello stimatore per valori elevati di n (usando ad esempio il TLC), bensì usa la sua distribuzione reale (in questo caso la distribuzione t_{n-1}).

Lezione 08/04, ultima modifica 20/05, Michele Nardin

Esempio di test bilaterale Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione avente media μ e varianza σ^2 finite. Considerato il fatto che non abbiamo informazioni sulla distribuzione delle variabili casuali, ci appoggeremo su di un test *approssimato*. Supponiamo di voler verificare la seguente ipotesi relativa alla media della distribuzione:

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

Seguendo la falsariga di quanto visto sopra, procediamo come segue:

(a) Prendiamo $\overline{X} = \dfrac{\sum X_i}{n}$, che è stimatore della media del campione. Non conoscendo la distribuzione esatta delle variabili, consideriamo la distribuzione approssimata: sotto H_0 abbiamo che $\overline{X} \overset{a}{\sim} N(\mu_0, \sigma^2/n)$ (questo, come già detto varie volte, per il TLC).

(b) Scegliamo una regola di decisione: usando anche la distribuzione asintotica di $\overline{X}$ sotto H_0, individuiamo la *regione di rifiuto del test*. Dobbiamo trovare quindi due valori $h, k \in \mathbb{R}$ tali per cui *rifiuto* H_0 se $\overline{X_n} \leq h$ o $\overline{X_n} \geq k$.

(c) Scelto il livello di confidenza α, si ha che h, k vanno scelti in modo tale per cui

$$\alpha = P(\overline{X} \leq h \vee \overline{X} \geq k \mid \mu = \mu_0)$$

Grazie alla normalità della distribuzione asintotica, possiamo supporre che la distribuzione di $\overline{X_n}$ sia simmetrica, per lo meno da un certo n in poi (ricordiamo che, per campioni con numerosità superiori a 30, l'approssimazione è già buona).

Questo ci permette di scrivere

$$\frac{\alpha}{2} = P(\overline{X} \leq h \mid \mu = \mu_0) = P(\overline{X} \geq k \mid \mu = \mu_0)$$

Dato che

$$\frac{\overline{X} - \mu_0}{S_n/\sqrt{n}} \xrightarrow{D} N(0,1)$$

denotato con $z_{\alpha/2}$ il quantile di ordine $\alpha/2$ della normale standard, la regione critica (di rifiuto) sarà

$$C = \left\{ \underline{x} = (x_1, ..., x_n) \in X : \left| \frac{\overline{X} - \mu_0}{S_n/\sqrt{n}} \right| \geq z_{\alpha/2} \right\}$$

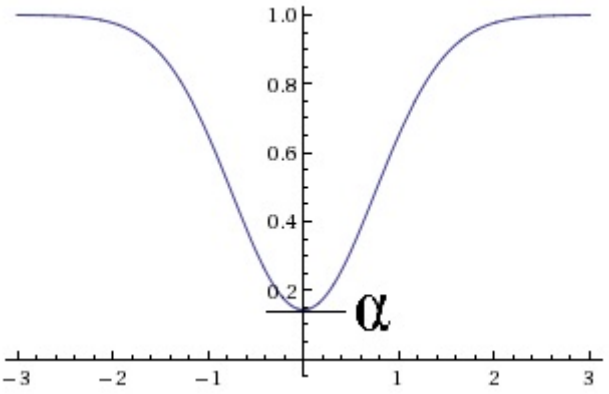
da cui

$$k = \frac{S_n}{\sqrt{n}} z_{\alpha/2} + \mu_0 (= -h)$$

Consideriamo inoltre la funzione di potenza associata, sempre sfruttando l'approssimazione normale ed il fatto che, per n grande, $S_n \approx \sigma$:

$$\gamma(\mu) = P(\overline{X} \leq \mu_0 - \frac{S_n}{\sqrt{n}} z_{\alpha/2} \mid \mu \neq \mu_0) + P(\overline{X} \geq \mu_0 + \frac{S_n}{\sqrt{n}} z_{\alpha/2} \mid \mu \neq \mu_0)$$
$$\approx \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} - z_{\alpha/2}\right) + \left[1 - \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} - z_{\alpha/2}\right)\right]$$

Di seguito possiamo osservare il grafico della funzione:



Osservazione: i test costruiti usando t di Student sono più conservativi rispetto a quelli costruiti con approssimazione Normale, nel senso che è più facile che un test venga accettato usando la prima rispetto alla seconda. Questo è dato dal fatto che la distribuzione t di student ha le code più pesanti rispetto alla normale! (in particolare fissato un $\alpha \in (0,1)$, i quantili della normale standard $z_{\alpha/2}$ sono più piccoli dei quantili della t di Student con $n-1$ gradi di libertà $t_{\alpha/2;n-1}$, cioè $|z_{\alpha/2}| < |t_{\alpha/2;n-1}|$)

Lezione del 12/04, ultima modifica 20/05, Andrea Gadotti

Esempio di test t-Student per due campioni

Campione $(X_1, ..., X_{n_1})$ da $N(\mu_1, \sigma^2)$ e $(Y_1, ..., Y_{n_2})$ da $N(\mu_2, \sigma^2)$ indipendenti. La varianza σ è la stessa ma non è nota.
Supponiamo di avere elementi per pensare che:

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

Abbiamo che $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1/n_1 + \sigma_2/n_2)$.
Prendiamo ora

$$S_p^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n - 2}$$

dove $n = n_1 + n_2$.
Abbiamo che:

$$T := \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \overset{H_0}{\sim} t_{n-2}$$

dove abbiamo usato anche il fatto che S_p^2 è una chi-quadro con $n-2$ gradi di libertà (vedi pag. 19).
In conclusione, avremo che la nostra regione critica sarà $C = \{(\underline{x}, \underline{y}) \mid T \geq t_{n-2;\alpha}\}$.

Esempio con bernoulliana Abbiamo $(X_1, ..., X_n)$ campione casuale da $b(1, p)$ (ricordiamo: media p, varianza $p(1-p)$). Le nostre ipotesi sono:

$$\begin{cases} H_0: & p = p_0 \\ H_1: & p = p_1 \end{cases}$$

con $p_1 < p_0$. Consideriamo

$$\hat{p}_n := \frac{\sum X_i}{n} \overset{a}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Abbiamo quindi:

$$Z := \frac{\hat{p}_n - p_0}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}$$

A questo punto imponiamo: $\alpha = P(Z \leq z_\alpha \mid H_0)$.

Esempio di test sulla varianza Campione casuale $(X_1, ..., X_n)$ da $N(\mu, \sigma^2)$, μ e σ^2 non noti. Le nostre ipotesi sono:

$$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 = \sigma_1^2 \end{cases}$$

con $\sigma_1^2 > \sigma_0^2$. Notiamo che $\frac{n-1}{\sigma_0^2} S_n^2 \overset{H_0}{\sim} \chi_{n-1}^2 =: W$. (Nota: come in molti esempi precedenti, il fatto che sia intelligente tirare fuori queste osservazioni che portano all'analisi di distribuzioni conosciute è calato dall'alto, almeno per ora).
Imponiamo:

$$\alpha = P\left(\frac{n-1}{\sigma^2} S_n^2 \mid \sigma^2 = \sigma_0^2\right) = P\left(\frac{n-1}{\sigma_0^2} S_n^2\right)$$

Quindi $k = w_{\alpha;n-1}$. In conclusione, rifiuto H_0 se $W \geq w_{\alpha;n-1}$.
Esempio In riferimento al caso generale appena trattato, supponiamo di avere $n = 25$, $\sigma_0^2 = 15$, $\sigma_1^2 = 20$, $s_n^2 = 17,4$ e $\alpha = 0,05$. Allora:

$$k = w_{0,05;(25-1)} = w_{0,05;24} = 36,415 > 27,84 = \frac{25-1}{15} 17,4 = \frac{n-1}{\sigma_0^2} s_n^2 = w$$

In conclusione, non rifiutiamo H_0.

Esempio con due campioni normali Campione $(X_1, ..., X_{n_1})$ da $N(\mu_1, \sigma_1^2)$ e $(Y_1, ..., Y_{n_2})$ da $N(\mu_2, \sigma_2^2)$ indipendenti. μ_i e σ_i^2 non noti. Le nostre ipotesi sono:

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Nota: l'ipotesi dell'uguaglianza delle due varianze prende il nome di omoschedasticità.
Sappiamo che $S_i^2 \xrightarrow{P} \sigma_i^2$. Inoltre

$$W := \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} = \frac{\frac{(n_2-1)S_2^2}{\sigma_2^2}\frac{1}{n_2-1}}{\frac{(n_1-1)S_1^2}{\sigma_1^2}\frac{1}{n_1-1}} \sim F_{(n_2-1),(n_1-1)}$$

(vedi pag. 20)
Sotto H_0 abbiamo chiaramente che $W := \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} = \frac{S_2^2}{S_1^2}$. La nostra regola di decisione consisterà nel rifiutare H_0 a favore di H_1 se $\frac{S_2^2}{S_1^2}$ è "lontano" da 1, ovvero se $W < k_1$ o $W > k_2$, dove k_1 e k_2 dipendono dalla distribuzione di $W = \frac{S_2^2}{S_1^2}$ e dal valore di α. Perciò, dividendo equamente in due parti la probabilità di errore, le due equazioni risultano:

$$\alpha/2 = P(W > k_2 \mid \sigma_1^2 = \sigma_2^2) \quad \text{e} \quad 1 - \alpha/2 = P(W < k_1 \mid \sigma_1^2 = \sigma_2^2)$$

In conclusione,

$$C = \{(\underline{x_1}, \underline{x_2}) \ : \ \frac{S_2^2}{S_1^2} < w_{(n_2-1),(n_1-1);\alpha/2} \quad \text{o} \quad \frac{S_2^2}{S_1^2} > w_{(n_2-1),(n_1-1);1-\alpha/2}$$

In riferimento al caso generale appena trattato, supponiamo di avere $n_1 = 14$, $n_2 = 10$ $s_1^2 = 17,4$, $\sigma_1^2 = 20$, $s_2^2 = 37,9$ e $\alpha = 0,05$.
Come nel caso generale, diciamo $W := S_2^2/S_1^2 \sim F_{(n_2-1),(n_1-1)}$.
Abbiamo che $w_{(10-1),(14-1);0,025} = 3,31$ e $w_{(10-1),(14-1);0,975} = 1/w_{13,19;0,025} = 1/3,76 = 0,26$.
Poiché $s_2^2/s_1^2 = 37,9/17,4 = 2,178$, decidiamo di non rifiutare H_0.

Capitolo 2

Seconda parte del corso