Indice

1 L'inferenza statistica

1.1 Introduzione ai test di pura significatività

Continua in questo quaderno lo studio del rapporto esistente tra variabili statistiche (campionarie) e variabili casuali (v.c.): intendiamo qui affrontare il problema della plausibilità empirica di ipotesi a priori fatte su una v.c. X.

Il tipo di problemi che vogliamo affrontare consiste nel cercare di rispondere alla domanda se, fatta una qualche ipotesi a priori H_0 sulla v.c. X, da cui il campione è estratto, non vi sia nei dati evidenza che H_0 è probabilmente falsa, ovvero non plausibile.

Vediamo qualche esempio.

Esempio 1.1.1: per mantenere sotto controllo la stabilità di un manufatto si misura in tempi diversi la distanza tra due suoi punti; ad ogni epoca la distanza è misurata più volte per evitare errori grossolani e per migliorare la stima del valore medio $D(t)$; si hanno così a disposizione varie medie campionarie $\overline{D}(t_1), \overline{D}(t_2), \ldots$ che di solito sono tra loro differenti; ci si chiede se i valori empirici $\overline{D}(t_i) - \overline{D}(t_k)$ siano tali da rendere non plausibile l'ipotesi H_0 che per i valori teorici valga $\overline{D}(t_1) = \overline{D}(t_2) = \ldots$, cioè il manufatto sia rimasto stabile.

Esempio 1.1.2: un produttore produce dei pezzi dei quali misura il valore di una certa caratteristica C. In base a considerazioni di tolleranza nell'impiego di quei pezzi, il produttore considera regolare il processo di produzione se i valori della caratteristica C sono estratti da una v.c. con media c e con s.q.m. σ_c.

Per verificare se la produzione è regolare (H_0), il produttore esamina un campione $C_1, C_2, \ldots C_N$ della produzione e calcola la media campionaria m_c e lo s.q.m. campionario $\overline{s}_c$ e si chiede se questi dati empirici mettano in evidenza la falsità di H_0, oppure se essa è plausibile in base ai dati.

Esempio 1.1.3: un economista esamina lo sviluppo di due grandi città,

A e B, di diverse regioni e fa l'ipotesi, (H_0), che tali città abbiano seguito un analogo sviluppo produttivo. Suddividendo le attività produttive in settori $i = 1, \ldots, N$, esamina ad esempio la percentuale di addetti in ogni città per tali settori $a_{A1}, a_{A2}, \ldots, a_{AN}; a_{B1}, a_{B2}, \ldots, a_{BN} (\sum a_{Ai} = = 1, \sum a_{Bi} = 1)$. Poiché molti ed incontrollabili sono i fattori che influenzano ognuna di tali percentuali, si può ipotizzare (H_0) che esse siano estrazioni a coppie dalle stesse variabili casuali
$\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_N (\sum \mathcal{A}_i = 1)$, cioè $a_{A1}, a_{B1} \sim \mathcal{A}_1; a_{A2}, a_{B2} \sim \mathcal{A}_2$; ecc. Ci si chiede se l'ipotesi che le due distribuzioni $\{a_{Ai}\}, \{a_{Bi}\}$ siano tra loro uguali sia confermata plausibilmente dai dati o se questi forniscano l'evidenza della falsità di H_0.

Il punto di vista specifico che assumeremo nella prima parte di questo quaderno è quello dei puri test di significatività, cioè una volta specificata H_0 si cercherà di costruire una statistica $S(\underline{X}^{(N)})$, la cui distribuzione sarà per ipotesi nota se H_0 è verificata e che tenda ad assumere valori "grandi" quando intuitivamente si ritiene che ci si allontani da H_0. Così si creeranno delle regioni critiche dello spazio campionario R^N, quelle cioè dove $S \geq c$ convenientemente grande, in modo che quando il valore estratto dal campione $\underline{x}$ appartiene a tale zona, cioè quando $S(\underline{x}) \geq c$ si consideri confutata, cioè poco plausibile, H_0. Naturalmente in questo modo si corre il rischio di rifiutare H_0 anche in casi in cui essa è vera, e precisamente tale rischio sarà dato da

$$P\{S \geq c | H_0\} = \alpha \ ; \tag{1.1.1}$$

α viene chiamato il livello di significatività del test. è bene notare che scrivendo $P\{S \geq c|H_0\}$ si è voluto sottolineare che questa probabilità va calcolata in base alla distribuzione di S, nota se si accetta l'ipotesi H_0; viceversa non si suppone di conoscere la distribuzione di S sotto ipotesi alternative H_A, punto di vista che assumeremo più avanti.

La (1.1.1) può essere usata in tre diversi modi:

a) fissato il valore critico c, si calcola il corrispondente livello di significatività

$$\alpha = 1 - F_S(c) \tag{1.1.2}$$

(F_S funzione di distribuzione di S nota in base ad H_0);

b) fissato un livello di significatività α (i valori più usati nei test sono 1%, 5%) si calcola il corrispondente valore critico invertendo la (1.1.2);

c) calcolato il valore osservato s di S relativo ad una certa estrazione $\underline{x}, s = S(\underline{x})$, calcola in corrispondenza il *livello di significatività osservato*

$$\alpha_0 = 1 - F_S(s) \ . \tag{1.1.3}$$

Si può notare che così si avrà una α_0 per ogni campione dato, cioè α_0 viene ad essere una estrazione dalla statistica (v.c.)

$$A = 1 - F_S[S(\underline{X}^{(N)})] \ . \tag{1.1.4}$$

La (1.1.4) è ovviamente una variabile uniforme sull'intervallo [0,1], se H_0 è vera.

1.2 Test parametrici, ipotesi semplici, ipotesi composte; intervalli fiduciari

Il caso più comune di applicazione della teoria dei test è quello in cui l'ipotesi H_0 consiste nell'affermare che la v.c. X ha distribuzione $f(x;\theta)$, dipendente da un parametro θ, e che θ assume un certo valore θ_0: diciamo allora che si ha una *ipotesi semplice* H_0.

In tal caso sembra naturale assumere come statistica S una funzione di θ e di un suo stimatore corretto $T(\underline{X})$; ad esempio

$$S(\underline{X}) = |T(\underline{X}) - \theta| \tag{1.2.1}$$

oppure

$$S(\underline{X}) = \left| \frac{T(\underline{X})}{\theta} - 1 \right| \ , \tag{1.2.2}$$

essendo chiaro che in entrambi i casi un alto valore di S può essere preso come indicazione della non accettabilità di H_0.

Esempio 1.2.1: da precedenti analisi ci si è convinti che in condizioni standard una linea telefonica ha una distribuzione dell'intervallo tra una telefonata e la successiva di tipo esponenziale
$f(x;\theta) = (1/\theta)e^{-x/\theta}, (x \geq 0)$; inoltre i rilevamenti per lunghi periodi hanno dato per θ un certo valore θ_0.

In un giorno a caso si decide di esaminare un campione di numerosità N della v.c. X, per verificare se (H_0) il modello è mutato ed in particolare se il traffico medio della linea, misurato da θ, sia variato. Poiché θ è la media di X, e poiché tale variabile è definita solo per valori positivi, si può pensare ad usare la statistica

$$S = \left| \frac{\mathcal{M}}{\theta} - 1 \right| \ .$$

In effetti se H_0 è vera, X/θ_0 è una esponenziale di media 1, cioè $2X/\theta_0 = = \chi^2_{(2)} = 2\Gamma(k = 1,\ \rho = 1)$, così che si può porre

$$2N\frac{\mathcal{M}}{\theta_0} = \chi^2_{(2N)} = 2\Gamma(k = N,\ \rho = 1) \ ,$$

cioè la distribuzione di S, sotto l'ipotesi H_0, è nota. Pertanto, fissato un livello di significatività α, si potrà trovare una regione critica dalla relazione

$$P\{|2N\frac{\mathcal{M}}{\theta_0} - 2N| \geq c|H_0\} = P\{|\chi^2_{(2N)} - 2N| \geq c\} = \alpha \ .$$

Spesso anziché una regione critica di questo tipo, cioè simmetrica attorno a $2N$, si preferisce, per l'intrinseca asimmetria della distribuzione di χ^2, che è diversa da zero solo sul semiasse positivo, una regione definita da

$$\frac{2N\mathcal{M}}{\theta_0} \leq c_1 \ ; \ \frac{2N\mathcal{M}}{\theta_0} \geq c_2$$

con

$$P\{\chi^2_{(2N)} \leq c_1\} = P\{\chi^2_{(2N)} \geq c_2\} = \alpha/2 \ . \qquad (1.2.3)$$

Osservazione 1.2.1: si noti che è possibile pensare a test per una ipotesi semplice anche quando θ è un vettore di parametri anziché una singola variabile. In tal caso l'ipotesi H_0 dovrà specificare θ_0, cioè tutte le componenti di θ. Così nell'Esempio 1.1.2 si aveva $X = \mathcal{N}[c, \sigma_c^2]$ con $\theta = (c, \sigma_c^2)$, e l'ipotesi H_0 consisteva nello specificare $\theta_0 = (c_0, \sigma^2_{co})$. In un caso come questo naturalmente anche $T(\underline{X})$ dovrà essere un vettore di stimatori di θ, cioè

$$T(\underline{X}) = (\mathcal{M}, \overline{\mathcal{S}}^2) \ .$$

È possibile dimostrare che per questo caso una statistica utile è

$$S(\underline{X}) = N \left| \frac{\mathcal{M} - c_0}{\sigma_{c_0}} \right|^2 + (N-1)\frac{\overline{\mathcal{S}}^2}{\sigma^2_{c_0}} \ ,$$

che risulta essere una $\chi^2_{(N)}$, a N gradi di libertà.

Tuttavia un approccio di questo tipo è poco usato in genere perché, se sulla base dei valori campionari e del livello di significatività α si decidesse di rifiutare H_0, non sarebbe chiaro se ciò sia dovuto a una o all'altra componente di θ: nel caso dell'Esempio 1.1.2, se sia la media campionaria $\mathcal{M}$ a non andare d'accordo con c_0, oppure $\overline{\mathcal{S}}^2$ con σ^2_{co}. Si preferirebbe allora cercare di costruire test separati per ogni componente, ciò che ci porta direttamente al problema delle ipotesi composte.

Prendiamo ora in esame il caso in cui la distribuzione $f(x; \theta, \lambda)$ dipenda da più parametri (θ, λ), ma si voglia sottoporre a test un'ipotesi H_0 solo su $\theta : H_0(\theta = \theta_0)$. Si dice allora che λ *è un parametro di disturbo (nuisance parameter) e l'ipotesi* H_0 *è chiamata composta.*

Si noti che quanto segue vale sia che λ sia monodimensionale o a più dimensioni, cioè se vi siano uno o più di uno parametri di disturbo.

Per trattare questo problema, cerchiamo in primo luogo degli stimatori corretti T, L rispettivamente di θ e λ; questi avranno una distribuzione

$f(t, l; \theta, \lambda)$ che dipende tanto da $\theta = E\{T\}$, che da $\lambda = E\{L\}$. Occorre allora cercare una funzione $S(T, L)$ la cui distribuzione sia indipendente da $\lambda, f(s, \theta)$: quando questa sia stata trovata, mediante tale statistica sarà possibile sottoporre a test l'ipotesi $\theta = \theta_0$, ricercando delle regioni critiche per una qualche funzione di S.

Naturalmente questo procedimento è più complesso e va visto caso per caso.

Esempio 1.2.2: si supponga che sia $X = \mathcal{N}[\mu, \sigma^2]$, e si voglia verificare l'ipotesi $H_0, \mu = \mu_0$: si dovrà allora esaminare la distribuzione congiunta di $(\mathcal{M}; \mathcal{S}^2)$ e cercare una statistica funzione di tali variabili, possibilmente funzione di μ ma non di σ^2, che sia anche distribuita indipendentemente da σ^2.

Riesaminando il procedimento teorico che ci porta a disegnare un test parametrico per l'ipotesi $H_0(\theta = \theta_0)$, si vede che in sostanza occorre definire una certa funzione $S(\underline{X}; \theta)$ che sotto l'ipotesi H_0 ha una distribuzione nota

$$S(\underline{X}; \theta_0) = Y \ , \ \text{(v.c. nota)} \tag{1.2.4}$$

così che fissato il livello di significatività α ed il valore y_α per cui

$$P\{Y \geq y_\alpha\} = \alpha \ ,$$

sia definita una regione critica dello spazio campionario, per cui se risulta

$$s = S(\underline{x}; \theta_0) \geq y_\alpha \ , \tag{1.2.5}$$

H_0 viene rifiutata, al livello α.

Notiamo che spesso accade che Y risulti indipendente da θ, anche se ciò non è essenziale in quanto segue.

Notiamo anche che se, per un certo θ_0 e per un campione dato $\underline{x}, H_0$ non è rigettata, cioè se

$$S(\underline{x}, \theta_0) < y_\alpha \ , \tag{1.2.6}$$

viene spontaneo osservare che, qualora avessimo fatto una differente ipotesi $\theta = \theta_1$ abbastanza vicina a θ_0, e posto che tutte le funzioni in gioco siano continue, si avrebbe ancora, per lo stesso campione $\underline{x}$,

$$S(\underline{x}, \theta_1) < y_\alpha \ . \tag{1.2.7}$$

È bene osservare che in caso Y dipenda da θ, il valore y_α in (1.2.7) sarà differente da quello in (1.2.6). Ci si può allora chiedere quali siano tutti i valori di θ, per cui un test basato su $S(\underline{x}; \theta)$, sul vettore campionario $\underline{x}$ e sul livello di significatività α, darebbe risposta positiva (di accettazione), cioè tutti i θ tali per cui

$$S(\underline{x}, \theta) < y_\alpha \tag{1.2.8}$$

sia verificata. L'intervallo di valori di θ per cui la (1.2.8) è verificata, si chiama *intervallo fiduciario al livello di significatività* α.

Osservazione 1.2.2: si noti che fornire un intervallo fiduciario per θ equivale, seppure in un'accezione diversa da quella del Quaderno n. 2, a dare una stima di θ; infatti la teoria che fissa i criteri con cui fornire gli intervalli fiduciari è chiamata anche teoria della stima per intervalli.

Esempio 1.2.3: sia $X = \mathcal{N}[\mu, 1]$ e sia $\underline{x}$ un vettore di N valori campionari estratti da X; si supponga di voler verificare l'ipotesi $H_0 : \mu = \mu_0$ al livello di significatività α. Si decida inoltre di usare la statistica

$$S = |\mathcal{M} - \mu| \tag{1.2.9}$$

per la verifica di tale ipotesi.

Se H_0 è vera, $\mathcal{M} = \mathcal{N}[\mu; 1/N]$, così che $S(\underline{x}, \mu)$ è sostanzialmente una "seminormale" (cfr. fig. 1.2.1), cosa che potrebbe anche scriversi nella forma

$$\sqrt{N}S = \frac{|\mathcal{M} - \mu|}{1/\sqrt{N}} = |Z| \ . \tag{1.2.10}$$

Dunque le regioni critiche di $\underline{X}$ saranno definite dalla relazione

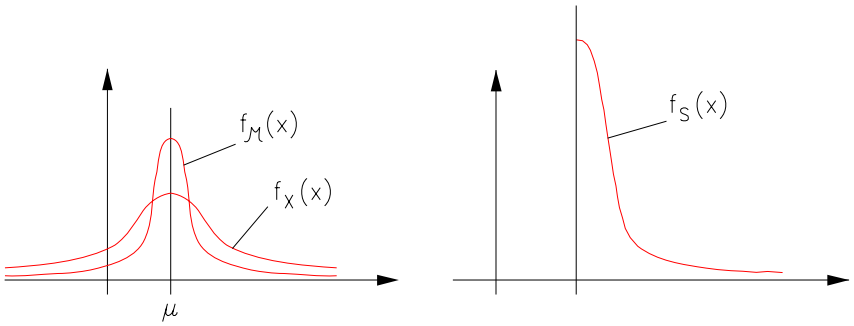


Figura 1.2.1:

$$S(\underline{x};\theta) \geq Z_{\alpha/2}$$
$$P\{|Z| \geq Z_{\alpha/2}\} = P\{Z \leq -Z_{\alpha/2}\} + P\{Z \geq Z_{\alpha/2}\} = \alpha$$
$$P\{Z \geq Z_{\alpha/2}\} = \alpha/2$$

cioè

$$|\mathcal{M}(\underline{x}) - \mu| \geq \frac{Z_{\alpha/2}}{\sqrt{N}} \ . \tag{1.2.11}$$

Pertanto se H_0 (cioè $\mu = \mu_0$) è vera, così che $|\mathcal{M} - \mu_0| = (1/\sqrt{N})|Z|$, e se il valore empirico della media è

$$m = \mathcal{M}(\underline{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i \ , \tag{1.2.12}$$

se vale

$$|m - \mu_0| \geq \frac{Z_{\alpha/2}}{\sqrt{N}} \ , \tag{1.2.13}$$

allora H_0 viene rigettata, in caso contrario, cioè se

$$|m - \mu_0| < \frac{Z_{\alpha/2}}{\sqrt{N}} \ ,$$

H_0 viene accettata.

Notiamo che fissato $\underline{x}$ è fissata anche la media empirica m (1.2.12), perciò anche considerando un'altra ipotesi $H_1(\mu = \mu_1)$ si può decidere sull'accettazione o meno di tale ipotesi in base al valore di $|m-\mu_1|$ confrontato con $Z_{\alpha/2}/\sqrt{N}$: ne segue che l'intervallo fiduciario, di livello α, corrispondente alla media empirica m, è l'insieme dei valori μ per cui

$$|m - \mu| \leq \frac{Z_{\alpha/2}}{\sqrt{N}} \ .$$

Osservazione 1.2.3: vogliamo notare che nel caso di una variabile X discreta, in generale anche S sarà discreta e quindi non si potrà fissare un livello di significatività arbitrario in quanto, posto

$$\alpha_i = P\{S \geq s_i\} \ , \tag{1.2.14}$$

anche i valori α_i saranno un insieme discreto.

Tuttavia fissato un α si potrà trovare un $s_i(\alpha)$ tale che la corrispondente significatività del test sia almeno pari ad α, cioè

$$P\{S \geq s_i(\alpha)\} = \alpha_i \qquad \alpha_i = \min_j\{\alpha_j \geq \alpha\} \ . \tag{1.2.15}$$

Esempio 1.2.4: sia X una variabile binaria

$$X = \begin{cases} 0 & 1 \\ p & q \end{cases} \ ;$$

sia

$$\underline{x}^N = \left| \begin{array}{c} x_1 \\ \vdots \\ x_N \end{array} \right|$$

un campione estratto da X, e si voglia sottoporre a test l'ipotesi $H_0 : p = p_0$ $(q = q_0 = 1 - p_0)$.

Notando che $p = E\{X\}$, si può pensare ad una statistica funzione della media campionaria.

Più precisamente, useremo $N\mathcal{M} = \sum_{i=1}^{N} X_i = \mathcal{K}$, notando che $\mathcal{K}$ ha una distribuzione binomiale

$$\mathcal{K} = B(N, p) \ ;$$

essendo N noto e fissando $p = p_0$ per H_0, la distribuzione di $\mathcal{K}$ risulterà nota.

Resta il problema di fissare S. In effetti se $p_0 = 1/2$, $\mathcal{K}$ ha una distribuzione simmetrica, mentre per $p_0 \neq 1/2$ la distribuzione diventa asimmetrica (cfr. fig. 1.2.2). Pertanto un criterio del tipo (cfr. (1.2.15))

$$P\{Np - k \leq K \leq Np + k\} \leq 1 - \alpha$$

va bene per $p_0 = 1/2$ o valori vicini (soprattutto quando N è elevato). Per valori di p_0 che danno una notevole asimmetria, si potrebbe definire un intervallo di accettazione, $k_1 \leq \mathcal{K} \leq k_2$, con

$$\begin{array}{lll} k_1 = \sup k & , & P(\mathcal{K} \leq k) \leq \alpha/2 \\ k_2 = \inf k & , & P(k \leq \mathcal{K}) \geq \alpha/2 \end{array} \ .$$

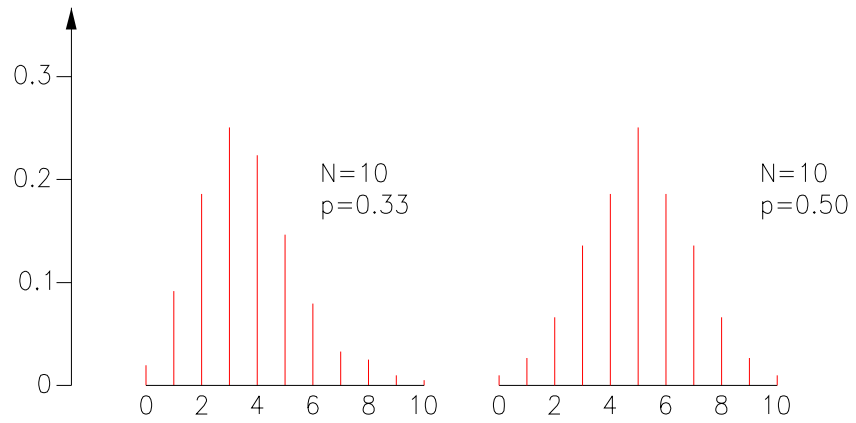


Figura 1.2.2: Variabili binomiali

1.3 Test sulle medie di campioni numerosi

In questo paragrafo consideriamo il problema di applicare test a ipotesi sulla media μ, mediante l'osservazione di campioni numerosi. La numerosità del campione ci permette di conoscere la distribuzione asintotica della media campionaria, indipendentemente dalla distribuzione di partenza.

Ai punti a) e b) considereremo i test su medie e differenze di medie, quando le varianze in gioco siano note a priori (H_0 semplice). Il caso in cui la varianza sia un parametro di disturbo (H_0 composta) è trattato al punto c).

a) Si supponga che X sia una v.c. con media μ incognita e s.q.m. σ noto. Consideriamo un campione bernoulliano $\underline{x}^{(N)}$ tratto da X con N abbastanza elevato; vogliamo sottoporre a verifica l'ipotesi semplice $H_0 : \mu = \mu_0$ sulla base del vettore di dati sperimentali $\underline{x}$. La statistica naturale in questo caso è una funzione di $\mathcal{M} = (1/N)\sum X_i$; se N è elevato si può applicare il teorema centrale della statistica e supporre che

$$\mathcal{M} \sim \mathcal{N}\left[\mu; \frac{\sigma^2}{N}\right] \ . \tag{1.3.1}$$

Pertanto se H_0 è vera, vale la relazione

$$\frac{\mathcal{M} - \mu_0}{\sigma/\sqrt{N}} \sim Z \ , \tag{1.3.2}$$

così che volendo assumere come statistica base del test

$$S = |\mathcal{M} - \mu| \ ,$$

si ha che l'intervallo di accettazione di H_0, al livello di significatività α, è appunto dato da quei valori empirici di $m = 1/N \sum_i x_i$ per cui

$$\frac{|m - \mu_0|}{\sigma/\sqrt{N}} \leq Z_{\alpha/2} \quad , \quad (P(Z \geq Z_{\alpha/2}) = \alpha/2) \ , \tag{1.3.3}$$

ovvero

$$\mu_0 - \frac{\sigma}{\sqrt{N}} Z_{\alpha/2} \leq m \leq \mu_0 + \frac{\sigma}{\sqrt{N}} Z_{\alpha/2} \ .$$

Corrispondentemente, fissato m, l'intervallo fiduciario al livello α, I_α, è dato dall'insieme dei μ per cui

$$|m - \mu| \leq \frac{\sigma}{\sqrt{N}} Z_{\alpha/2} \ ,$$

cioè

$$m - \frac{\sigma}{\sqrt{N}} Z_{\alpha/2} \leq \mu \leq m + \frac{\sigma}{\sqrt{N}} Z_{\alpha/2} \ . \tag{1.3.4}$$

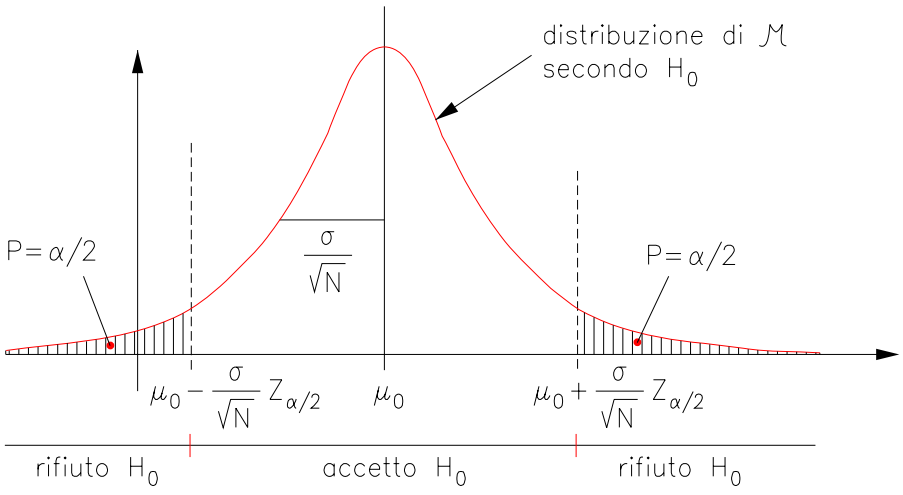


Figura 1.3.1:

Osservazione 1.3.1: si noti che, a parità di α, tanto l'intervallo di accettazione, quanto l'intervallo fiduciario, si restringono all'aumentare di N, volendo ciò dire che, con l'aumentare dell'informazione, una media empirica m che ha una distanza fissata da μ_0 diventa sempre meno probabile e finisce prima o poi per diventare un elemento di evidenza contro H_0. È questo un diverso modo di esprimere la consistenza dello stimatore $\mathcal{M}$.

b) Prendiamo in considerazione il caso in cui vi siano due v.c. indipendenti X ed Y con medie rispettivamente μ_X e μ_Y incognite e con varianze σ_X^2 e σ_Y^2 note. Si estraggono due campioni indipendenti dalle due variabili, con numerosità N_X ed N_Y elevate e si vuole sottoporre a test l'ipotesi $H_0 : \mu_X = \mu_Y$.

La statistica naturale da cui partire è

$$\mathcal{M}_X - \mathcal{M}_Y = \frac{1}{N_X} \sum X_i - \frac{1}{N_Y} \sum Y_i \ , \tag{1.3.5}$$

così che se vale H_0, indipendentemente dalle distribuzioni originarie di

X ed Y, applicando il teorema centrale della statistica si può scrivere $(N_X, N_Y \to \infty)$

$$\mathcal{M}_X - \mathcal{M}_Y \sim \mathcal{N}\left[0, \frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}\right] \ : \tag{1.3.6}$$

come si vede, sotto l'ipotesi H_0, la distribuzione di $\mathcal{M}_X - \mathcal{M}_Y$ è completamente specificata e dunque è possibile costruire dei test. In particolare, si può porre la (1.3.6) nella forma equivalente

$$\frac{\mathcal{M}_X - \mathcal{M}_Y}{\sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}}} = Z \ , \tag{1.3.7}$$

e derivare la regione critica per la statistica

$$S = \frac{|\mathcal{M}_X - \mathcal{M}_Y|}{\sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}}} \ ,$$

ovvero definire l'intervallo di accettazione di H_0 al livello di significatività α, come quello in cui le due medie empiriche m_X, m_Y soddisfano la relazione

$$-\sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}} Z_{\alpha/2} \leq m_X - m_Y \leq \sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}} Z_{\alpha/2} \ . \tag{1.3.8}$$

Quando la (1.3.8) non è verificata H_0 è rifiutata al livello di significatività α.

Osservazione 1.3.2: anziché verificare l'ipotesi $H_0 : \mu_X = \mu_Y$, può capitare di dover verificare l'ipotesi $H_0 : \mu_X - \mu_Y = \mu_0$. Modificando la (1.3.6) nella forma

$$\mathcal{M}_X - \mathcal{M}_Y \sim \mathcal{N}\left[\mu_0, \frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}\right] \ ,$$

si arriva ad un intervallo di accettazione del tipo (1.3.8) in cui a $m_X - m_Y$ va sostituita l'espressione $m_X - m_Y - \mu_0$. Naturalmente la stessa espressione può essere usata eventualmente per trovare un intervallo fiduciario per μ.

c) Consideriamo ora il caso in cui per un campione numeroso si vuol verificare l'ipotesi $H_0 : \mu = \mu_0$, senza conoscerne a priori la varianza σ^2. Si potrebbe pensare semplicemente di sostituire nella (1.3.2) $\overline{\mathcal{S}^2}$ a σ^2, cioè di porre

$$\frac{\mathcal{M} - \mu}{\frac{\overline{\mathcal{S}}}{\sqrt{N}}} = \frac{1}{\sqrt{N}} \frac{\sum (X_i - \mu)}{\overline{\mathcal{S}}} \sim Z \ . \tag{1.3.9}$$

La (1.3.9) tuttavia non è di immediata derivazione dal teorema centrale della statistica perché le variabili $(X_i - \mu)/\overline{\mathcal{S}}$, se sono identicamente distribuite, non sono però indipendenti, in quanto tutte funzioni della stessa statistica $\overline{\mathcal{S}}$. D'altro canto possiamo scrivere

$$S_N = \frac{\mathcal{M} - \mu}{\frac{\overline{\mathcal{S}}}{\sqrt{N}}} = \frac{\mathcal{M} - \mu}{\frac{\sigma}{\sqrt{N}}} \cdot \frac{\sigma}{\overline{\mathcal{S}}} \ ; \tag{1.3.10}$$

vogliamo dimostrare che anche S_N tende in legge ad una normale standardizzata.

Notiamo infatti che, posto

$$\Gamma_N = \frac{\mathcal{M}_N - \mu}{\sigma/\sqrt{N}} \ , \ \Delta_N = \frac{\overline{S}_N}{\sigma} \ , \ \left(S_N = \frac{\Gamma_N}{\Delta_N} \right) \ ,$$

si ha per $N > N_\varepsilon$ opportuno

$$P_N\{|\Delta_N - 1| \leq \varepsilon\} \geq 1 - \varepsilon \ ,$$

per ε fisso, ma arbitrario.

D'altro canto, presi $a, b > 0$,

$$\begin{array}{rl} & P_N\{a \le S_N \le b\} = P_N\{a\Delta_N \le \Gamma_N \le b\Delta_N\} = \\ = & P_N\{[a\Delta_N \le \Gamma_N \le b\Delta_N] \cap [|\Delta_N - 1| \le \varepsilon]\} + \\ + & P_N\{[a\Delta_N \le \Gamma_N \le b\Delta_N] \cap [|\Delta_N - 1| > \varepsilon]\} \le \\ \le & P_N\{a(1-\varepsilon) \le \Gamma_N \le b(1+\varepsilon)\} + \varepsilon \end{array} \qquad (1.3.11)$$

Poiché, per $N \to \infty$

$$P_N\{a(1-\varepsilon) \le \Gamma_N \le b(1+\varepsilon)\} \to P\{a(1-\varepsilon) \le Z \le b(1+\varepsilon)\}$$

per N'_ε opportuno sarà

$$P_N\{a \le S_N \le b\} \le P\{a(1-\varepsilon) \le Z \le b(1+\varepsilon)\} + 2\varepsilon \ . \qquad (1.3.12)$$

D'altra parte, per la (1.3.11),

$$\begin{array}{rl} & P_N\{a \le S_N \le b\} \ge P_N\{[a\Delta_N \le \Gamma_N \le b\Delta_N] \cap [|\Delta_N - 1| \le \varepsilon]\} \ge \\ \ge & P_N\{[a(1+\varepsilon) \le \Gamma_N \le b(1-\varepsilon)] \cap [|\Delta_N - 1| \le \varepsilon]\} = \\ = & P_N\{a(1+\varepsilon) \le \Gamma_N \le b(1-\varepsilon)\} + \\ - & P\{[a(1+\varepsilon) \le \Gamma_N \le b(1-\varepsilon)] \cap [|\Delta_N - 1| > \varepsilon]\} \\ \ge & P_N\{a(1+\varepsilon) \le \Gamma_N \le b(1-\varepsilon)\} - \varepsilon \ . \end{array}$$

poiché d'altro canto

$$P_N\{a(1+\varepsilon) \le \Gamma_N \le b(1-\varepsilon)\} \to P\{a(1+\varepsilon) \le Z \le b(1-\varepsilon)\}$$

per $N > N''_\varepsilon$ opportuno sarà

$$P_N\{a \le S_N \le b\} \ge P\{a(1+\varepsilon) \le Z \le b(1-\varepsilon)\} - 2\varepsilon \ ,$$

che combinata con la (1.3.12) dice che per tutti gli N sufficientemente grandi

$$
\begin{aligned}
& P\{a(1+\varepsilon) \leq Z \leq b(1-\varepsilon)\} - 2\varepsilon \leq P_N\{a \leq S_N \leq b\} \leq \\
\leq \quad & P\{a(1-\varepsilon) \leq Z \leq b(1+\varepsilon)\} + 2\varepsilon \ ;
\end{aligned}
$$

per l'arbitrarietà di ε resta perciò provato che

$$
\lim_{N\to\infty} P_N\{a \leq S_N \leq b\} = P\{a \leq Z \leq b\} \tag{1.3.13}
$$

per tutti gli $a, b > 0$. Con ragionamenti analoghi si prova il caso generale.

Dunque vale la convergenza in legge.

Osservazione 1.3.3: è utile notare che se pure la (1.3.9) è valida, il suo grado di approssimazione sarà peggiore che nel caso in cui la varianza è nota, infatti la variabile $(\mathcal{M}-\mu)(\overline{\mathcal{S}}|\sqrt{N})^{-1}$ ha una differenza da una Z dovuta sia alla convergenza di $\mathcal{M}$ a μ, che alla convergenza di $\overline{\mathcal{S}}$ a σ.

Osservazione 1.3.4: lo stesso procedimento può essere applicato al confronto tra medie, $H_0 : \mu_X = \mu_Y$, quando non si conoscano le varianze teoriche σ_X^2, σ_Y^2. Poiché per $N_X, N_Y \to \infty$ si ha in probabilità

$$
\sqrt{\frac{\overline{\mathcal{S}}_X^2}{N_X} + \frac{\overline{\mathcal{S}}_Y^2}{N_Y}} \xrightarrow{P} \sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}} \ ,
$$

si potrà come in (1.3.9) porre

$$
\frac{(\mathcal{M}_X - \mathcal{M}_Y) - (\mu_X - \mu_Y)}{\sqrt{\frac{\overline{\mathcal{S}}_X^2}{N_X} + \frac{\overline{\mathcal{S}}_Y^2}{N_Y}}} \ , \tag{1.3.14}
$$

il che appunto ci permette di eseguire dei test su $\mu_X - \mu_Y$. Si può anche notare che vi sono casi in cui si può ragionevolmente supporre che $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, mentre si vuole sottoporre a test $\mu_X - \mu_Y$. In questo caso è opportuno dare un'unica stima di σ^2 basandosi su entrambi i campioni $\underline{x}^{(N)}$ e $\underline{y}^{(N)}$, in modo che lo stimatore sia più attendibile.

Ciò di solito è fatto prendendo una nuova stima che sia una combinazione lineare di $\overline{\mathcal{S}}_X^2$ ed $\overline{\mathcal{S}}_Y^2$

$$\overline{\mathcal{S}}^2 = a\overline{\mathcal{S}}_X^2 + b\overline{\mathcal{S}}_Y^2 \ .$$

Questa sarà non deviata se

$$\sigma^2 = E\{\overline{\mathcal{S}}^2\} = (a+b)\sigma^2 \ ,$$

cioè se

$$a + b = 1 \ . \tag{1.3.15}$$

Poiché la varianza di $\overline{\mathcal{S}}^2$ è data da

$$\sigma^2(\overline{\mathcal{S}}^2) = a^2\sigma^2(\overline{\mathcal{S}}_X^2) + b^2\sigma^2(\overline{\mathcal{S}}_Y^2) \tag{1.3.16}$$

si può pensare di minimizzare (1.3.16) sotto la condizione (1.3.15). Il minimo lo si ha per

$$a = \frac{c}{\sigma^2(\overline{\mathcal{S}}_X^2)} \quad , \quad b = \frac{c}{\sigma^2(\overline{\mathcal{S}}_Y^2)} \tag{1.3.17}$$

dove c è scelto in modo che valga la (1.3.15).

D'altra parte, ricordando la (1.4.5) del Quaderno n. 2, si può vedere che valgono le relazioni asintotiche

$$\sigma^2(\overline{\mathcal{S}}_X^2) = \frac{\text{cost}}{N_X - 1} + o\left(\frac{1}{N_X}\right) \quad ; \quad \sigma^2(\overline{\mathcal{S}}_Y^2) = \frac{\text{cost}}{N_Y - 1} + o\left(\frac{1}{N_Y}\right)$$

inoltre per la (1.4.6) del Quaderno n. 2 tale relazione è esatta per le distribuzioni normali, con

$$\text{cost} \ = 2\sigma^4 \ .$$

Ne segue che è vantaggioso, almeno asintoticamente, porre

$$a = c(N_X - 1) \quad , \quad b = c(N_Y - 1) \ ,$$

ovvero imponendo $a + b = 1$,

$$a = \frac{N_X - 1}{N_X + N_Y - 2} \quad , \quad b = \frac{N_Y - 1}{N_X + N_Y - 2} \ .$$

Lo stimatore congiunto della varianza così ottenuto è

$$\overline{\mathcal{S}}^2 = \frac{N_X - 1}{N_X + N_Y - 2}\overline{\mathcal{S}}_X^2 + \frac{N_Y - 1}{N_X + N_Y - 2}\overline{\mathcal{S}}_Y^2 \ . \tag{1.3.18}$$

La relazione che permette di vagliare un'ipotesi su $\mu_X - \mu_Y$, diviene

$$\frac{(\mathcal{M}_X - \mathcal{M}_Y) - (\mu_X - \mu_Y)}{\overline{\mathcal{S}}\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}} \sim Z \ . \tag{1.3.19}$$

1.4 Test di buon adattamento

Vogliamo risolvere in questo paragrafo il problema dell'inferenza statistica per il confronto tra una distribuzione campionaria ed una corrispondente distribuzione teorica nota in base all'ipotesi H_0. Si avranno due casi: H_0 sarà semplice se la distribuzione teorica sarà definita univocamente; H_0 sarà composta se H_0 specificherà solo una famiglia parametrica di distribuzioni $f(x;\theta)$, in cui θ diviene un parametro (o più) di disturbo. Il confronto tra distribuzione empirica e distribuzione teorica può avvenire sia per confronto delle rispettive funzioni di distribuzione, sia comparando tra loro un istogramma con una funzione teorica di densità "raggruppata" per classi (cioè omogenea all'interno di ogni classe e con aree dei rettangoli esattamente uguali a quelle teoriche).

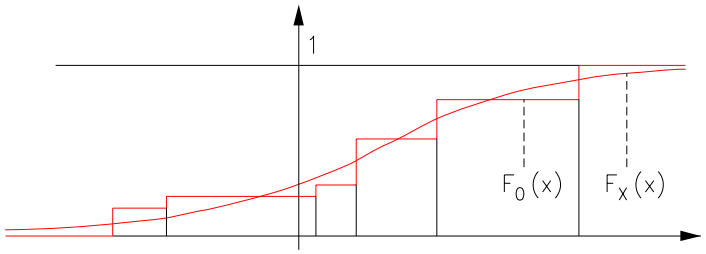


Figura 1.4.1: Confronto tra funzioni di distribuzione.

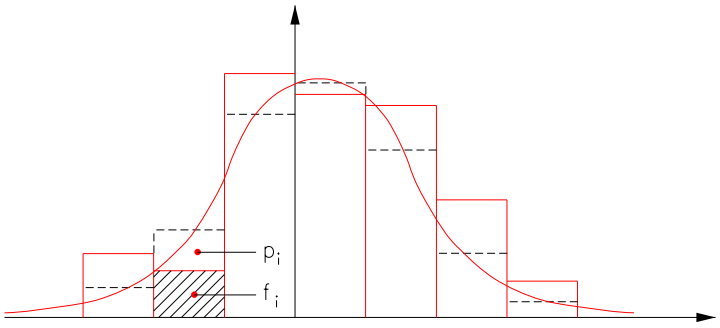


Figura 1.4.2: Confronto fra istogramma empirico (——) e istogramma "teorico" (- - - -).

a) Test di Kolmogorov

Data la funzione cumulativa di frequenza $F_0(x)$, si consideri come statistica

$$D = \sup_x |F_0(x) - F_X(x)| \ ; \tag{1.4.1}$$

si può osservare che D deve essere necessariamente un valore assunto dalla (1.4.1) in uno dei punti di salto della $F_0(x)$.

Si noti che, per ogni x, $F_0(x)$ è una statistica in quanto, introducendo la funzione di Heaviside $h(x) = 0,\ x < 0;\ h(x) = 1,\ x \geq 0$, si può scrivere

$$F_0(x) = \sum_{i=1}^{N} h(x - X_i) \ .$$

Con argomenti che esulano da questo ambito, si può provare che per $N \to \infty$ la distribuzione asintotica di D è definita da

$$\lim P\{D \leq \frac{x}{\sqrt{N}}\} = 1 + 2\sum_{k=1}^{+\infty}(-1)^k e^{-2k^2x^2} \ . \tag{1.4.2}$$

Osservazione 1.4.1: come si vede la distribuzione asintotica di $\sqrt{N}D_{(N)}$ non dipende dalla distribuzione teorica $F_X(x)$: in realtà ciò è vero per ogni N perché una qualunque trasformazione (monotona) di X cambierebbe nello stesso modo la $F_X(x)$ e la $F_0(x)$, lasciando inalterate le differenze (1.4.1). Così ad esempio definendo $Y = F_X(X)$, si ha una variabile uniformemente distribuita su [0,1] e D diventa

$$D = \sup_{y \in [0,1]} |F_0[F_X^{-1}(y)] - y| \ . \tag{1.4.3}$$

b) Test del χ^2

Diviso l'asse in m intervalli, si inizia cercando la distribuzione congiunta di numeri empirici N_i di estrazioni che cadono nei vari intervalli $I_i (i = 1, \ldots, m)$.

In effetti già sappiamo che ogni N_i è asintoticamente normale con media $\nu_i = Np_i$ e varianza $\sigma_i^2 = \nu_i(1 - \nu_i/N)$ (cfr. Quaderno n. 1, paragrafo 18).

Tuttavia la distribuzione congiunta del vettore $\underline{N}^+ = [N_1, \ldots, N_m]$ va ricercata direttamente e non può essere dedotta dalle distribuzioni marginali in quanto gli N_i non sono tra loro dipendenti poiché deve valere la relazione lineare

$$\sum_{i=1}^{m} N_i = \sum_{i=1}^{m} \nu_i = N \sum_{i=1}^{m} p_i = N \ . \tag{1.4.4}$$

D'altra parte, definendo le funzioni binarie (o contatori) della v.c. campionaria X_k per l'intervallo I_i

$$C_i(X_k) = \begin{cases} 1 & \text{se} \quad X_k \in I_i \\ 0 & \text{se} \quad X_k \in I_i \end{cases} \ , \tag{1.4.5}$$

ovvero

$$P\{C_i(X_k) = 1\} = p_i \quad ; \quad P\{C_i(X_k) = 0\} = 1 - p_i \ , \tag{1.4.6}$$

si viene a creare un vettore $\underline{C}(X_k) = \begin{vmatrix} C_1(X_k) \\ \vdots \\ C_m(X_k) \end{vmatrix}$ che descrive un avvenimento con m possibili risultati, ognuno con la sua probabilità p_i; infatti il vettore $\underline{C}(X_k)$ deve necessariamente avere una componente uguale ad 1 e tutte le altre nulle. Come si vede si ha una generalizzazione del semplice gioco a testa e croce.

Ora possiamo rappresentare il vettore $\underline{N}$ come

$$\underline{N} = \begin{vmatrix} N_1 \\ \vdots \\ N_m \end{vmatrix} = \begin{vmatrix} \sum_{k=1}^{N} C_1(X_k) \\ \vdots \\ \sum_{k=1}^{N} C_m(X_k) \end{vmatrix} = \sum_{k=1}^{N} \underline{C}(X_k) \ ; \tag{1.4.7}$$

per comodità passiamo da $\underline{N}$ al vettore $\underline{Y}$ di componenti

$$Y_i = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \frac{C_i(X_k) - p_i}{\sqrt{p_i}} \ , \tag{1.4.8}$$

ciascuna delle quali ha media nulla.

Ricerchiamo la distribuzione di $\underline{Y}$ tramite la sua funzione generatrice dei momenti, ovvero

$$\begin{aligned} G_Y(\underline{t}) &= E\{e^{\underline{t}^+ \underline{Y}}\} = E\{e^{\sum_{i=1}^m t_i Y_i}\} = \\ &= E\left\{ \exp\left(\sum_{k=1}^{N} \sum_{i=1}^{m} \frac{t_i}{\sqrt{N}} \frac{C_i(X_k) - p_i}{\sqrt{p_i}} \right) \right\} = \\ &= E\left\{ \Pi_{k=1}^{N} \exp\left(\sum_{i=1}^{m} \frac{t_i}{\sqrt{N}} \frac{C_i(X_k) - p_i}{\sqrt{p_i}} \right) \right\} = \\ &= \left[E\left\{ \exp\left(\sum_{i=1}^{m} \frac{t_i}{\sqrt{N}} \frac{C_i(X_k) - p_i}{\sqrt{p_i}} \right) \right\} \right]^N = \ . \end{aligned}$$

Ora, posto

$$\widetilde{t} = \sum_{i=1}^{m} t_i \sqrt{p_i} \ , \tag{1.4.9}$$

risulta

$$\begin{aligned} G_Y(\underline{t}) &= \left[e^{-\frac{\widetilde{t}}{\sqrt{N}}} E\left\{ e^{\frac{1}{\sqrt{N}} \sum_{i=1}^m \frac{t_i}{\sqrt{p_i}} C_i(X)} \right\} \right]^N = \qquad (1.4.10) \\ &= \left[e^{-\frac{\widetilde{t}}{\sqrt{N}}} \sum_{i=1}^{m} p_i e^{\frac{1}{\sqrt{N}} \frac{t_i}{\sqrt{p_i}}} \right]^N = \\ &= \left[\sum_{i=1}^{m} p_i e^{\frac{1}{\sqrt{N}} \left(\frac{t_i}{\sqrt{p_i}} - \widetilde{t} \right)} \right]^N = \\ &= \left\{ \sum_{i=1}^{m} p_i \left[1 + \frac{1}{\sqrt{N}} \left(\frac{t_i}{\sqrt{p_i}} - \widetilde{t} \right) + \frac{1}{2N} \left(\frac{t_i}{\sqrt{p_i}} - \widetilde{t} \right)^2 + O_{-3/2} \right] \right\}^N \ , \end{aligned}$$

dove $O_{-3/2} = 0(N^{-3/2})$.

Ora sfruttando il fatto che $\sum_{i=1}^{m} p_i = 1,\ \sum p_i(\frac{t_i}{\sqrt{p_i}} - \widetilde{t}) = 0$ per la definizione (1.4.9) ed usando la nota relazione $(1 + \frac{x}{N})^N \to e^x$, dalla (1.4.10) ricaviamo l'espressione asintotica

$$\begin{aligned} G_Y(\underline{t}) &\sim e^{1/2\sum_{i=1}^{m} p_i \left(\frac{t_i}{\sqrt{p_i}} - \widetilde{t}\right)^2} \\ &= e^{1/2\left[\sum_{i=1}^{m} t_i^2 - \left(\sum_{i=1}^{m} t_i\sqrt{p_i}\right)^2\right]} \ . \end{aligned} \tag{1.4.11}$$

Vogliamo ora dimostrare che $\underline{Y}$ è un vettore normale standardizzato con una distribuzione singolare, più precisamente con il supporto in quella varietà $m-1$ dimensionale che è definita dai vettori che soddisfano la relazione di ortogonalità

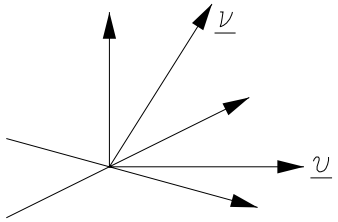
$$\sum_{i=1}^{m} \sqrt{p_i}\ Y = \underline{a}^+ \underline{Y} = 0 \tag{1.4.12}$$

$$(\underline{a}^+ = [\sqrt{p_i} \ldots \sqrt{p_m}] \quad , \quad |\underline{a}|^2 = \sum_{i=1}^{m} p_i = 1) \ .$$

Che il vettore $\underline{Y}$ definito dalla (1.4.8) soddisfi effettivamente la (1.4.12) è di immediata verifica quando si osservi che

$$\sum_{i=1}^{m} C_i(X_k) \equiv 1 \ ;$$

per il resto della dimostrazione conviene invece usare un sistema di coordinate più comode. Perciò supponiamo che $\underline{V}$ sia un vettore $m-1$ dimensionale, normale e standardizzato; formiamo poi il vettore m-dimensionale



$$\underline{U} = \left| \begin{array}{c} \underline{V} \\ \hline O \end{array} \right| \ .$$

Per definizione $\underline{U}$ è una normale standardizzata singolare con supporto in R^{m-1}. Ora sia $\underline{t}$ un vettore costante di R^m, decomposto secondo la formula

$$\underline{t} = \left| \begin{array}{c} \underline{\tau} \\ 0 \end{array} \right| + \eta \left| \begin{array}{c} \underline{O} \\ 1 \end{array} \right| \left(\underline{\tau} = \left| \begin{array}{c} \tau_1 \\ \vdots \\ \tau_{m-1} \end{array} \right| , \ \underline{O} = \left| \begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \right| \right) \ ;$$

è chiaro che

$$\underline{t}^+ \underline{U} = \underline{\tau}^+ \underline{V} \ .$$

Pertanto

$$G_{\underline{U}}(\underline{t}) = G_{\underline{V}}(\underline{\tau}) = e^{1/2 \sum_{i=1}^{m-1} \tau_i^2} \ ,$$

che, introdotto il vettore $\underline{\varepsilon} = \left| \begin{array}{c} \underline{O} \\ 1 \end{array} \right|$, può essere scritto come

$$G_{\underline{V}}(\underline{\tau}) = e^{1/2[|t|^2 - (\underline{t}^+ \underline{\varepsilon})^2]} \ . \tag{1.4.13}$$

Ora sia R una qualsiasi rotazione d'assi che porti $\underline{\varepsilon}$ in $\underline{a} = R\underline{\varepsilon}$; una tale rotazione esiste sempre purché $|\underline{a}| = R\underline{\varepsilon}$, ciò che in realtà è per la (1.4.13). Per un teorema generale sappiamo che la nuova variabile $\underline{Y} = R\underline{U}$ sarà anch'essa normale e standardizzata mentre il suo supporto sarà la varietà ortogonale ad $\underline{a}$ (così come R^{m-1}, cioè la varietà ortogonale ad $\underline{\varepsilon}$, era il supporto di $\underline{V}$). Quanto alla funzione generatrice di $\underline{Y}$, posto $\underline{\lambda} = R\underline{t}$, si ha $\underline{\lambda}^+ \underline{Y} = \underline{t}^+ \underline{U}$, così che dalla (1.4.13)

$$\begin{aligned} G_{\underline{Y}}(\underline{\lambda}) &= G_{\underline{U}}(\underline{t}) = e^{1/2[|t|^2 - (\underline{t}^+ \underline{\varepsilon})^2]} \\ &= e^{1/2[|\underline{\lambda}|^2 - (\underline{\lambda}^+ \underline{a})^2]} \ . \end{aligned} \tag{1.4.14}$$

Confrontando la (1.4.14) con la (1.4.11) si vede che esse sono identiche ovvero le due v.c. $\underline{Y}$, una definita da (1.4.7), l'altra definita da $\underline{Y} = R\underline{U}$ danno la stessa distribuzione, che è quanto volevamo provare.

In particolare allora sarà

$$|\underline{Y}|^2 = |\underline{U}|^2 = |\underline{V}|^2 = \chi^2_{m-1}$$

e d'altro canto,

$$\begin{aligned} |\underline{Y}|^2 &= \sum_{i=1}^{m} Y_i^2 = \sum_{i=1}^{m} \frac{1}{N} \frac{[\sum_{k=1}^{N} C_i(X_k) - Np_i]^2}{p_i} = \\ &= \sum_{i=1}^{m} \frac{(N_i - \nu_i)^2}{\nu_i} \ , \end{aligned}$$

così che vale la relazione

$$\sum_{i=1}^{m} \frac{(N_i - \nu_i)^2}{\nu_i} \cong \chi^2_{m-1} \ . \tag{1.4.15}$$

In base a questa relazione il test sull'ipotesi

$$H_0 \equiv \{P\{X \in I_i\} = \frac{\nu_i}{N} \ , \ 1 = 1, 2 \ldots, m\}$$

può essere effettuato al livello di significatività α, verificando se

$$\begin{aligned} &\sum_{i=1}^{m} \frac{(N_i - \nu_i)^2}{\nu_i} \leq \chi^2_\alpha \\ &P(\chi^2_{m-1} \leq \chi^2_\alpha) = 1 - \alpha \ . \end{aligned} \tag{1.4.16}$$

Se la (1.4.16) è verificata H_0 è accettata, in caso contrario H_0 è rifiutata.

Osservazione 1.4.2: qualora l'ipotesi H_0, sulla distribuzione di X, fosse composta (ad esempio si dica che X è normale senza specificare media e varianza) si pone il problema di eliminare l'influenza dei parametri di disturbo. La cosa risulta abbastanza semplice quando tali parametri siano stimabili con momenti campionari, così una volta fissati i valori dei parametri, restano anche definite le probabilità teoriche p_i ed il numero teorico di estrazioni per intervalli, ν_i.

Naturalmente, stimando θ dal campione si creano dei legami tra le variabili N_i. In effetti si pensi al caso della media; condizionando la media teorica ad essere uguale a quella empirica si dice che, approssimativamente

$$\frac{1}{N}\sum N_i\xi_i = \frac{1}{N}\sum \nu_i\xi_i \ ,$$

dove ξ_i sono i punti medi degli intervalli I_i.

Se si fosse usato anche il momento del secondo ordine (caso della normale con media e varianza incognita), si avrebbe che

$$\frac{1}{N}\sum N_i\xi_i^2 = \frac{1}{N}\sum \nu_i\xi_i^2 \ .$$

Si noti che queste due relazioni possono essere scritte come

$$\begin{cases} \sum \delta N_i\xi_i = 0 \\ \sum \delta N_i\xi_i^2 = 0 \end{cases} \quad (\delta N_i = N_i - \nu_i) \tag{1.4.17}$$

il che equivale a vincolare il vettore

$$\frac{\delta N_i}{\sqrt{\nu_i}}$$

alla intersezione dei tre piani risultanti dalle (1.4.17) e dalla $\sum_i \delta N_i = 0$ così che esso risulterà distribuito come una normale standardizzata di dimensioni

$$\dim = m - 1 - (\text{n°parametri stimati}) = m - 1 - h \ .$$

Pertanto in questo caso la (1.4.16) è sostituita dalla relazione

$$\sum_{i=1}^{m} \frac{(N_i - \nu_i)^2}{\nu_i} \sim \chi^2_{(m-1-h)} \ . \tag{1.4.18}$$

Osservazione 1.4.3: nell'uso delle formule asintotiche (1.4.16), (1.4.18) è bene usare qualche cautela, garantendosi ad esempio che la divisione in intervalli sia tale per cui ν_i è piccolo rispetto ad N, ma nello stesso tempo non troppo piccolo: ad esempio può porsi il limite $\nu_i \geq 5$. Naturalmente un test di buon adattamento tipo χ^2 ha senso solo per campioni abbastanza numerosi, almeno di alcune decine di elementi.

1.5 Campioni normali

Data l'importanza delle distribuzioni normali in statistica, vale la pena di approfondire il caso in cui $X = \mathcal{N}[\mu, \sigma^2]$. Si vogliono costruire specifici test per μ e σ^2, nonché confronti tra medie e varianze per i campioni normali.

Poiché le statistiche con cui si possono costruire test per μ e σ^2 sono naturalmente $\mathcal{M}$ ed $\mathcal{S}^2$ ci si pone il problema della distribuzione congiunta di tali variabili. Dimostriamo il seguente teorema.

Teorema 1.5.1: se X è normale ed $\underline{X}^{(N)}$ una variabile campionaria bernoulliana, allora $\mathcal{M}$ ed $\mathcal{S}^2$ sono variabili tra loro indipendenti; inoltre

$$\mathcal{M} = \mathcal{N}\left[\mu, \frac{\sigma^2}{N}\right] \tag{1.5.1}$$

$$\mathcal{S}^2 = \frac{\sigma^2}{N}\chi^2_{(N-1)} \ . \tag{1.5.2}$$

Per dimostrare il teorema cominciamo a considerare, per un i qualunque fissato, la coppia di variabili

$$\begin{cases} \mathcal{M} = \frac{1}{N}\sum_i X_j \\ X_i - \mathcal{M} = X_i - \frac{1}{N}\sum_i X_j \ . \end{cases} \tag{1.5.3}$$

Come è ovvio (1.5.3) definisce una trasformazione lineare da $\underline{X}^{(N)}$ alla variabile doppia $(\mathcal{M}, X_i - \mathcal{M})$.

Poiché $\underline{X}^{(N)}$ è normale,

$$\underline{X}^{(N)} = \mathcal{N}[\mu\underline{e}; \sigma^2 I] \quad (\underline{e}^+ = [1 \quad 1 \ldots 1]) \ , \tag{1.5.4}$$

anche $(\mathcal{M}, X_i - \mathcal{M})$ sarà normale: in particolare si ha

$$E\{X_i - \mathcal{M}\} = 0 \ . \tag{1.5.5}$$

Da (1.5.5) segue che la covarianza tra $\mathcal{M}$ e $X_i - \mathcal{M}$ può essere calcolata come

$$\begin{aligned} & E\{(\mathcal{M} - \mu)(X_i - \mathcal{M})\} = \\ = \quad & E\{\frac{1}{N}\sum_j (X_j - \mu)X_i\} - E\{(\mathcal{M} - \mu)\mathcal{M}\} = \\ = \quad & \frac{1}{N}E\{(X_j - \mu)^2\} - E\{(\mathcal{M} - \mu)^2\} = \\ = \quad & \frac{1}{N}\sigma^2 - \frac{1}{N}\sigma^2 = 0 \ . \end{aligned} \tag{1.5.6}$$

Siccome $\mathcal{M}, X_i - \mathcal{M}$ sono congiuntamente normali, la (1.5.6) implica anche che esse siano variabili stocasticamente indipendenti. D'altronde la varianza campionaria

$$\mathcal{S}^2 = \frac{1}{N}\sum (X_i - \mathcal{M})^2$$

è essenzialmente funzione solo degli scarti $X_i - \mathcal{M}$ e pertanto è anch'essa una variabile stocasticamente indipendente da $\mathcal{M}$. Detto ciò non resta che studiare separatamente le distribuzioni di $\mathcal{M}$ ed $\mathcal{S}^2$: quanto ad $\mathcal{M}$, già sappiamo che la (1.5.1) è vera.

Passiamo quindi a notare che si può scrivere l'identità (pitagorica)

$$\frac{1}{N}\sum(X_i - \mu)^2 = (\mathcal{M} - \mu)^2 + \mathcal{S}^2$$

ovvero

$$\sum_{i=1}^{N} \frac{(X_i - \mu)^2}{\sigma^2} = \frac{(\mathcal{M} - \mu)^2}{\frac{\sigma^2}{N}} + \frac{N}{\sigma^2}\mathcal{S}^2 \ . \tag{1.5.7}$$

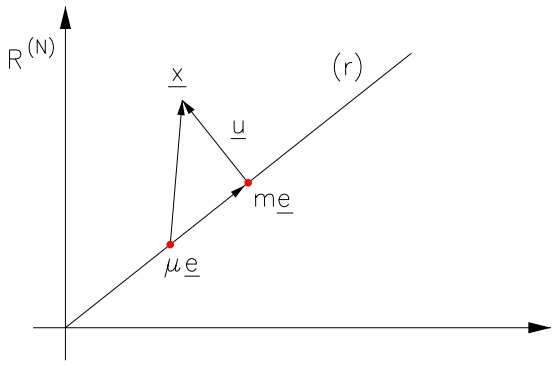
Al primo membro si ha una variabile $\chi^2_{(N)}$ (cfr. Quaderno n. 1, Esempio 13.2) al secondo membro la prima variabile è una $\chi^2_{(1)}$ e la seconda variabile indipendente dalla prima. Pertanto per una osservazione fatta nel Quaderno n. 1, paragrafo 13, sul teorema della somma di χ^2 indipendenti (13.19) si ha anche che

$$\frac{N}{\sigma^2}\mathcal{S}^2 = \chi^2_{(N-1)} \ ,$$

che coincide con la (1.5.2).

Osservazione 1.5.1: da un punto di vista geometrico il teorema ora dimostrato e le (1.5.1), (1.5.2), (1.5.7) hanno una interpretazione intuitiva. Sia $R^{(N)}$ lo spazio dei campioni e sia (r) la retta descritta dai vettori $\lambda\underline{e}$(λ reale): notiamo che, per la (1.5.4), la media di $\underline{\underline{X}}^{(N)}$ deve stare su (r).

Decomponiamo il vettore degli scarti teorici $\underline{X} - \mu\underline{e}$ nella somma di $(\mathcal{M} - \mu)\underline{e}$ e del vettore degli scarti campionari



$$
\begin{aligned}
|\underline{x} - \mu\underline{e}|^2 &= \sum (x_i - \mu)^2 \\
|\mathcal{M}\underline{e} - \mu\underline{e}|^2 &= N(\mathcal{M} - \mu)^2 \\
|\underline{u}|^2 &= \sum (x_i - \mathcal{M})^2
\end{aligned}
$$

Figura 1.5.1

$$
\begin{aligned}
&\underline{\mathcal{U}} = \underline{X} - \mathcal{M}\underline{e} \\
&\underline{X} - \mu\underline{e} = (\mathcal{M} - \mu)\underline{e} + \underline{\mathcal{U}} \ : \qquad (1.5.8)
\end{aligned}
$$

poiché $\underline{\mathcal{U}}$ è sempre ortogonale ad (r), in quanto

$$
\underline{e}^{+}\underline{\mathcal{U}} = \sum \mathcal{U}_i = \sum (X_i - \mathcal{M}) \equiv 0 \ ,
$$

vale tra i moduli dei tre vettori la relazione pitagorica

$$
|\underline{X} - \mu\underline{e}|^2 = |(\mathcal{M} - \mu)\underline{e}|^2 + |\underline{\mathcal{U}}|^2 \ ,
$$

che coincide esattamente con la (1.5.7).

Inoltre la variabile $\underline{X} - \mu\underline{e}$ è normale con matrice di covarianza isotropa, perciò la decomposizione (1.5.8) in componenti ortogonali porta ancora a variabili normali con matrice di covarianza isotropa $\sigma^2 I$ (cfr. Quaderno n. 1, Esempio 17.1): $(\mathcal{M} - \mu)\underline{e}$ ci dà la componente lungo (r), mentre $\underline{\mathcal{U}}$ ci dà la componente perpendicolare ad (r), perciò $\underline{\mathcal{U}}$ è indipendente da $(\mathcal{M} - \mu)\underline{e}$ e inoltre la matrice di covarianza di $\underline{\mathcal{U}}$, nel suo sottospazio, cioè nella varietà $(N-1)$-dimensionale ortogonale ad (r), è ancora $\sigma^2 I_{(N-1)}$.

Ne segue che

$$\frac{N\mathcal{S}^2}{\sigma^2} = \frac{|\underline{\mathcal{U}}|^2}{\sigma^2} = \chi^2_{(N-1)}$$

per l'osservazione 17.6 nel Quaderno n. 1.

Osservazione 1.5.2: dalla distribuzione campionaria di $\underline{X}^{(N)}$ si può anche derivare la distribuzione di altre statistiche. In primo luogo notiamo che avendo trovato la distribuzione di $\mathcal{S}^2$ è nota anche quella della varianza campionaria corretta

$$\overline{\mathcal{S}}^2 = \frac{\sigma^2}{N-1}\chi^2_{(N-1)} \ . \tag{1.5.9}$$

Inoltre si può anche ricavare la distribuzione rigorosa ad esempio dei coefficienti di skewness e curtosi campionari, cioè

$$B = \frac{\overline{\mathcal{M}}_{(3)}}{\mathcal{S}^3} = \frac{(1/N)\sum(X_i - \mathcal{M})^3}{[(1/N)\sum(X_i - \mathcal{M})^2]^{3/2}} \tag{1.5.10}$$

$$\Gamma = \frac{\overline{\mathcal{M}}_{(4)}}{\mathcal{S}^4} = \frac{(1/N)\sum(X_i - \mathcal{M})^4}{[(1/N)\sum(X_i - \mathcal{M})^2]^2} \ . \tag{1.5.11}$$

Noi qui ci limitiamo a riportare che per N grande valgono le distribuzioni asintotiche

$$B \sim \mathcal{N}[0, \quad 6/N] \tag{1.5.12}$$

$$\Gamma \sim \mathcal{N}[3, \quad 24/N] \ . \tag{1.5.13}$$

1.6 Test per le medie di campioni normali

Riconsideriamo i due problemi di verifica di ipotesi

a) per un campione di numerosità N

$$H_0 : \mu = \mu_0 \tag{1.6.1}$$

b) per due campioni di numerosità N_X ed N_Y, di eguale varianza $\sigma_X^2 = \sigma_Y^2$,

$$H_0 : \mu_X = \mu_Y \ , \tag{1.6.2}$$

sfruttando la conoscenza delle distribuzioni campionarie studiate nel paragrafo 1.5.

a) Supponiamo che $X = \mathcal{N}[\mu, \sigma^2]$, e di voler disegnare un test per l'ipotesi (1.6.1). Il problema in questo caso è che H_0 è una ipotesi composta con σ^2 come parametro di disturbo, così per il test desiderato non si può semplicemente usare la statistica $|\mathcal{M} - \mu|$. D'altro canto per le (1.5.1) e (1.5.2) si può scrivere

$$\frac{\mathcal{M} - \mu}{\frac{\sigma}{\sqrt{N}}} = Z \quad , \quad \frac{\overline{\mathcal{S}}}{\sigma} = \sqrt{\frac{\chi^2_{(N-1)}}{(N-1)}} \tag{1.6.3}$$

e le due variabili $\mathcal{M}$ ed $\overline{\mathcal{S}}$ sono tra loro indipendenti in base al Teorema 1.5.1.

Dividendo tra loro le due relazioni (1.6.3) e ricordando l'Esempio 13.3 del Quaderno n. 1, si trova

$$\frac{\mathcal{M} - \mu}{\frac{\overline{\mathcal{S}}}{\sqrt{N}}} = t_{(N-1)} \qquad (t \text{ di Student}) \ . \tag{1.6.4}$$

La (1.6.4) lega le variabili campionarie $\mathcal{M}, \overline{\mathcal{S}}$ ed il parametro μ ad una variabile nota; si ha quindi una relazione adatta alla verifica di ipotesi su μ.

In particolare possiamo ritenere che valori grandi della statistica campionaria

$$S(\underline{x}) = \frac{|m - \mu_0|}{\overline{s}/\sqrt{N}} \ . \tag{1.6.5}$$

diano una evidenza della non plausibilità dell'ipotesi $H_0 : \mu = \mu_0$.

Il test su H_0 pertanto si effettuerà fissando dapprima il valore della significatività α prescelto, in secondo luogo cercando il valore critico per cui

$$P\{|t_{(N-1)}| \geq t_{\alpha/2}\} = 2P\{t_{(N-1)} \geq t_{\alpha/2}\} = \alpha \ ,$$

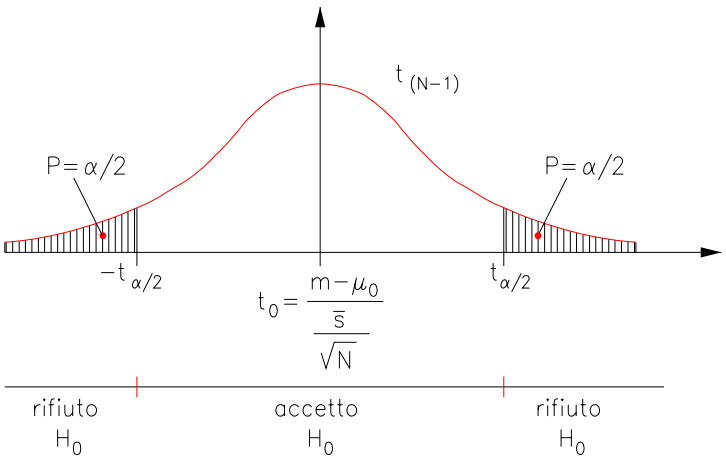


Figura 1.6.1:

infine, calcolando il valore empirico osservato

$$t_0 = \frac{m - \mu_0}{\overline{s}/\sqrt{N}} \ : \tag{1.6.6}$$

se l'ipotesi H_0 è corretta t_0 è una estrazione da una $t_{(N-1)}$ di Student a $N-1$ gradi di libertà e dunque deve essere

$$|t_0| \leq t_{\alpha/2} \tag{1.6.7}$$

con probabilità $P = 1 - \alpha$.

Pertanto, al livello di significatività α, se la (1.6.7) è verificata accetto H_0, in caso contrario rifiuto H_0. La decisione su H_0 è quindi presa verificando se la relazione

$$|m - \mu_0| \leq \frac{\overline{s}}{\sqrt{N}} t_{\alpha/2} \tag{1.6.8}$$

è soddisfatta oppure no; al solito la stessa relazione può servire per definire un intervallo fiduciario di μ.

Fissati i valori campionari m, s e fissato α (e quindi $t_{\alpha/2}$) l'insieme dei μ_0 per cui la (1.6.8) è verificata ci dà l'intervallo richiesto.

Osservazione 1.6.1: notiamo che, poiché $t_{(N)} \to Z$ in legge per $N \to \infty$, la (1.6.4) è in perfetto accordo con la (1.3.9).

b) supponiamo ora che $X = \mathcal{N}[\mu_X, \sigma^2]$ e $Y = \mathcal{N}[\mu_Y, \sigma^2]$. Si sono estratti da X un campione di numerosità N_X e da Y, in modo indipendente, un campione di numerosità N_Y; si vuole verificare l'ipotesi

$$H_0 : \mu_X = \mu_Y \ , \tag{1.6.9}$$

(o una qualsiasi altra ipotesi del tipo $\mu_X - \mu_Y = \mu_0$).

Ora per i due campioni si ha

$$\mathcal{M}_X = \mathcal{N}\left[\mu_X, \frac{\sigma^2}{N_X}\right] \quad , \quad \mathcal{M}_Y = \mathcal{N}\left[\mu_Y, \frac{\sigma^2}{N_Y}\right] \ ,$$

così che per sottoporre a test la (1.6.9) si può pensare di prendere come statistica

$$\mathcal{M}_X - \mathcal{M}_Y = \mathcal{N}\left[\mu_X - \mu_Y, \left(\frac{1}{N_X} + \frac{1}{N_Y}\right)\sigma^2\right] \ ,$$

ovvero la relazione

$$\frac{(\mathcal{M}_X - \mathcal{M}_Y) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}} = Z \ . \tag{1.6.10}$$

Come si vede però nella (1.6.10) σ è un parametro di disturbo che occorre eliminare.

A questo scopo osserviamo che

$$\frac{(N_X - 1)}{\sigma^2}\overline{\mathcal{S}}^2_X = \chi^2_{(N_x-1)} \quad , \quad \frac{(N_Y - 1)}{\sigma^2}\overline{\mathcal{S}}^2_Y = \chi^2_{(N_y-1)} \ ;$$

sommando le due espressioni si trova

$$\frac{(N_X - 1)}{\sigma^2}\overline{\mathcal{S}}^2_X + \frac{(N_Y - 1)}{\sigma^2}\overline{\mathcal{S}}^2_Y = \chi^2_{(N_x+N_y-2)} \ . \tag{1.6.11}$$

Ciò tra l'altro conferma che

$$\begin{aligned} \overline{\mathcal{S}}^2 &= \frac{N_X - 1}{N_X + N_Y - 2}\overline{\mathcal{S}}^2_X + \frac{N_Y - 1}{N_X + N_Y - 2}\overline{\mathcal{S}}^2_Y = \\ &= \frac{\sigma^2}{N_X + N_Y - 2}\chi^2_{(N_X+N_Y-2)} \end{aligned}$$

è uno stimatore corretto di σ^2.

Inoltre $\overline{\mathcal{S}}^2$ è funzione di $\overline{\mathcal{S}}^2_X, \overline{\mathcal{S}}^2_Y$ che sono indipendenti entrambi tanto da $\mathcal{M}_X$ quanto da $\mathcal{M}_Y$, perciò anche $\overline{\mathcal{S}}^2$ è indipendente da tali variabili.

Da tutto ciò si deriva la relazione

$$\frac{(\mathcal{M}_X - \mathcal{M}_Y) - (\mu_X - \mu_Y)}{\overline{\mathcal{S}}\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}} = \frac{Z}{\sqrt{\frac{\chi^2_{(N_X+N_Y-2)}}{N_X+N_Y-2}}} = t_{(N_X+N_Y-2)} \qquad (1.6.12)$$

che è la relazione ricercata, adatta alla verifica di ipotesi su $\mu_X - \mu_Y$.

In particolare per il test di $H_0 : \mu_X = \mu_Y$, si usa il valore empirico

$$\frac{|m_x - m_y|}{\overline{s}\sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}} = t_0 \ : \qquad (1.6.13)$$

se H_0 è corretta, t_0 è un'estrazione da una $t_{(N_X+N_Y-2)}$, a $N_X + N_Y - 2$ gradi di libertà, e quindi sarà

$$|t_0| \leq t_{\alpha/2} \qquad (N_X + N_Y - 2 \ \text{ gradi di libertà}) \qquad (1.6.14)$$

con probabilità $P = 1 - \alpha$. Se si verifica questo caso, H_0 è accettata, se invece risulta

$$|t_0| > t_{\alpha/2} \qquad (1.6.15)$$

H_0 è rifiutata perché il valore empirico grande di $|t_0|$ viene preso come evidenza contro tale ipotesi.

1.7 Test sulla varianza campionaria

Il Teorema 1.5.1 ci permette direttamente di disegnare dei test per le varianze campionarie di campioni normali: infatti per la (1.5.9) sappiamo che

$$(N-1)\frac{\overline{\mathcal{S}}^2}{\sigma^2} = \chi^2_{(N-1)} \ . \qquad (1.7.1)$$

Pertanto se si pone l'ipotesi

$$H_0 : \sigma^2 = \sigma_0^2 \ ,$$

la distribuzione di $\overline{\mathcal{S}}^2$ viene completamente specificata e, fissato un livello di significatività α, è possibile disegnare una regione critica ed un intervallo di accettazione di H_0. Di solito come statistica per giudicare sulla plausibilità di H_0 si sceglie, implicitamente,

$$S = (N-1)\left(\frac{\overline{\mathcal{S}}^2}{\sigma^2} - 1\right) \ ,$$

ritenendo che se S è grande in valore assoluto, ci stiamo allontanando da H_0; tuttavia, poiché la distribuzione di base è qui quella di una $\chi^2_{(N-1)}$ che è asimmetrica, si preferisce usare un intervallo di accettazione pure asimmetrico.

Fissato il livello di significatività α ed ammesso che H_0 sia vera, si calcola il valore campionario

$$\frac{(N-1)\overline{s}^2}{\sigma_0^2} = \chi_0^2 \ ; \tag{1.7.2}$$

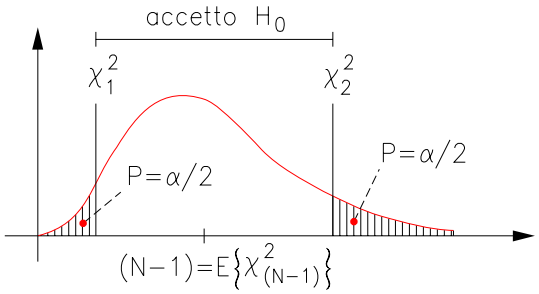


Figura 1.7.1:

in base ad H_0 tale valore dovrà trovarsi tra χ^2_1 e χ^2_2 con probabilità 1-α

$$\begin{aligned} P(\chi^2_{(N-1)} \leq \chi^2_1) &= \alpha/2 \\ P(\chi^2_{(N-1)} \geq \chi^2_2) &= \alpha/2 \ . \end{aligned} \tag{1.7.3}$$

Perciò H_0 verrà accettata se

$$\chi^2_1 \leq \frac{(N-1)\overline{s}^2}{\sigma^2} \leq \chi^2_2 \ , \tag{1.7.4}$$

rifiutata in caso contrario.

Osservazione 1.7.1: la relazione (1.7.4), che definisce il criterio di accettazione di H_0, permette nello stesso tempo di definire gli intervalli fiduciari per σ^2. Fissato il livello α ed usando gli stessi χ^2_1, χ^2_2 delle (1.7.3) si ha per tale intervallo

$$\frac{N-1}{\chi^2_2}\overline{s}^2 \leq \sigma^2 \leq \frac{N-1}{\chi^2_1}\overline{s}^2 \ , \tag{1.7.5}$$

ovvero, in termini di s.q.m.

$$\sqrt{\frac{N-1}{\chi^2_2}}\overline{s} \leq \sigma \leq \sqrt{\frac{N-1}{\chi^2_1}}\overline{s} \ . \tag{1.7.6}$$

È interessante notare che tale intervallo si riduce in ampiezza per $N \to \infty$, ma assai lentamente. Così, ad esempio, se $\overline{s} = 1$ ed $\alpha = 5\%$,

$$\begin{array}{rrrr} N = 2 & 0,45 & \leq \sigma \leq & 31,62 \\ 5 & 0,60 & \leq \sigma \leq & 2,87 \\ 10 & 0,69 & \leq \sigma \leq & 1,83 \\ 20 & 0,76 & \leq \sigma \leq & 1,46 \end{array}$$

Questa tabella dà l'idea di quanto sia in effetti variabile la stima empirica $\overline{s}$ rispetto a σ, se si pensa che ancora con un campione di numerosità 20,

valori più grandi di 0,76 $\overline{s}$ e più piccoli di 1,46 $\overline{s}$ continuano a passare il test, al livello del 5%.

Un'altro problema che è possibile risolvere sfruttando la (1.7.1) è quello di confrontare tra loro due varianze campionarie. Più precisamente si considerano due variabili normali indipendenti $X = \mathcal{N}[\mu_X, \sigma_X^2]$, $Y = = [\mu_Y, \sigma_Y^2]$ e due campioni estratti da queste, rispettivamente di numerosità N_X ed N_Y. In base alla (1.7.1) possiamo scrivere

$$\begin{cases} \overline{\mathcal{S}}_X^2/\sigma_X^2 = \chi^2_{(N_X-1)}/(N_X-1) \\ \overline{\mathcal{S}}_Y^2/\sigma_Y^2 = \chi^2_{(N_Y-1)}/(N_Y-1) \end{cases} \tag{1.7.7}$$

dove le due variabili χ^2 sono tra loro indipendenti perché funzioni di variabili indipendenti. Dividendo tra loro le due relazioni e ricordando l'Esempio 13.4 nel Quaderno n. 1, si ha

$$\begin{aligned} \frac{\overline{\mathcal{S}}_X^2}{\overline{\mathcal{S}}_Y^2} \cdot \left(\frac{\sigma_Y^2}{\sigma_X^2}\right) &= \frac{\frac{\chi^2_{(N_X-1)}}{N_X-1}}{\frac{\chi^2_{(N_Y-1)}}{N_Y-1}} = \\ &= F_{(N_X-1, N_Y-1)} \qquad (F \text{ di Fisher}) \ . \end{aligned} \tag{1.7.8}$$

Questa relazione ci dice che se si fissa per ipotesi H_0 il valore del parametro σ_Y^2/σ_X^2

$$H_0 \left(\frac{\sigma_Y^2}{\sigma_X^2} = k\right) \ , \tag{1.7.9}$$

allora la distribuzione della statistica $\overline{\mathcal{S}}_X^2/\overline{\mathcal{S}}_Y^2$ è completamente fissata e, sulla base del suo valore campionario, si potrà eseguire un test al livello di significatività α prefissato.

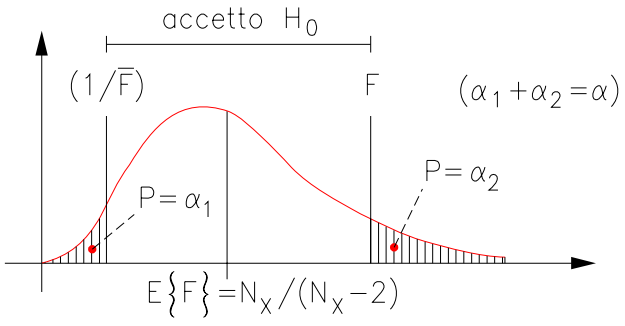

accetto H_0
$(1/\overline{F})$
F
$(\alpha_1+\alpha_2=\alpha)$
$P=\alpha_1$
$P=\alpha_2$
$E\{F\}=N_X/(N_X-2)$

Figura 1.7.2:

Ricordiamo che anche la F di Fisher ha un andamento asimmetrico, come quello in fig. 1.7.2. così che risulta naturale scegliere un intervallo asimmetrico.

Data la natura della variabile da sottoporre a test (rapporto tra due varianze campionarie), si usa spesso un intervallo del tipo

$$\frac{1}{F_{\alpha/2}} \leq \frac{\overline{s}_X^2}{\overline{s}_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \leq F_{\alpha/2} \quad : \tag{1.7.10}$$

notando che la (1.7.10) può anche essere scritta come

$$\sup \left\{ \frac{\overline{s}_X^2}{\overline{s}_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \ , \ \frac{\overline{s}_Y^2}{\overline{s}_X^2} \cdot \frac{\sigma_X^2}{\sigma_Y^2} \right\} \leq F_{\alpha/2} \ , \tag{1.7.11}$$

si riconosce che il limite $F_{\alpha/2}$ può essere derivato dalle usuali tabelle.

Pertanto il criterio di accettazione di H_0 è essenzialmente (1.7.11); fissato α, e dunque anche $F_{\alpha/2}$, se la (1.7.11) è verificata H_0 è accettata, in caso contrario rifiutata.

Osservazione 1.7.2: se si sceglie il rapporto tra le varianze campionarie in modo che risulti sempre maggiore di 1, nel determinare F_α da usarsi in (1.7.11), occorre porre attenzione all'uso corretto dei gradi di libertà, nel caso che i due campioni abbiano numerosità diverse. Infatti F_α sarà il valore critico di F ad $N_X - 1$, $N_Y - 1$ gradi di libertà rispettivamente del numeratore e del denominatore, se

$$\frac{\overline{s}_X^2}{\overline{s}_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \geq 1 \ ; \tag{1.7.12}$$

al contrario, F_α andrà presa con $N_Y - 1$ ed $N_X - 1$ gradi di libertà di numeratore e denominatore se

$$\frac{\overline{s}_Y^2}{\overline{s}_X^2} \cdot \frac{\sigma_X^2}{\sigma_Y^2} \geq 1 \ . \tag{1.7.13}$$

Osservazione 1.7.3. il caso più tipico di test di confronto tra varianze campionarie è quello in cui si fa l'ipotesi

$$H_0 : \sigma_X^2 = \sigma_Y^2 \ , \tag{1.7.14}$$

ovvero $\sigma_X^2/\sigma_Y^2 = 1$. Questo test in particolare deve sempre essere applicato quando si voglia confrontare tra loro le medie empiriche di due campioni, usando la relazione (1.6.13). Infatti la (1.6.13) è valida solo nel caso che la (1.7.14) sia verificata.

1.8 Test sul coefficiente di correlazione

Sia data una variabile casuale normale doppia,

$$\begin{vmatrix} X \\ Y \end{vmatrix} = \mathcal{N} \left[\begin{vmatrix} \mu_X \\ \mu_Y \end{vmatrix} \ , \ \begin{vmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{vmatrix} \right] \ . \tag{1.8.1}$$

Se da tale variabile si estrae un campione di numerosità N, si potrà costruire il corrispondente spazio campionario (a 2N dimensioni) e le statistiche funzioni della variabile campionaria. Tra queste ha particolare interesse il coefficiente di correlazione campionario

$$R_{XY} = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X \mathcal{S}_Y} \quad [1] \tag{1.8.2}$$

Questo coefficiente ha una distribuzione che dipende solo da ρ, seppure in modo non semplice. In particolare è possibile trovare la distribuzione esplicita di R quando $\rho = 0$; mentre si riesce a trovare una distribuzione asintotica per una funzione di R, quando $\rho \neq 0$.

Più precsamente riportiamo, senza dimostrazone, il seguente teorema.

Teorema 1.8.1: nelle ipotesi fatte sulla variabile doppia
a) se $\rho = 0$

$$\frac{R}{\sqrt{1-R^2}}\sqrt{N-2} = t_{(N-2)} \tag{1.8.3}$$

[1] Si osservi che se al numeratore si usa la stima deviata così come al denominatore, si ha per R lo stesso risultato che si avrebbe usando per entrambi le stime corrette, cioè $R = \overline{\mathcal{S}}_{XY}/\overline{\mathcal{S}}_X\overline{\mathcal{S}}_Y$.

(t di Student a $N-2$ gradi di libertà),

b) se $\rho \neq 0$ per N grande si ha (in legge)

$$\frac{1}{2}\log\frac{1+R}{1-R} \cong \mathcal{N}\left[\frac{1}{2}\log\frac{1+\rho}{1-\rho}\ ,\ \frac{1}{N-3}\right]\ . \tag{1.8.4}$$

Le (1.8.3) e (1.8.4) sono relazioni che ci permettono di sottoporre a test ipotesi del tipo

$$H_0 : \rho = \rho_0\ ,$$

in quanto quando ρ_0 è fissato, la distribuzione di R (o meglio di una funzione di R) è fissata.

Conviene trattare separatamente i casi a) e b).

a) Si pone l'ipotesi fondamentale

$$H_0 : \rho = 0\ . \tag{1.8.5}$$

Pertanto se H_0 è corretta, il valore empirico

$$\frac{r}{\sqrt{1-r^2}}\sqrt{N-2} = t_0 \tag{1.8.6}$$

è una estrazione da una t di Student a $N-2$ gradi di libertà. Inoltre notiamo che per $r \to \pm 1,\ t_0 \to \pm\infty$, così che un valore alto di t_0 deriva da un valore alto di r, che può essere assunto come indicazione contraria a $\rho = 0$. Quindi useremo come intervallo di accettazione

$$\begin{aligned} &|t_0| \leq t_{\alpha/2} \\ &P(t_{(N-2)} \geq t_{\alpha/2}) = \alpha/2\ , \end{aligned} \tag{1.8.7}$$

con t_0 dato dalla (1.8.6).

Osservazione 1.8.1: il test di ipotesi (1.8.5), data la normalità della distribuzione, serve anche come test di indipendenza stocastica di X da Y.

b) Si pone l'ipotesi

$$H_0 : \rho = \rho_0 \ ,$$

nel qual caso il valore campionario

$$\frac{\sqrt{N-3}}{2} \log \frac{1+r}{1-r} \cdot \frac{1-\rho_0}{1+\rho_0} = Z_0 \ , \tag{1.8.8}$$

è una estrazione da una normale standardizzata.

Se si osserva il grafico di Z_0 come funzione di r, riportato qualitativamente in fig. 1.8.1, si vede che grandi valori positivi o negativi di Z_0 corrispondono a valori di r molto maggiori o minori di ρ_0, così che il test può essere eseguito con l'intervallo di accettazione

$$\begin{aligned} &|Z_0| \leq Z_{\alpha/2} \\ &[P(Z \geq Z_{\alpha/2}) = \alpha/2] \end{aligned} \tag{1.8.9}$$

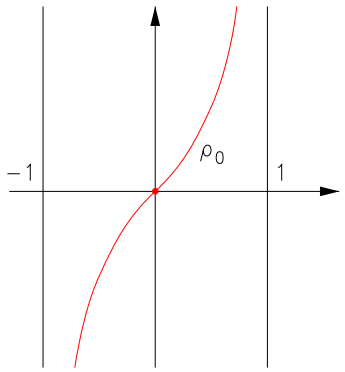


Figura 1.8.1:

Osservazione 1.8.2: notiamo che un test sul coefficiente di correlazione ha senso solo per campioni che hanno una certa numerosità, come si vede notando che se $H_0 : \rho = 0$ è vera, H_0 è accettata al livello $\alpha = 5\%$ se il coefficiente empirico soddisfa $|r| \leq r_{\alpha/2}$, secondo la seguente tabella

N	5	10	20	30	60	120
$\overline{r}_{\alpha/2}$	0,83	0,63	0,44	0,36	0,25	0,18

dunque ad esempio con un campione di 60 elementi si ha ancora il 5% di probabilità di ottenere un coefficiente di correlazione empirico di modulo maggiore del 25%.

1.9 Un test semplice di normalità

Abbiamo già menzionato nel paragrafo 1.4 il problema di sottoporre a verifica l'ipotesi che un certo campione sia tratto da una distribuzione che appartiene ad una famiglia parametrica. Nel paragrafo 1.4 abbiamo dato una soluzione di tipo generale al problema con due possibili test: quello di Kolmogorov e quello del χ^2.

Se ora facciamo l'ipotesi specifica di voler verificare l'appartenenza ad una distribuzione normale $H_0\{X \sim N[\mu, \sigma^2]\}$, è possibile istituire dei test che, sebbene assai semplici, sono in genere assai efficaci. Si noti che in questo caso entrambi μ e σ^2 hanno la funzione di parametri di disturbo; occorre quindi cercare una statistica che sia indipendente da tali parametri. Tali sono ad esempio i coefficienti campionari di skewness e di curtosi (cfr. Osservazione 1.5.2), che, come ricordato al paragrafo 9 del Quaderno n. 1, sono invarianti per trasformazioni lineari della variabile X.

Ricordando che, quando X è normale,

$$B = \frac{\overline{\mathcal{M}}_{(3)}}{\mathcal{S}^3} \sim \mathcal{N}\left[0, \frac{6}{N}\right]$$

e

$$\Gamma = \frac{\overline{\mathcal{M}}_{(4)}}{\mathcal{S}} \sim \mathcal{N}\left[3, \frac{24}{N}\right] \ ,$$

è facile costruire due test osservando che se H_0 è vera, allora

$$\frac{B}{\sqrt{\frac{6}{N}}} \sim Z \ , \ \frac{\Gamma - 3}{\sqrt{\frac{24}{N}}} \sim Z \ .$$

Inoltre un valore assoluto grande di B o di $\Gamma - 3$ può essere preso come indicazione che

$$\beta \sim E\{B\} \neq 0$$

oppure che

$$\gamma \sim E\{\Gamma\} \neq 3 \ ,$$

il che indicherebbe una non normalità della distribuzione.

Pertanto, fissato un livello di significatività α, ed il corrispondente valore critico $Z_{\alpha/2}$, per la normale standardizzata, si accetta H_0 se i valori empirici

$$b = \overline{m}_{(3)}/s^3 \quad , \quad g = \overline{m}_{(4)}/s^4$$

sono tali da verificare le relazioni

$$\left| \frac{b}{\sqrt{\frac{6}{N}}} \right| \leq Z_{\alpha/2} \tag{1.9.1}$$

$$\left| \frac{g-3}{\sqrt{\frac{24}{N}}} \right| \leq Z_{\alpha/2} \ : \tag{1.9.2}$$

in caso contrario l'ipotesi H_0 è rifiutata.

Osservazione 1.9.1: in realtà l'uso simultaneo di due test modifica la significatività α, in quanto si richiede che contemporaneamente siano verificate la (1.9.1) e (1.9.2): infatti se A_B è l'insieme in cui si verifica la (1.9.1)

$$P(\underline{X}^{(N)} \in A_B) = 1 - \alpha$$

ed anche se in A_Γ è verificata la (1.9.2)

$$P(\underline{X}^{(N)} \in A_\Gamma) = 1 - \alpha \ ,$$

perciò in generale (a meno che $P(A_B - A_\Gamma) + P(A_\Gamma - A_B) = 0$)

$$1 - \alpha \geq P(\underline{X}^{(N)} \in A_B \cap A_\Gamma) \geq 1 - 2\alpha \ ,$$

così che la vera significatività sta tra α e 2α.

In questo caso si può mostrare che essa è prossima a 2α.

Osservazione 1.9.2: si noti che, presi singolarmente, i due test (1.9.1) e (1.9.2) servono ad evidenziare due diversi tipi di deviazione rispetto ad una distribuzione normale: infatti l'indice di skewness servirà a mettere in evidenza particolari asimmetrie della distribuzione empirica, mentre l'indice di curtosi ci dirà se la distribuzione empirica tende ad avere delle code che sono più alte o più basse della normale, ovvero se la probabilità di ottenere valori più distanti dalla media di $\lambda\sigma$, con $\lambda \sim 3, 4, 5 \ldots$ è più alta o più bassa di quella normale.

1.10 La verifica di ipotesi, in presenza di ipotesi alternative

Fino ad ora abbiamo trattato il problema di verificare la plausibilità di un'ipotesi fondamentale H_0, sulla base di una statistica campionaria S

che intuitivamente tendesse ad assumere valori grandi, quando H_0 non fosse verificata. La distribuzione di S era nota per ipotesi solo quando H_0 era verificata, perciò restava indefinita la specifica maniera in cui si sospettava che H_0 fosse contraddetta: solo a posteriori si poteva giudicare che una certa S fosse utile a mettere in evidenza certe deviazioni rispetto ad altre. Vogliamo ora considerare il caso in cui si vuole verificare H_0 specificando in quale direzione può succedere che ci si allontani da H_0 stessa: ciò viene fatto stabilendo un'ipotesi alternativa H_A.

Tipicamente quando il test è prametrico e l'ipotesi fondamentale specifica

$$H_0 : \theta = \theta_0 \ , \tag{1.10.1}$$

si ha che l'ipotesi alternativa è data nella forma

$$H_A : \theta = \theta_A \ . \tag{1.10.2}$$

Come abbiamo già fatto per H_0, anche per H_A si distingue il caso in cui tale ipotesi sia semplice, come in (1.10.2), o quando sia composta, cioè quando si ha una famiglia di ipotesi di tipo (1.10.2).

Ad esempio è comune il caso in cui come alternativa si ponga $H_A : \theta = = \theta_A (\forall \theta_A > \theta_0)$, o più sinteticamente

$$H_A : \theta > \theta_0 \ . \tag{1.10.3}$$

In questo caso H_A indica che il senso di allontanamento sospettato da H_0 è in direzione $\theta > \theta_0$.

Osservazione 1.10.1: è importante capire che, contrariamente ai test di pura significatività, ora abbiamo a disposizione due distribuzioni

$$\begin{array}{ll} f_0(x) & \text{che vale se è vera } H_0 \\ f_A(x) & \text{che vale se è vera } H_A \ : \end{array}$$

questa aggiunta di informazione ci darà nuovi elementi di giudizio. Sottolineiamo anche che non è affatto necessario che f_0 ed f_A appartengano alla stessa famiglia parametrica: si veda l'Esempio 1.10.2.

Osservazione 1.10.2: per il momento le due ipotesi H_0 ed H_A non sono affatto sullo stesso piano poiché il problema che abbiamo posto non è di decidere tra due alternative, bensì di decidere se H_0 non è plausibile, rispetto ad un allontanamento in direzione H_A.

Esempio 1.10.1: sia $\underline{x}^{(N)}$ un campione tratto da una $\mathcal{N}[\mu, \sigma^2]$ con μ incognita e σ^2 nota: le ipotesi in alternativa sono

$$\begin{array}{l} H_0 : \mu = 0 \\ H_A : \mu = 1 \end{array} \qquad \text{(test con ipotesi alternativa semplice)} \tag{1.10.4}$$

oppure

$$\begin{array}{l} H_0 : \mu = 0 \\ H_A : \mu > 0 \ . \end{array} \qquad \text{(test con ipotesi alternativa composta)} \tag{1.10.5}$$

È chiaro che per verificare H_0 contro H_A potremo ancora usare la media empirica $\mathcal{M}$ e che potremo usare il concetto di livello di significatività α: tuttavia è chiaro che non si sceglierà un insieme di accettazione del tipo

$$\mu_0 - Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mathcal{M} \leq \mu_0 + Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \ ,$$

perché qualora si avesse un valore campionario

$$\mathcal{M} < \mu_0 - Z_{\alpha/2}\sigma/\sqrt{N}$$

questo non sarebbe affatto un'indicazione che ci siamo allontanati da H_0 in direzione H_A né per la (1.10.4) né per la (1.10.5).

Dunque il problema sarà, fissato α, trovare una forma conveniente dell'intervallo di accettazione di H_0.

Esempio 1.10.2: dato un campione $\underline{x}^{(N)}$ ci si chiede se tale campione è normalmente distribuito, (H_0), oppure se esso sia distribuito, a meno

di una trasformazione lineare, secondo una t di Student (H_A) con un opportuno numero ν di gradi di libertà. Si noti che in questo esempio H_0 è composta poiché $X \sim \mathcal{N}[\mu, \sigma^2]$ con (μ, σ^2) parametri di disturbo, e analogamente H_A è composta con μ, σ^2 e anche ν come parametri di disturbo.

Volendo togliere la dipendenza da μ e σ^2, diventa logico usare una statistica invariante per trasformazioni lineari. Nel caso in esempio è vantaggioso usare il coefficiente di curtosi: in particolare notiamo che

$$\gamma(t_\nu) = 3\frac{\nu - 2}{\nu - 4} \qquad (> 3) \ ,$$

così che le ipotesi in alternativa diventano

$$H_0 : \gamma = 3 \quad , \quad H_A : \gamma > 3 \ . \tag{1.10.6}$$

Ancora una volta il problema sarà di trovare un intervallo di accettazione di H_0, fissato il livello di significatività α ed usando H_A per definire la forma più vantaggiosa dell'intervallo.

Si noti che in questo esempio H_0 ed H_A si riferiscono a famiglie parametriche diverse.

Per comprendere come l'ipotesi alternativa, ovvero la distribuzione $f_A(x)$ e la corrispondente likelihood $L_A(\underline{x})$, possano influenzare la scelta dell'intervallo di accettazione, osserviamo che finora i test sono stati disegnati con il seguente criterio: fissata H_0 e una statistica S, fissato anche un livello di significatività α, si è cercato un intervallo I_α tale che

$$P\{S \in I_\alpha | H_0\} = 1 - \alpha \ , \tag{1.10.7}$$

e si è deciso sulla base del "buon senso" che $S \in I_\alpha^c$ fornisse un'evidenza contro l'ipotesi H_0. Notiamo che la condizione $S \in I_\alpha$, può essere tradotta nello spazio campionario $R^{(N)}$, definendo l'insieme $\Omega_{0,\alpha}$, tale che

$$\underline{x} \in \Omega_{0,\alpha} \leftrightarrow S \in I_\alpha \ ,$$

perciò la (1.10.7) può essere sostituita dalla condizione

$$P\{\underline{X} \in \Omega_{0,\alpha} | H_0\} = 1 - \alpha \ . \tag{1.10.8}$$

Naturalmente per ogni α fissato esistono molte possibili regioni del tipo $\Omega_{0,\alpha}$: ad esempio (cfr. Esempio 1.10.1), se $X \in N[\mu, \sigma^2]$, σ^2 noto ed $S = \mathcal{M}$, I_α è dato da

$$P\{\mathcal{M} \in I_\alpha | \mu = \mu_0\} = 1 - \alpha \ ,$$

ovvero, con $\alpha_1 + \alpha_2 = \alpha$, (cfr. fig. 1.10.1)

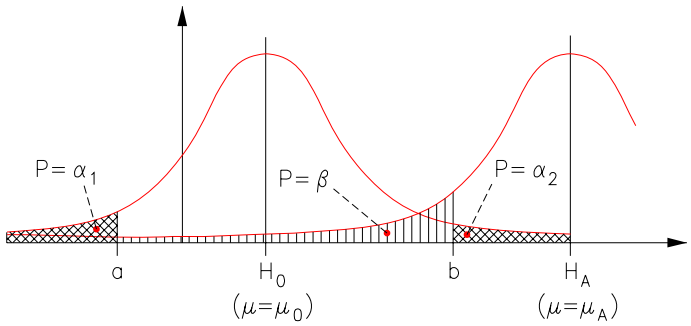


Figura 1.10.1:

$$\mu_0 - \sigma Z_{\alpha_1} \leq \mathcal{M} \leq \mu_0 + \sigma Z_{\alpha_2} \tag{1.10.9}$$

Ora supponiamo che sia definita l'ipotesi H_A e supponiamo ad esempio che essa sia semplice,

$$H_A : \mu = \mu_A \qquad (\mu_A > \mu_0) \ . \tag{1.10.10}$$

Fissata H_A, cioè $L_A(\underline{x})$, si potrà conoscere la distribuzione di S sotto H_A, e così si potrà calcolare

$$\beta = P\{S \in I_\alpha | H_A\} = P\{\underline{X} \in \Omega_{0,\alpha} | H_A\} \ . \tag{1.10.11}$$

Questo parametro β, che dipende dalla scelta delle famiglie $\Omega_{0,\alpha}$, rappresenta la probabilità che venga accettata H_0 (perché $\underline{x} \in \Omega_{0,\alpha}$) quando invece è vera H_A.

Il valore $1 - \beta$ è detto potenza del test, al livello di significatività α,

$$1 - \beta = \text{pow } \{H_A, \Omega_{0,\alpha} H_0\} \ , \tag{1.10.12}$$

ed è la probabilità di rifiutare H_0 quando in effetti H_A è vera.

È chiaro che un test sarà tanto più efficiente nel mettere in evidenza una deviazione da H_0 in direzione H_A, quanto più grande sarà la potenza (1.10.12).

Esempio 1.10.3: (soluzione del problema dell'Esempio 1.10.1). Sia $X \sim \mathcal{N}[\mu, \sigma^2]$ (σ^2 nota), e sia I_α la famiglia di intervalli (1.10.9) per la media campionaria, definita per ogni $0 \leq \alpha_1 \leq \alpha$. La potenza, rispetto all'ipotesi $H_A : \mu = \mu_A$, dell'intervallo $I_\alpha(\alpha_1)$ può essere trovata dalle seguenti relazioni:

$$\begin{aligned} P\{\mathcal{M} \leq a | H_0\} = \alpha_1 \quad &\rightarrow \quad a = \mu_0 - Z_{\alpha_1} \sigma / \sqrt{N} \\ P\{\mathcal{M} \geq b | H_0\} = \alpha_2 \quad &\rightarrow \quad b = \mu_0 + Z_{\alpha_2} \sigma / \sqrt{N} \end{aligned}$$

$$\alpha_1 + \alpha_2 = \alpha \qquad F_Z(Z_{\alpha_1}) = 1 - \alpha_1 \qquad , \qquad F_Z(Z_{\alpha_2}) = 1 - \alpha_2 \ ,$$

Z_{α_1} funzione monotona decrescente di α_1 $(\alpha_1 \to 0,\ Z_{\alpha_1} \to +\infty)$
Z_{α_2} funzione monotona crescente di α_1 $(\alpha_1 \to 0,\ Z_{\alpha_2} \to Z_\alpha)$

$$\begin{aligned} P\{\mathcal{M} \le a|H_A\} &= P\{Z \le \frac{a-\mu_A}{\sigma/\sqrt{N}}\} = P\left\{Z \le -Z_{\alpha_1} - \frac{\mu_A-\mu_0}{\sigma/\sqrt{N}}\right\} = \\ &= 1 - F_Z\left[Z_{\alpha_1} + \frac{\mu_A-\mu_0}{\sigma/\sqrt{N}}\right] \\ P\{\mathcal{M} \ge b|H_A\} &= P\{Z \ge \frac{b-\mu_A}{\sigma/\sqrt{N}}\} = P\{Z \ge Z_{\alpha_2} - \frac{\mu_A-\mu_0}{\sigma/\sqrt{N}} = \\ &= 1 - F_Z\left[Z_{\alpha_2} - \frac{\mu_A-\mu_0}{\sigma/\sqrt{N}}\right] \ . \end{aligned}$$

Ora, chiamando

$$c = \frac{\mu_A-\mu_0}{\sigma/\sqrt{N}} > 0 \ ,$$

si osserva che, per definizione, la potenza del test, contro l'ipotesi H_A, è, secondo le (1.10.11), (1.10.12),

$$\begin{aligned} 1-\beta &= 1 - F_Z(Z_{\alpha_1}+c) + 1 - F_Z(Z_{\alpha_2}-c) = \\ &= 1 - F_Z(Z_{\alpha_1}) + 1 - F_Z(Z_{\alpha_2}) + \\ &+ [F_Z(Z_{\alpha_2}) - F_Z(Z_{\alpha_2}-c)] - [F_Z(Z_{\alpha_1}+c) - F_Z(Z_{\alpha_1})] \ . \end{aligned}$$

Ma $1 - F_Z(Z_{\alpha_1}) = \alpha_1$ e $1 - F_Z(Z_{\alpha_2}) = \alpha_2$, così che la somma dei primi due termini è α, costante; inoltre i termini in parentesi quadra sono positivi perché la $F_Z(z)$ è monotona. Perciò si tratta di muovere α_1, e di conseguenza α_2, in modo da rendere $F_Z(Z_{\alpha_1}+c) - F_Z(Z_{\alpha_1})$ il più piccolo possibile e $F_Z(Z_{\alpha_2}) - F_Z(Z_{\alpha_2}-c)$ il più grande possibile, cioè di avere Z_{α_1} molto grande e Z_{α_2} quanto più piccolo si può.

Ciò avviene quando $\alpha_1 \to 0,\ \alpha_2 \to \alpha,\ Z_{\alpha_1} \to +\infty$, perché

$$\begin{aligned} &F_Z(Z_{\alpha_1}+c) - F_Z(Z_{\alpha_1}) \to 0 \\ &F_Z(Z_{\alpha_2}) - F_Z(Z_{\alpha_2}-c) \to F_Z(Z_\alpha) - F_Z(Z_\alpha - c) \ . \end{aligned}$$

Il risultato è perciò che l'esigenza di massimizzare la potenza, cioè di rendere minimo il rischio β di accettare H_0 quando H_A è vera, ci porta automaticamente a scegliere l'intervallo di accettazione H_0, contro H_A, nella forma ad una coda

$$-\infty < \mathcal{M} \leq b = \mu_0 + Z_\alpha \sigma/\sqrt{N} \tag{1.10.13}$$

Osservazione 1.10.3: si noti che nel determinare l'intervallo di accettazione l'unica cosa che conta è che $\mu_A > \mu_0$, cioè che $c > 0$, da cui dipende il fatto che si sceglie un intervallo di accettazione non limitato inferiormente e con un limite superiore b (vedi (1.10.13)) indipendente dallo specifico valore di μ_A. È chiaro che se si avesse avuta l'alternativa $H_A(\mu_A < \mu_0)$, per simmetria l'intervallo di accettazione sarebbe diventato semplicemente

$$\mu_0 - Z_\alpha \sigma/\sqrt{N} < \mathcal{M} < +\infty \ .$$

1.11 Il lemma di Neyman-Pearson per alternative semplici. Test uniformemente più potenti

Riprendiamo il problema definito nel paragrafo precedente, ed esemplificato nell'Esempio 1.10.3, proponendolo dapprima nella seguente forma (alternativa semplice): data una variabile campionaria $\underline{X}^{(N)}$ e fissato un livello di significatività α, si considerano due ipotesi semplici H_0 (ipotesi fondamentale) ed H_A (ipotesi alternativa), inoltre si divide lo spazio campionario R^N in due insiemi $\Omega_{0,\alpha}$ e $\Omega_{A,\alpha} = \Omega^c_{0,\alpha}$ tali che

$$P\{\underline{X} \in \Omega_{0,\alpha} | H_0\} = 1 - \alpha \tag{1.11.1}$$

$$\Omega_{0,\alpha} \cup \Omega_{A,\alpha} = R^N \ ; \tag{1.11.2}$$

la suddivisione di R^N in due specifici $\Omega_{0,\alpha}, \Omega_{A,\alpha}$ con le caratteristiche (1.11.1) e (1.11.2) è detta anche "disegno" del test. In un certo disegno fissato, $\Omega_{0,\alpha}$ è la regione di accettazione di H_0, mentre $\Omega_{A,\alpha}$, la regione di accettazione di H_A, deve essere una regione critica di H_0 di grandezza α, cioè

$$P\{\underline{X} \in \Omega_{A,\alpha}|H_0\} = \alpha \ . \tag{1.11.3}$$

Per ogni α esistono molti possibili disegni del test e si vuole trovare quello ottimale, nel senso che sia massima la potenza del test, ovvero

$$P\{\underline{X} \in \Omega_{A,\alpha}|H_A\} = 1 - \beta = \max \ . \tag{1.11.4}$$

Il problema è risolto da un teorema, noto come *lemma di Neyman-Pearson*, che afferma quanto segue.

Teorema 1.11.1: per ogni α la regione critica $\Omega_{A,\alpha}$, ottimale, ovvero soddisfacente (1.11.4), ha la forma

$$\mathrm{lr}_{A0}(\underline{x}) = \frac{L_A(\underline{x})}{L_0(\underline{x})} \geq c_\alpha \ , \tag{1.11.5}$$

per quel valore di c_α, se esiste, per cui è verificata la (1.11.3).

La funzione $\mathrm{lr}_{A0}(\underline{x})$ si chiama *likelihood ratio* (rapporto di verosimiglianza) e rappresenta appunto il rapporto tra la likelihood secondo H_A e quella secondo H_0, valutata nel punto $\underline{x}$.

In effetti sia $\overline{\Omega}_{A,\alpha}$ la zona critica della forma (1.11.5), dove la costante c_α è determinata della condizione (1.11.3) e sia $\Omega_{A,\alpha}$ una qualsiasi altra regione critica di grandezza α: allora

$$\begin{aligned} \alpha &= P\{\underline{X} \in \overline{\Omega}_{A,\alpha}|H_0\} = \int_{\overline{\Omega}_{A,\alpha}} L_0(\underline{x})d\underline{x} = \\ &= P\{\underline{X} \in \Omega_{A,\alpha}|H_0\} = \int_{\Omega_{A,\alpha}} L_0(\underline{x})d\underline{x} \ . \end{aligned} \tag{1.11.6}$$

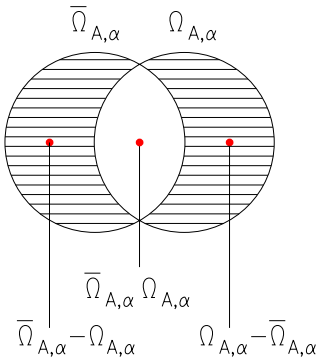

$\overline{\Omega}_{A,\alpha}$
$\Omega_{A,\alpha}$
$\overline{\Omega}_{A,\alpha}$ $\Omega_{A,\alpha}$
$\overline{\Omega}_{A,\alpha} - \Omega_{A,\alpha}$
$\Omega_{A,\alpha} - \overline{\Omega}_{A,\alpha}$

Figura 1.11.1:

Notando che valgono le relazioni (cfr. fig. 1.11.1)

$$\begin{cases} \overline{\Omega}_{A,\alpha} = \overline{\Omega}_{A,\alpha} \cdot \Omega_{A,\alpha} + (\overline{\Omega}_{A,\alpha} - \Omega_{A,\alpha}) \\ \Omega_{A,\alpha} = \overline{\Omega}_{A,\alpha} \cdot \Omega_{A,\alpha} + (\Omega_{A,\alpha} - \overline{\Omega}_{A,\alpha}) \ , \end{cases} \tag{1.11.7}$$

si vede che dalla (1.11.6) si ha

$$\int_{\overline{\Omega}_{A,\alpha} - \Omega_{A,\alpha}} L_0(\underline{x}) d\underline{x} = \int_{\Omega_{A,\alpha} - \overline{\Omega}_{A,\alpha}} L_0(\underline{x}) d\underline{x} \ . \tag{1.11.8}$$

D'altro canto

$$\begin{array}{lll} \text{in} & [\overline{\Omega}_{A,\alpha} - \Omega_{A,\alpha}] \subset \overline{\Omega}_{A,\alpha} & \text{si ha} \quad L_A \geq c_\alpha L_0 \\ \text{in} & [\Omega_{A,\alpha} - \overline{\Omega}_{A,\alpha}] \subset [\overline{\Omega}_{A,\alpha}]^c & \text{si ha} \quad L_A \leq c_\alpha L_0 \ , \end{array}$$

così che

$$\int_{\overline{\Omega}_{A,\alpha} - \Omega_{A,\alpha}} L_A(\underline{x}) d\underline{x} \geq \int_{\Omega_{A,\alpha} - \overline{\Omega}_{A,\alpha}} L_A(\underline{x}) d\underline{x} \ . \tag{1.11.9}$$

Usando ancora le (1.11.7), le (1.11.9) danno

$$\int_{\overline{\Omega}_{A,\alpha}} L_A(\underline{x}) d\underline{x} \geq \int_{\Omega_{A,\alpha}} L_A(\underline{x}) d\underline{x}$$

ovvero

$$P\{\underline{X} \in \overline{\Omega}_{A,\alpha} | H_A\} \geq P\{\underline{X} \in \Omega_{A,\alpha} | H_A\} \tag{1.11.10}$$

il che dimostra appunto che la potenza di $\overline{\Omega}_{A,\alpha}$ è maggiore di quella di $\Omega_{A,\alpha}$.

Esempio 1.11.1: riprendiamo l'Esempio 1.10.3, ritrovandone la soluzione mediante il teorema appena dimostrato. Sia dunque

$$\begin{array}{ll} H_0 & X = \mathcal{N}[\mu_0, \sigma^2] \quad (\sigma^2 \text{ nota}) \\ H_A & X = \mathcal{N}[\mu_A, \sigma^2] \quad (\sigma^2 \text{ nota}) \end{array}$$

con $\mu_A > \mu_0$.

Formata la likelihood ratio si vede che

$$\begin{aligned} \mathrm{lr}_{A0}(\underline{x}) \quad &= \frac{\frac{1}{(\sqrt{2\pi}\sigma)^N} \exp\left\{-\frac{\sum(x_i-\mu_A)^2}{2\sigma^2}\right\}}{\frac{1}{(\sqrt{2\pi}\sigma)^N} \exp\left\{-\frac{\sum(x_i-\mu_0)^2}{2\sigma^2}\right\}} = \\ &= \exp\left\{\frac{2N\mathcal{M}(\mu_A-\mu_0) - N(\mu_A^2-\mu_0^2)}{2\sigma^2}\right\} \ . \end{aligned}$$

Ne segue che la regione $\overline{\Omega}_{A,\alpha}$, data dalla condizione (1.11.5), deve corrispondere a

$$\mathcal{M}(\mu_A - \mu_0) \geq \text{ (costante opportuna)}$$

ovvero

$$\mathcal{M} \geq b_\alpha \ ,$$

perché $\mu_A > \mu_0$.

Ora la costante b_α va determinata imponendo

$$P\{\mathcal{M} \geq b_\alpha | H_0\} = \alpha \ ,$$

ovvero

$$\frac{b_\alpha - \mu_0}{\sigma/\sqrt{N}} = Z_\alpha \ , \qquad (1.11.11)$$

che coincide proprio con la (1.10.13).

Osservazione 1.11.1: come si vede dalla (1.11.1), la zona critica, $\mathcal{M} \geq \mu_0 + Z_\alpha(\sigma/\sqrt{N})$, non dipende da μ_A, se non per il segno della disuguaglianza, e dunque essa è valida anche per una H_A composta, $H_A(\mu_A > \mu_0)$. Al contrario, tale zona dipende da σ, perciò essa non costituirebbe la soluzione nel caso considerassimo tanto H_0 quanto H_A composte, con σ come parametro di disturbo.

Osservazione 1.11.2: quando, come nell'Esempio 1.11.1, la zona critica ottimale $\Omega_{A,\alpha}$ non dipende dalla particolare ipotesi alternativa $H_A(\theta = \theta_A)$, si dice che il test trovato è *uniformemente più potente.*

In generale ricordando la definizione (1.10.12) si vede che ad ogni test di ipotesi $H_0(\theta = \theta_0)$, definito tramite le sue zone di accettazione $\Omega_{0,\alpha}$ e le sue zone critiche $\Omega_{A,\alpha} = [\Omega_{0,\alpha}]^c$, si può associare, una volta definita un'alternativa $H_A(\theta = \theta_A)$, la funzione di potenza

$$P\{\Omega_{A,\alpha}|H_A\} = 1 - \beta = \text{pow } (\theta_A, \alpha) \ . \tag{1.11.12}$$

Questa funzione dipende in generale dalle due variabili θ_A e α, ma è di solito considerata per vari valori fissati di α come funzione di θ_A.

Ad esempio, riprendendo il caso dell'Esempio 1.11.1, dove le zone critiche erano definite da

$$\begin{aligned} &\Omega_{0,\alpha} \to m \leq b_\alpha = \mu_0 + Z_\alpha \sigma/\sqrt{N} \\ &\Omega_{A,\alpha} \to m > b_\alpha = \mu_0 + Z_\alpha \sigma/\sqrt{N} \ , \end{aligned}$$

si vede che

$$\begin{aligned} &\text{pow } (\mu_A, \alpha) = P\{\mathcal{M} > b_\alpha | \mu = \mu_A\} = \qquad (1.11.13) \\ = \quad &P\left\{ \frac{\mathcal{M} - \mu_A}{\sigma/\sqrt{N}} > -\frac{\mu_A - \mu_0}{\sigma/\sqrt{N}} + Z_\alpha \right\} = 1 - F_Z\left\{ Z_\alpha - \frac{\mu_A - \mu_0}{\sigma/\sqrt{N}} \right\} \ . \end{aligned}$$

Naturalmente questa funzione è definita per $\mu_A > \mu_0$ e, come è ovvio, pow $(\mu_0, \alpha) = 1 - F_Z(Z_\alpha) = \alpha$. La situazione è rappresentata graficamente in fig. 1.11.2.

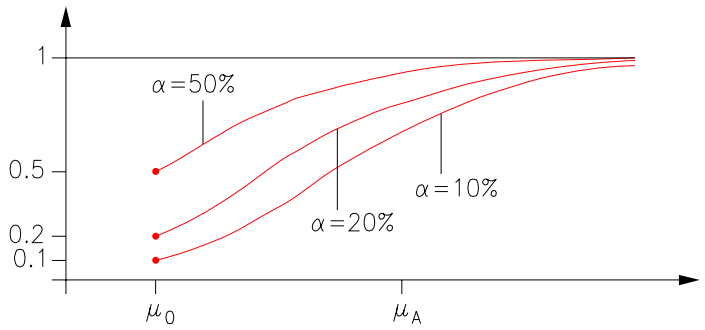


Figura 1.11.2: Pow (μ_A, α).

Il possibile uso della funzione di potenza sta nel giudicare il rendimento di due diversi test, rispetto alla massimizzazione della potenza: infatti tra due test preferiamo sempre quello che ha, a parità di α, funzione di potenza maggiore.

Osservazione 1.11.3: ci si potrebbe chiedere, a prima vista, come mai per ogni α fissato non si applichi il lemma di Neyman-Pearson e non si trovi l'insieme critico ottimale $\overline{\Omega}_{A,\alpha}$ per ogni valore di θ_A. Ma si noti che questo $\overline{\Omega}_{A,\alpha}$ per l'appunto dipende da θ e quindi non può fornire una unica regione critica per H_0 : solo quando $\overline{\Omega}_{A,\alpha}$ è lo stesso per tutti i θ_A, si può pensare che viene costruito un vero test per H_0 e si ha allora il test uniformemente più potente.

Osservazione 1.11.4: come si sarà notato fin qui si sono considerate prevalentemente delle ipotesi alternative unilaterali, come $\mu_A > \mu_0$ per l'Esempio 1.11.1, per le quali è naturale trovare delle zone critiche ad una sola coda come la (1.11.5).

Ci si può chiedere se non si dia il caso in cui l'ipotesi alternativa H_A sia della forma

$$H_A : \theta_A \neq \theta_0 \quad , \quad X \sim f_X(x;\theta) \ . \tag{1.11.14}$$

Si può osservare che in effetti questo caso è diverso da quello dei test di pura significatività, in quanto seppure H_A non specifica il valore di θ (ipotesi composta), specifica però il fatto che X appartiene alla stessa famiglia parametrica (1.11.14). Così l'Esempio 1.11.1 si potrebbe porre

$$\begin{array}{lll} x = \mathcal{N}[\mu, \sigma^2] & & (\sigma^2 \text{ nota}) \\ H_0 & \mu = \mu_0 & \\ H_A & \mu \neq \mu_0 \ . & \end{array} \tag{1.11.15}$$

Pensando anche solo a questo esempio si comprende però che non si può più invocare un criterio di massima potenza, perché l'esigenza di massimizzare la potenza per $\mu > \mu_0$ porta a un insieme d'accettazione di tipo $(1/N)\sum X_i \leq \mu_0 + c_\alpha$, mentre per $\mu < \mu_0$ si trova all'opposto $(1/N)\sum X_i \geq \mu_0 - c_\alpha$. In casi come questi in cui vale una regola di simmetria della variabile campionaria su cui è basato il test, un ragionevole

compromesso è assumere una zona di accettazione simmetrica, dividendo il rischio α in due parti, ognuna di probabilità $\alpha/2$, corrispondenti alle due code che indicano le due possibili direzioni di allontanamento da H_0. Così nell'Esempio 1.11.1 si prenderebbe, per verificare (1.11.15), un intervallo

$$|(1/N)\sum X_i - \mu_0| \leq c_{\alpha/2} \ ,$$

determinando poi $c_{\alpha/2}$ in base alla condizione

$$P\{(1/N)\sum X_i \geq \mu_0 + c_{\alpha/2}|H_0\} = \alpha/2 \ .$$

Come si vede si torna così in sostanza agli stessi test che avevamo catalogato come test di pura significatività.

1.12 Test con ipotesi alternativa e con parametri di disturbo

Vi sono molti importanti casi in cui tanto H_0 quanto H_A sono ipotesi composte per la presenza di un parametro di disturbo, di valore non specificato.

La teoria si fa qui più complessa e noi svilupperemo completamente solo un esempio significativo cercando di coglierne gli aspetti generali.

Esempio 1.12.1: sia $\underline{X}^{(N)}$ la variabile campionaria di un campione bernoulliano, normale

$$X = \mathcal{N}[\mu, \sigma^2] \tag{1.12.1}$$

e si voglia sottoporre a test l'alternativa tra ipotesi

$$\begin{aligned} H_0 \quad & \mu = \mu_0 \\ H_A \quad & \mu = \mu_A > \mu_0 \ , \end{aligned} \tag{1.12.2}$$

senza conoscere il valore di σ^2 (parametro di disturbo).

L'idea chiave è quella di andare a condizionare la likelihood ratio a superfici tali che, almeno sotto H_0, la distribuzione di $\underline{X}$ sia indipendente dal parametro di disturbo, σ^2 : questo è possibile farlo cercando una statistica Q, sufficiente per σ^2, per cui sulla superficie $Q = q$

$$L_{|Q=q}(\underline{x}; \mu_0, \sigma^2) = K(\underline{x}; \mu_0, q) \tag{1.12.3}$$

indipendentemente da σ^2, per definizione di sufficienza.

Se ora, condizionato a $Q = q$, possiamo trovare un test (ottimale secondo il lemma di Neyman-Pearson) per l'alternativa (1.12.2) e se per caso tale test non dipende da q, siamo arrivati ad una soluzione accettabile proprio perché valida indipendentemente dal valore dato alla variabile condizionante: se per di più il test non dipende dallo specifico valore alternativo μ_A, si dice che si ha un test similare uniformemente più potente.

Limitandoci all'Esempio 1.12.1, si può notare che sotto H_0

$$\begin{aligned} L_0(\underline{x}) &= c \cdot \exp[-1/(2\sigma^2)|\underline{x} - \mu_0 \underline{e}|^2] \, , \\ (\underline{e}^+ &= [1\ 1\ 1\ \ldots\ 1] \quad ; \quad |\underline{e}| = \sqrt{N}) \end{aligned}$$

così che la statistica sufficiente per σ^2 è

$$Q = |\underline{X} - \mu_0 \underline{e}|^2 \ . \tag{1.12.4}$$

La superficie $Q = q$ è perciò una sfera dello spazio campionario, di centro $\mu_0 \underline{e}$ e di raggio $\sqrt{q}$ (cfr. fig. 1.12.1).

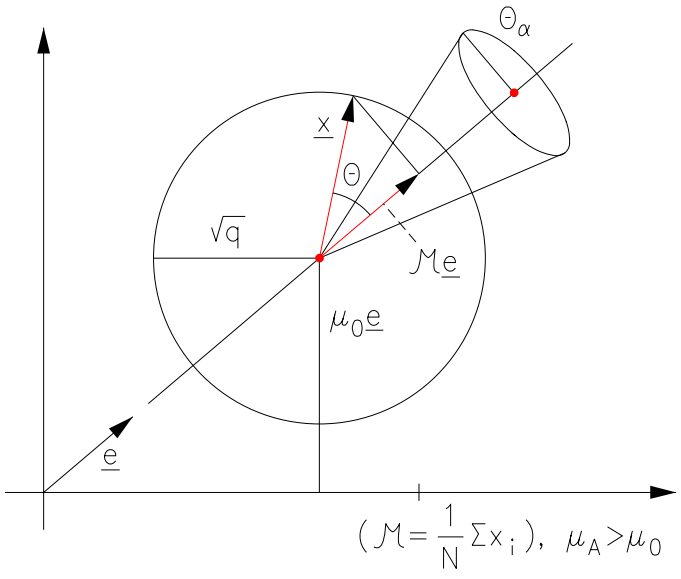

Θ_α
$\underline{x}$
Θ
$\sqrt{q}$
$\mathcal{M}\underline{e}$
$\mu_0\underline{e}$
$\underline{e}$
$(\mathcal{M}=\frac{1}{N}\Sigma x_i),\ \mu_A>\mu_0$

Figura 1.12.1:

Come si vede, su tale sfera, $L_0 = \text{cost}$, cioè la distribuzione sotto l'ipotesi H_0 è uniforme e quindi la distribuzione condizionata dipende solo da q e non da σ^2.

Valutiamo ora $L_A(\underline{x})$ sulla stessa sfera.

A tal scopo notiamo che (cfr. fig. 1.12.1)

$$\begin{aligned} |\underline{x} - \mu_A \underline{e}|^2 &= |\underline{x} - \mu_0 \underline{e}|^2 + 2(\underline{x} - \mu_0 \underline{e}) \cdot \underline{e}(\mu_0 - \mu_A) + (\mu_0 - \mu_A)^2 |\underline{e}|^2 = \\ &= q^2 + 2\sqrt{N}\sqrt{q}\cos\theta(\mu_0 - \mu_A) + N(\mu_0 - \mu_A)^2 \end{aligned}$$

dove θ è l'angolo tra il vettore $\underline{x} - \mu_0 \underline{e}$ ed il vettore $\underline{e}$, così che sulla nostra sfera

$$\text{lr}_{A,0}(\underline{x}|Q = q) = \frac{L_A(\underline{x}|Q = q)}{L_0(\underline{x}|Q = q)} \sim \exp\left\{ \frac{\sqrt{N}\sqrt{q}}{\sigma^2}(\mu_A - \mu_0)\cos\theta \right\} . \tag{1.12.5}$$

Ne deriva che l'insieme critico ottimale vincolato alla sfera $Q = q$ e definito dalla relazione

$$\text{lr}_{A,0}(\underline{x}|Q = q) \geq c_\alpha ,$$

per la (1.12.5), considerato che $\mu_A > \mu_0$, è dato dalla calotta

$$\theta \leq \theta_\alpha , \tag{1.12.6}$$

per θ_α opportuno. Poiché la distribuzione di θ è ovviamente indipendente, sotto H_0 (cioè quando $\mu = \mu_0$), tanto da μ_A quanto da q, siamo in presenza di un test similare ed uniformemente più potente. Le zone critiche, date dalla (1.12.6), sono nello spazio campionario dei coni di vertice $\mu_0 \underline{e}$ e di semiapertura θ_α : se il campione cade in una tale zona, si rifiuta H_0, in caso contrario H_0 è accettata.

Occorre ora trovare la distribuzione di θ o di una sua funzione.

A tale scopo basta osservare che (cfr. fig. 1.12.1)

$$\begin{aligned} |\cot\theta| &= \frac{|(\mathcal{M}-\mu_0)|\sqrt{N}}{|\underline{X}-\mathcal{M}\underline{e}|} = \\ &= \frac{|\mathcal{M}-\mu_0|}{\frac{\sqrt{\sum(X_i-\mathcal{M})^2}}{\sqrt{N}}} = \\ &= \frac{1}{\sqrt{N-1}}\frac{|\mathcal{M}-\mu_0|}{\frac{\overline{s}}{\sqrt{N}}} \quad : \end{aligned}$$

pertanto, ricordando il Teorema 1.5.1, si vede che

$$\sqrt{N-1}\cot\theta = t_{(N-1)} \qquad (t \text{ di Student a } N-1 \text{ gradi di libertà}). \tag{1.12.7}$$

Di conseguenza l'insieme critico

$$\theta \leq \theta_\alpha$$

può essere riscritto come

$$\sqrt{N-1}\cot\theta \geq \sqrt{N-1}\cot\theta_\alpha = T_{(N-1),\alpha} \ ,$$

ovvero

$$\frac{m-\mu_0}{\frac{\overline{s}}{\sqrt{N}}} \geq t_{(N-1),\alpha} \tag{1.12.8}$$

che corrisponde, fatto su una sola coda, al test di pura significatività visto nel paragrafo 1.6.

1.13 Test localmente più potenti

Proseguiamo l'analisi della costruzione di test parametrici. Finora abbiamo visto l'utilità dell'uso del concetto di potenza ed abbiamo definito i test uniformemente più potenti come quelli che rendono massima la potenza rispetto a tutti i valori assunti dal parametro sotto l'ipotesi alternativa H_A. I test di questo tipo, però, sono più l'eccezione che la regola, proprio perché la richiesta che una zona critica sia ottimale per tutti i valori θ_A è molto forte. Ricordiamo anche che un modo grafico per valutare due test è quello di paragonare le loro curve di potenza, a parità di α, naturalmente. Il test che ha la curva più alta è preferibile.

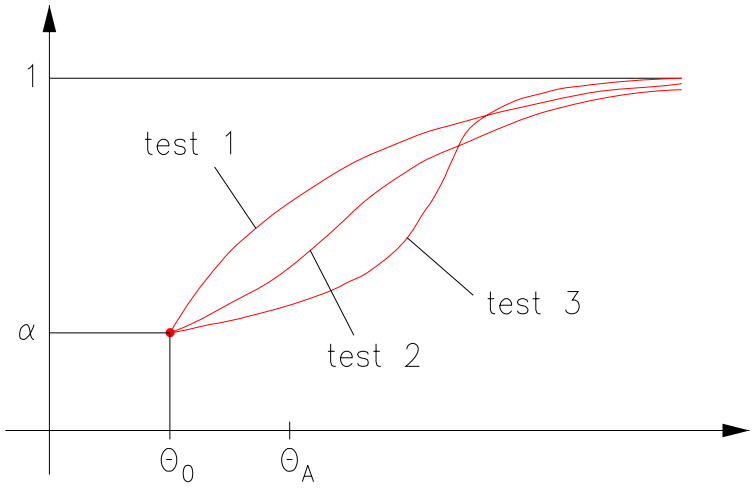


Figura 1.13.1:

In fig. 1.13.1, ad esempio, si vede chiaramente che il test 1 è preferibile al test 2, ma la situazione è diversa per il test 3 in quanto la curva di potenza è in parte sotto e in parte sopra quella del test 1.

In molti casi tuttavia i valori alternativi θ_A che più ci interessano, e quindi rispetto cui si vuole massimizzare la potenza, sono proprio quelli più vicini a θ_0 : in questo senso possiamo dire che il test 1 è per lo meno localmente (attorno a θ_0) migliore del test 3.

Si noti che, com'è anche evidente in fig. 1.13.1, per tutte le curve di potenza è sempre

$$\text{pow } (\theta_0, \alpha) = P\{X \in \Omega_{A,\alpha} | \theta_A = \theta_0\} = \alpha \ , \tag{1.13.1}$$

così che, per α fissato, tutte le curve di potenza devono partire dal punto (θ, α).

Cerchiamo ora di formalizzare questo ragionamento grafico. Se ci limitiamo ad una ottimizzazione di tipo locale possiamo dire che cerchiamo quella curva di potenza

$$\text{pow } (\theta_A; \alpha) = P\{\underline{X} \in \Omega_{A,\alpha} | \theta = \theta_A\} = \int_{\Omega_{A,\alpha}} L(\underline{x}; \theta_A) d\underline{x} \ , \tag{1.13.2}$$

che passa per (θ_0, α), ovvero soddisfa (1.13.1), per cui è massima la pendenza in θ_0

$$\frac{\partial}{\partial \theta_A} \text{pow } (\theta_A, \alpha)|_{\theta_A = \theta_0} = \int_{\Omega_{A,\alpha}} \frac{\partial L(\underline{x}, \theta_0)}{\partial \theta} d\underline{x} = \max \ . \tag{1.13.3}$$

In definitiva si tratta di trovare $\Omega_{A,\alpha}$ che rende massimo il funzionale (1.13.3), sotto la condizione

$$\int_{\Omega_{A,\alpha}} L(\underline{x}, \theta_0) d\underline{x} = \alpha \ . \tag{1.13.4}$$

Questo problema è formalmente identico a quello risolto dal lemma di Neyman-Pearson e quindi ha la stessa soluzione, cioè $\Omega_{A,\alpha}$ è l'insieme in cui vale la relazione

$$\frac{\frac{\partial}{\partial\theta}L(\underline{x};\theta_0)}{L(\underline{x};\theta_0)} \geq c_\alpha \ , \tag{1.13.5}$$

dove c_α è definito poi dalla condizione (1.13.4).

Si può notare anche che la condizione (1.13.5), introducendo la variabile (cfr. Quaderno n. 2, paragrafo 1.7)

$$U(\underline{x};\theta) = \partial_\theta \log L(\underline{x};\theta)$$

può essere riscritta come

$$U(\underline{x};\theta_0) \geq c_\alpha \ . \tag{1.13.6}$$

Il richiamo non è puramente formale, in quanto la v.c. $U(\underline{X};\theta_0)$, sotto l'ipotesi $\theta = \theta_0$, ha una serie di proprietà note: in particolare ricordiamo che (cfr. Quaderno n. 2, paragrafo 1.7)

$$\begin{aligned} &E\{U(\underline{X},\theta_0)|\theta=\theta_0\} = 0 \\ &\sigma^2\{U|\theta=\theta_0\} = -E\{\partial_\theta U(\underline{X};\theta_0)|\theta=\theta_0\} = \mathcal{I}(\theta_0) \ . \end{aligned} \tag{1.13.7}$$

Inoltre, per campioni numerosi, sappiamo che, per il teorema centrale, U è asintoticamente normale

$$U \sim \mathcal{N}[0, \mathcal{I}(\theta_0)] \tag{1.13.8}$$

così che, seppure in forma approssimata, la costante c_α è determinata dalla relazione

$$c_\alpha \sim Z_\alpha \sqrt{\mathcal{I}(\theta_0)} \ .$$

Pertanto il test localmente ottimale cercato è basato sulle zone critiche

$$U(\underline{x},\theta_0) \geq Z_\alpha \sqrt{\mathcal{I}(\theta_0)} \ : \tag{1.13.9}$$

quando la (1.13.9) è verificata H_0 è rifiutata, contro l'alternativa $H_A(\theta_A > \theta_0)$; in caso contrario H_0 è accettata.

Osservazione 1.13.1: qualora si volesse calcolare la potenza del test (1.13.9), si può notare che per definizione di potenza e per definizione di $\Omega_{A,\alpha}$, si ha

$$\begin{aligned} \text{pow } (\theta_A;\alpha) &= P\{\underline{X} \in \Omega_{A,\alpha} | \theta = \theta_A\} = \\ &= P\{U(\underline{X};\theta_0) \geq Z_\alpha \sqrt{\mathcal{I}(\theta_0)} | \theta = \theta_A\} : \end{aligned} \tag{1.13.10}$$

d'altro canto, in prima approssimazione, posto $\delta\theta = \theta_A - \theta_0$,

$$\begin{aligned} U(\underline{X};\theta_0) &\cong U(\underline{X};\theta_A) - \delta\theta \ \partial_\theta U(\underline{X};\theta_A) \cong \\ &\cong U(\underline{X};\theta_A) + \delta\theta \mathcal{I}(\theta_A) \ . \end{aligned} \tag{1.13.11}$$

Ma quando $\theta = \theta_A$, $U(\underline{X};\theta_A)$ è asintoticamente normale e più precisamente

$$U(\underline{X};\theta_A) \sim \mathcal{N}[0, \mathcal{I}(\theta_A)] \ :$$

pertanto, con l'approssimazione (1.13.11), si può asserire che, quando vale H_A

$$U(\underline{X};\theta_0) \sim \mathcal{N}[\delta\theta \mathcal{I}(\theta_A), \mathcal{I}(\theta_A)] \ . \tag{1.13.12}$$

Ne deriva che

$$\text{pow } (\theta_A,\alpha) \sim P\left\{ Z \geq Z_\alpha \sqrt{\frac{\mathcal{I}(\theta_0)}{\mathcal{I}(\theta_A)}} - \delta\theta \sqrt{\mathcal{I}(\theta_A)} \right\} \ , \tag{1.13.13}$$

Z indicando al solito la normale standard.

Si può notare che tale funzione, benché approssimata, soddisfa correttamente le seguenti proprietà

- pow $(\theta_0, \alpha) = \alpha$
- pow $(\theta_A, \alpha) \to 1$ quando θ_A, e quindi $\delta\theta$, tende a $+\infty$
- pow $(\theta_A, \alpha) \to 1$ per $N \to +\infty$, perché $\mathcal{I}(\theta) = 0(N)$ e l'argomento di (1.13.13) quindi va come $-0(\sqrt{N})$.

1.14 Decisione tra alternative

Fino ad ora abbiamo considerato l'ipotesi alternativa H_A come una direzione lungo la quale ci si deve chiedere se ci si allontana da H_0: l'idea di massimizzare la potenza e la teoria che ne deriva traduce esattamente l'esigenza di cautelarsi col test contro un ben preciso tipo di allontanamento da H_0. Tuttavia il test col suo insieme di accettazione nella forma

$$\mathrm{lr}_{A,0}(\underline{x}) \leq c_\alpha \ ,$$

non esprime affatto una scelta tra H_0 e H_A, proprio perché la potenza, anche se massima, può essere bassa.

Infatti, si consideri il caso di una famiglia $\mathcal{N}[\mu, 1]$ con l'alternativa

$$\begin{array}{ll} H_0 & \mu_0 = 0 \\ H_A & \mu_A = 0,1 \ : \end{array}$$

fissato ad esempio $\alpha = 5\%$, per un campione di numerosità 16, si ha accettazione di H_0 nell'intervallo per la media campionaria.

$$m \leq (1/4) Z_{0,05} = 0,41 \ :$$

ma naturalmente se esce una media campionaria $m = 0,1$ questa non può essere presa come indicazione che H_0 è giusta ed H_A sbagliata; in

effetti possiamo dire solo che il valore empirico $m = 0,1$ non è così significativamente diverso da zero, in direzione di $\mu_A = 0,1$, da poter dire che H_0 va rigettata col livello di significatività del 5%.

Naturalmente questa situazione è resa evidente dal fatto che il test risulta assai poco potente, in quanto

$$\text{pow } = P\{m > 0,41 | \mu = 0,1\} = P\{Z > 1,24\} = 10,75\%$$

Vogliamo ora però risolvere un problema diverso, cioè quello di scegliere tra le due ipotesi H_0 ed H_A.

Possiamo fare ciò basandoci sempre sulla teoria della likelihood ratio, ovvero sulla variabile U: a tale scopo considereremo un test, basato su un valore di discriminazione D, per cui

$$\begin{array}{ll} \text{lr}_{A,0} < D & \text{si sceglie } H_0 \\ \text{lr}_{A,0} > D & \text{si sceglie } H_A \ . \end{array} \tag{1.14.1}$$

Naturalmente in questo approccio sarebbe necessario conoscere la distribuzione di $\text{lr}_{A,0}(\underline{X})$ sotto varie ipotesi. L'impresa è ovviamente ardua in termini generali, tuttavia se ci si accontenta di risultati asintoticamente esatti, per $N \to \infty$, si può notare che

$$\begin{aligned} \text{lr}_{A,0}(\underline{X}) &= \frac{L(\underline{X};\theta_A)}{L(\underline{X};\theta_0)} \cong 1 + \delta\theta \frac{\partial_\theta L(\underline{X};\theta_0)}{L(\underline{X};\theta_0)} = \\ &= 1 + \delta\theta U(\underline{X};\theta_0) \end{aligned} \tag{1.14.2}$$

e, come si è già visto nel paragrafo 1.13, si ha il risultato distribuzionale approssimato, quando $\theta = \theta_A$

$$U(\underline{X};\theta_0) \sim \mathcal{N}[\delta\theta\mathcal{I}(\theta_A), \mathcal{I}(\theta_A)] \ . \tag{1.14.3}$$

Fissiamo quali sono le variabili che entrano in gioco:

N = numerosità del campione con cui si esegue il test

D = valore discriminante tra le due ipotesi

θ_0 = valore di θ secondo l'ipotesi H_0

θ_A = valore di θ secondo l'ipotesi H_A

α (rischio di primo tipo) = $P\{$scegliere H_0 quando è vera $H_0\}$ = $= P\{U(\underline{X};\theta_0) \geq D|\theta = \theta_0\}$

β (rischio di secondo tipo) = $P\{$scegliere H_0 quando è vera $H_A\}$ = $P\{U(\underline{X};\theta_0) \leq D|\theta = \theta_A\}$.

In base alle definizioni e supponendo di poter usare (1.14.2) e (1.14.3), si vede che tra le varie grandezze devono sussistere due relazioni

$$\begin{array}{ll} D = Z_\alpha \sqrt{\mathcal{I}(\theta_0)} & (\text{dalla definizione di } \alpha) \\ D = \mathcal{I}(\theta_A)\delta\theta - Z_\beta \sqrt{\mathcal{I}(\theta_A)} & (\text{dalla definizione di } \beta) \end{array} \tag{1.14.4}$$

notiamo che N appare implicitamente, in quanto

$$\begin{aligned} \mathcal{I}(\theta) &= E\{-\partial_\theta U(\underline{X};\theta)\} = E\{-\partial_\theta^2 \log L(\underline{X};\theta)\} = \\ &= N\ E\{-\partial_\theta^2 \log f(x;\theta)\} \end{aligned} \tag{1.14.5}$$

e ricordiamo anche che

$$\delta\theta = \theta_A - \theta_0\ . \tag{1.14.6}$$

Mediante le (1.14.4) (1.14.4) due delle grandezze in gioco possono essere ricavate dalle altre; a seconda della scelta delle variabili si hanno problemi con vari significati statistici, illustrati nei seguenti esempi.

Esempio 1.14.1: dati $\theta_0, \theta_A, \alpha, N$ trovare D, β; è essenzialmente il problema che abbiamo trattato finora, in cui D è calcolato puramente in base a α e invece β, e quindi la potenza $(= 1 - \beta)$ è calcolata di conseguenza.

Esempio 1.14.2: dati θ_0, θ_A, N, D trovare α, β; si mette l'accento su una scelta a priori di D, ad esempio $D = (1/2)(\theta_0 + \theta_A)$ così che si hanno uguali rischi di I e II tipo; α e β vengono calcolati di conseguenza.

Esempio 1.14.3: noti $\theta_0, N, \alpha, \beta$ trovare D, θ_A; l'accento è sulla sensibilità del test, ovvero sulla possibilità di discriminare un'ipotesi alternativa θ_A quanto più possibile vicina a θ_0, fissati i rischi di I e II tipo.

Esempio 1.14.4: dati $\theta_0, \theta_A, \alpha, \beta$ trovare N, D; l'accento è sulla numerosità del campione necessaria a discriminare tra due ipotesi θ_0, θ_A fissate con rischi di I e II tipo dati.

1.15 Cenni ai metodi non parametrici per i test di ipotesi

Talvolta la mancanza completa di informazioni sulle distribuzioni da cui si estraggono i nostri campioni può spingere alla ricerca di formulazioni di test che prescindono completamente dalla distribuzione sottostante e perciò chiamati test non parametrici o distribution free.

Ci metteremo qui nell'ipotesi semplice di due campioni bernoulliani $x_1, x_2, \ldots, x_{n_1}, y_1, y_2, \ldots, y_{n_2}$ ed analizzeremo il problema di confrontare le medie di X ed Y, supponendo che

$$X_i \sim f_X(x_i - \theta) \quad ; \quad Y_i \sim f_X(y_i) \tag{1.15.1}$$

cioè che X e Y siano tratte indipendentemente da distribuzioni con la stessa forma, ma si sospetta che le X possano avere una media diversa dalle Y; si cercherà perciò di testare l'ipotesi semplice

$$H_0 \ : \ \theta = 0 \ . \tag{1.15.2}$$

L'idea di base con cui si opera è quella di unire $x_1 \ldots x_{n_1} y_1 \ldots y_{n_2}$ in un unico campione $\{v_i\}$ di numerosità $N = n_1 + n_2$

$$v_1 = x_1, \ v_2 = x_2, \ \ldots v_{n_1} = x_{n_1}, v_{n_1+1} = y_1 \ldots v_N = y_{n_2} \tag{1.15.3}$$

osservando che se H_0 è vera, questo è un campione bernoulliano tratto dalla distribuzione incognita $f_X(v)$; ma allora estraendo da $\{v_i, \ i = 1 \ldots N\}$, n_1 elementi a caso, (proprio come nell'estrazione da un'urna senza rimpiazzo) dovremmo ottenere un campione con caratteristiche

statistiche completamente analoghe al campione $\{x_i,\ i = 1 \dots n_1\}$ che ci apparirà così come una tra molte possibili scelte, tutte equiprobabili.

Così disponendo ad esempio di un algoritmo automatico di ricampionamento si possono estrarre molti sottocampioni $\{V_{\omega_i}; i = 1, \dots n_1\}$ a caso e studiare ad esempio la distribuzione della statistica

$$\overline{v}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} v_{\omega_i} \tag{1.15.4}$$

ricostruendone empiricamente la forma; così si potranno determinare le code $\overline{v}_1 \leq V_{\alpha/2}^{\inf}$, $\overline{v}_1 \geq V_{\alpha/2}^{\sup}$, che portano una probabilità totale α, e se risulta

$$V_{\alpha/2}^{\inf} \leq \overline{x} = \frac{1}{n_1} \sum_{i=1}^{n} x_i \leq V_{\alpha/2}^{\sup} \tag{1.15.5}$$

accetteremo H_0, in caso contrario lo rifiuteremo.

Osservazione 1.15.1: potrebbe sembrare che, volendo confrontare la media delle X con quella delle Y, sarebbe più sensato costruire una statistica del tipo

$$\overline{v}_1 - \overline{v}_2 = \frac{1}{n_1} \sum_{i=1}^{n-1} v_{\omega i} - \frac{1}{n_2} \sum_{j=1}^{n_2} v_{\eta j} \tag{1.15.6}$$

dove $\eta \equiv \{\eta_1 \dots \eta_{n2}\}$ è il vettore di indici che rimane da $\{1,\ 2, \dots N\}$ tolto $\{\omega_1 \dots \omega_{n1}\}$. In realtà, poiché i valori $\{v_i\ ;\ i = 1 \dots N\}$ sono fissati, la statistica (1.15.6) è direttamente funzione di $\overline{v}_1$; infatti, posto $\overline{v} = 1/N \sum_{i=1}^{N} v_i$, si ha

$$\overline{v}_1 - \overline{v}_2 = \overline{v}_1 - \frac{1}{n_2}(N\overline{v} - n_1\overline{v}_1)\ , \tag{1.15.7}$$

che è appunto funzione di $\overline{v}_1$, in quanto $\overline{v}$ è costante.

Ora notiamo che il procedimento delineato richiede un notevole lavoro di calcolo, che aumenta assai rapidamente al crescere di n_1 e n_2 ; è naturale

pertanto chiedersi se non sia possibile trovare una qualche distribuzione approssimata per la statistica $\overline{v}_1$. A questo proposito, anche data la forma di $\overline{v}_1$, ci si può aspettare che valga un'approssimazione normale; in effetti si può vedere che questa è sensata se n_1 e n_2 non sono troppo diversi tra loro, ovvero se $\frac{|n_1-n_2|}{N}$ è piccolo. In questa ipotesi il problema è essenzialmente quello di trovare la $E\{\overline{v}_1\}$ e la $\sigma^2(\overline{v}_1)$; osserviamo che nella (1.15.4) l'elemento stocastico è la scelta del vettore $\omega = \{\omega_1 \dots \omega_{n_1}\}$ che varia su tutte le $L = N(N-1)\dots(N-n_1+1)$ possibili scelte, considerate tutte equiprobabili. Dunque le estrazioni ω sono un insieme numerabile i cui elementi indicheremo con
$\omega^\ell = \{\omega_1^\ell, \dots \omega_{n_1}^\ell\}$, $\ell = 1, \dots L$. Ma allora

$$\begin{aligned} E\{\overline{v}_1\} &= \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_1}\sum_{i=1}^{n_1} v_{\omega_i^\ell} = \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{L}\sum_{\ell=1}^{L} v_{\omega_i^\ell} \ . \end{aligned} \tag{1.15.8}$$

Ora, tra tutte le estrazioni ω, vi sono ad esempio quelle per cui $\omega_i^\ell = 1$, quelle per cui $\omega_i^\ell = 2$ e così via. È chiaro per motivi di simmetria che queste sono tutte altrettanto numerose e che ad esempio le ω per cui $\omega_i^\ell = 1$ sono $(N-1)\cdot(N-2)\dots(N-n_1+1) = L/N$.

Quindi la (1.15.8) diventa

$$\begin{aligned} E\{\overline{v}_1\} &= \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{L}\sum_{k=1}^{N}\frac{L}{N}v_k = \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{1}{N}\sum_{k=1}^{N} v_k\right) = \overline{v} \ . \end{aligned} \tag{1.15.9}$$

In modo perfettamente analogo si può calcolare

$$\sigma^2(\overline{v}_1) = E\{(\overline{v}_1 - \overline{v})^2\} = \tag{1.15.10}$$

$$= \frac{n_2}{n_1}\frac{1}{(N-1)N}\sum_{k=1}^{N}(v_k-\overline{v})^2 =$$
$$= \frac{n_2}{n_1}\frac{1}{N-1}s^2 \ ,$$

essendo $s^2 = (1/N)\sum_{k=1}^{N}(v_k-\overline{v})^2$ la varianza campionaria completa.

Quindi ora un test può essere ricavato, al solito, con la relazione approssimata

$$\frac{\overline{v}_1-\overline{v}}{\sqrt{\frac{n_2}{n_1(N-1)}}S} \sim Z \ . \tag{1.15.11}$$

Osservazione 1.15.2: un'idea sostanzialmente equivalente, ma assai comoda nella costruzione del test, è di passare dal campione originario $\{v_i \ ; \ i=1,\ldots,N\}$ al cosiddetto campione dei ranghi $\{R_i \ ; \ i=1,\ldots N\}$; notiamo che il rango R_i è il numero d'ordine di v_i quando il campione venga riordinato in ordine crescente. Così ad esempio al campione $\{v_1 = 1, \ v_2 = -1, \ v_3 = 4, \ v_4 = 2\}$, che riordinato diventa $\{v_2, v_1, v_4, v_3\}$, corrisponde il vettore dei ranghi $\{R_1 = 2, \ R_2 = 1, \ R_3 = 4, \ R_4 = 3\}$.

Come si vede, il vettore dei ranghi contiene tutti gli interi da 1 a N. La comodità di usare $R = \{R_i \ ; \ i = 1, 2 \ldots N\}$ invece di $\{v_i \ ; \ i = 1 \ldots N\}$, sta nel fatto che per R sono fissati i valori

$$\overline{R} = \frac{1}{N}\sum_{i=1}^{N}R_i = \frac{N+1}{2} \quad , \quad S^2 = \frac{1}{N}\sum_{i=1}^{N}R_i^2 - \left(\frac{N+1}{2}\right)^2 = \frac{N^2-1}{12} \tag{1.15.12}$$

così che la (1.15.11) diventa direttamente

$$\frac{\overline{R}_1 - \frac{N+1}{2}}{\sqrt{\frac{n_2}{n_1}\frac{N+1}{12}}} \sim Z \ . \tag{1.15.13}$$

Esempio 1.15.1: siano dati i due campioni

$$x_1 = 1,98\ x_2 = 0,94\ x_3 = -0,51\ x_4 = 1,01\ x_5 = 2,22\ x_6 = 1,99$$
$$y_1 = -1,02\ y_2 = 1,62\ y_3 = 0,25\ y_4 = 0,15\ y_5 = -1,67$$

tratti rispettivamente da una $\mathcal{N}[0,8\ ;\ 1]$ e da una $\mathcal{N}[0\ ;\ 1]$. Ci si chiede se la media del campione X è significativamente diversa da quella del campione Y al livello di significatività $\alpha - 5\%$.

Per prima cosa uniamo i due campioni in uno solo, di 11 elementi, e costruiamo il corrispondente vettore dei ranghi

		1	2	3	4	5	6
v	=	1,98	0,94	-0,51	1,01	2,22	1,99
R	=	9	6	3	7	11	10
		7	8	9	10	11	
v	=	-1,02	1,62	0,25	0,15	-1,67	
R	=	2	8	5	4	1	

Ora usando la (1.15.13) si trova che il valore sperimentale della statistica è

$$\frac{7,\overline{6} - 6}{\sqrt{5/6}} \cong 1,826 = Z_{sp}\ .$$

Notiamo che se $H_0 : \mu_X = \mu_Y$, senza alternative preferenziali risulta

$$|Z_{sp}| < Z_{0,025} = 1,96$$

e quindi H_0 va accettata (non può essere rifiutata), mentre se avessimo posto la questione se $\mu_X > \mu_Y$, si sarebbe avuto

$$Z_{sp} > Z_{0,025} = 1,645$$

ed H_0 sarebbe stata rifiutata.

Si noti che se invece della (1.15.13) si fosse usata la (1.15.11) si sarebbe trovato

$$Z_{sp} = 2,145$$

che è diverso dal precedente, ma non di tanto.

Osservazione 1.15.3: il metodo usato per confrontare μ_X e μ_Y tramite la (1.15.13) può essere facilmente generalizzato al confronto fra le dispersioni dei due campioni; basterà infatti in questo caso sostituire nei campioni $\{x_i,\ i = 1 \dots n_1\}$, $\{y_i\ ,\ i = 1 \dots n_2\}$ i campioni $\{|x_i - m_x|,\ i = 1 \dots n_1\}$, $\{|y_i - m_Y|,\ i = 1 \dots n_2\}$ che poi potranno essere uniti e trasformati in un vettore di ranghi R. Applicando a questo vettore la (1.15.13) si potrà poi sottoporre a test l'ipotesi che i ranghi dei due campioni siano uguali, contro quella che siano diversi, ad esempio $\overline{R}_1 > \overline{R}_2$ che indicherebbe una varianza maggiore di X rispetto ad Y.

2 L'inferenza per le stime della teoria dei minimi quadrati

2.1 Risultati distribuzionali per campioni normali

In questo paragrafo ci prepariamo i risultati di base che ci serviranno nei prossimi paragrafi per svariate applicazioni.

Fino ad ora la teoria dei minimi quadrati è stata svolta indipendentemente dalla distribuzione della variabile campionaria Y (vettore delle osservabili). Naturalmente, se vogliamo dedurre risultati distribuzionali per le stime ottenute con la teoria dei minimi quadrati, occorre fare un'ipotesi di partenza sulla distribuzione di Y: noi supporremo, d'ora in poi e per tutto questo capitolo, che

$$Y = \mathcal{N}[y, \sigma_0^2 Q] \ . \tag{2.1.1}$$

Inoltre rimarremo sempre nell'ambito dei modelli lineari, poiché la teoria generale per i modelli non lineari è troppo complessa ed incompleta: naturalmente le nostre conclusioni potranno valere, sia pure in forma approssimata, per quei modelli non lineari che però nel dominio di maggior densità di Y sono ben approssimati mediante equazioni linearizzate. Inoltre, per semplicità e data l'importanza del caso, ci riferiremo sempre al modello parametrico

$$y = Ax + a \ . \tag{2.1.2}$$

Premettiamo un lemma sulle variabili normali in R^n.

Lemma 2.1.1. sia, $(\underline{V} \in R^n)$, $\underline{V} = \mathcal{N}[0, C]$ e sia

$$\mathcal{R}^n = \mathcal{V}^1 \oplus \mathcal{V}^2 \qquad (\dim \mathcal{V}^i = n_i, \ n_1 + n_2 = n)$$

una decomposizione di $\mathcal{R}^n$ in due varietà complementari C^{-1}-ortogonali, ovvero sia

$$\begin{cases} \forall \underline{v} \in \mathcal{R}^n & \underline{v} = \underline{v}_1 + \underline{v}_2 \\ \underline{v}_1 \in \mathcal{V}^1 \quad , \quad \underline{v}_2 \in \mathcal{V}_2 \ , & \\ \underline{z}_1^+ C^{-1} \underline{v}_2 = 0 \ , & \end{cases} \tag{2.1.3}$$

allora, data la decomposizione della v.c. $\underline{V}$

$$\underline{V} = \underline{V}_1 + \underline{V}_2 \qquad (V_1 \in \mathcal{V}^1, \ V_2 \in \mathcal{V}^2) \tag{2.1.4}$$

si ha che $\underline{V}_1$ e $\underline{V}_2$ sono variabili normali stocasticamente indipendenti e per di più

$$\begin{cases} \underline{V}_1^+ C^{-1} \underline{V}_1 = \chi^2_{n_1} \\ \underline{V}_2^+ C^{-1} \underline{V}_2 = \chi^2_{n_2} \ , \end{cases} \tag{2.1.5}$$

così che dalla decomposizione (pitagorica)

$$\underline{V}_1^+ C^{-1} \underline{V}_1 + \underline{V}_2^+ C^{-1} \underline{V}_2 = \underline{V}^+ C^{-1} \underline{V}$$

si ricava anche la nota legge di decomposizione delle χ^2 indipendenti

$$\chi^2_{n_1} + \chi^2_{n_2} = \chi^2_n \ .$$

Per comprendere il lemma basta pensare che V può essere considerata come generata da una $\underline{Z} \in \mathcal{R}^n$ normale standardizzata

$$\underline{V} = C^{1/2} \underline{Z} \quad , \quad \underline{Z} = \mathcal{N}[0, I] \ . \tag{2.1.6}$$

Inoltre ci saranno due varietà $\mathcal{U}^1, \mathcal{U}^2$ nello spazio delle $\underline{Z}$, tali che

$$\begin{array}{l} \underline{z} = \underline{z}_1 + \underline{z}_2 \quad , \quad \underline{z}_1 \in \mathcal{U}^1 \quad , \quad \underline{z}_2 \in \mathcal{U}^2 \\ (\mathcal{V}^1 = C^{1/2} \mathcal{U}^1 \quad , \quad \mathcal{V}^2 = C^{1/2} \mathcal{U}^2) \ . \end{array} \tag{2.1.7}$$

Notiamo che in conseguenza della (2.1.3) si ha

$$\underline{z}_1^+ \underline{z}_2 = \underline{v}_1^+ C^{-1/2} C^{-1/2} \underline{v}_2 = \underline{v}_1^+ C^{-1} \underline{v}_2 = 0 \ ,$$

cioè $\mathcal{U}^1$ e $\mathcal{U}^2$ sono varietà ortogonali complementari dello spazio $\mathcal{R}^n$ in cui Z è normale standardizzata. Ma allora esisterà una scelta di assi per cui

$$\underline{z} = \begin{vmatrix} z_1 \\ \vdots \\ z_{n_1} \\ z_{(n_1+1)} \\ \vdots \\ z_n \end{vmatrix} = \begin{vmatrix} z_1 \\ \vdots \\ z_{n_1} \\ 0 \\ \vdots \\ 0 \end{vmatrix} + \begin{vmatrix} 0 \\ \vdots \\ 0 \\ z_{(n_1+1)} \\ \vdots \\ z_n \end{vmatrix} = \underline{z}_1 + \underline{z}_2 \ ,$$

con $\underline{z}_1 \in \mathcal{U}^1$, $\underline{z}_2 \in \mathcal{U}^2$. Da qui appare ovvio che $\underline{z}_1$ è stocasticamente indipendente da $\underline{z}_2$ perché

$$\frac{1}{\sqrt{2\pi}^n} e^{-\underline{z}^+ \underline{z}} = \frac{1}{\sqrt{2\pi}^{n_1}} e^{-\underline{z}_1^+ \underline{z}_1} \cdot \frac{1}{\sqrt{2\pi}^{n_2}} e^{-\underline{z}_2^+ z_2} \ ,$$

ed anche che

$$\begin{aligned} \underline{z}_1^+ \underline{z}_1 &= \underline{v}_1^+ C^{-1} \underline{v}_1 = z_1^2 + \ldots + z_{n_1}^2 = \chi_{n_1}^2 \\ \underline{z}_2^+ \underline{z}_2 &= \underline{v}_2^+ C^{-1} \underline{v}_2 = z_{n_1+1}^2 + \ldots + z_n^2 = \chi_{n_2}^2 \ , \end{aligned}$$

c.v.d.

Osservazione 2.1.1: applicando ripetutamente il Lemma 2.1.1 si ha che più in generale, alla decomposizione

$$\begin{cases} \mathcal{R}^n = \mathcal{V}^1 \oplus \mathcal{V}^2 \oplus \ldots \oplus \mathcal{V}^p \quad (\dim \mathcal{V}^i = n_i \quad , \quad n_1 + \ldots + n_p = n) \\ \underline{v} = \underline{v}_1 + \underline{v}_2 + \ldots + \underline{v}_p \\ \underline{v}_i \in \mathcal{V}^i \quad , \quad \underline{v}_i^+ C^{-1} \underline{v}_j = 0 \end{cases} \tag{2.1.8}$$

ed alla relativa decomposizione pitagorica

$$\underline{v}^+ C^{-1} \underline{v} = \underline{v}_1^+ C^{-1} \underline{v}_1 + \ldots + \underline{v}_p^+ C^{-1} \underline{v}_p \tag{2.1.9}$$

corrisponde una decomposizione della v.c. $\mathcal{V} = \mathcal{N}[0, C]$ in p componenti $\underline{V}_i (i = 1, \ldots, p)$ stocasticamente indipendenti, tali che

$$\underline{V}_i^+ C^{-1} \underline{V}_i = \chi^2_{n_i} \ , \tag{2.1.10}$$

così che la (2.1.9) rappresenta la regola di composizione di χ^2 indipendenti.

Osservazione 2.1.2: se anziché usare la matrice di covarianza C si usa una matrice proporzionale Q nel definire la Q^{-1}-ortogonalità, l'indipendenza continua a sussistere ed il risultato (2.1.10) diventa

$$\left\{ \begin{array}{lcl} \underline{V}_i^+ Q^{-1} \underline{V}_i & = & \sigma_0^2 \chi^2_{n_i} \\ \sum \underline{V}_i^+ Q^{-1} \underline{V}_i & = & \sigma_0^2 \chi^2_n \ . \end{array} \right. \tag{2.1.11}$$

Sulla scorta del Lemma 2.1.1 è assai semplice provare il basilare teorema che segue.

Teorema 2.1.1: sia dato un problema di m.q.

$$\left\{ \begin{array}{lcll} y = Ax + a & , & y = E\{Y\} & (\dim y = n, \ \dim x = m) \\ Y_0 (\text{osservazioni}) & , & C_{YY} = \sigma_0^2 Q & \end{array} \right.$$

e siano

$$\begin{array}{rcll} \hat{x} & = & N^{-1} A^+ Q^{-1} (Y_0 - a) & (N = A^+ Q^{-1} A) \\ \hat{y} & = & A\hat{x} + a & \\ U & = & Y_0 - A\hat{x} - a = Y_0 - \hat{y} & \end{array} \tag{2.1.12}$$

rispettivamente le stime (corrette) di (x, y) ed il vettore degli scarti delle equazioni: allora

$$\left\{ \begin{array}{l} \hat{x} = \mathcal{N}[x, \sigma_0^2 N^{-1}] \\ \hat{y} = \mathcal{N}[y, \sigma_0^2 A N^{-1} A^+] \\ U = \mathcal{N}[0, \sigma_0^2 (Q - A N^{-1} A^+)] \end{array} \right. \tag{2.1.13}$$

inoltre $(\hat{x}, \hat{y})$ ed U sono stocasticamente indipendenti e risulta

$$\begin{aligned} & (\hat{y} - y)^+ Q^{-1} (\hat{y} - y) = \\ = \; & (\hat{x} - x)^+ N (\hat{x} - x) = \sigma_0^2 \chi^2_m \qquad (2.1.14) \\ & U^+ Q^{-1} U = \sigma_0^2 \chi^2_{n-m} \; , \qquad (2.1.15) \end{aligned}$$

le due forme quadratiche essendo tra loro indipendenti, così che è anche

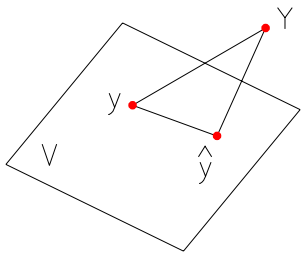
$$\hat{\sigma}_0^2 = \frac{U^+ Q^{-1} U}{n - m} = \frac{\sigma_0^2}{n - m} \chi^2_{n-m} \qquad (2.1.16)$$

con σ_0^2 in particolare indipendente da $\hat{x}$.

In effetti è sufficiente osservare che le (2.1.12) definiscono trasformazioni lineari del vettore Y per poter asserire che $\hat{x}, \hat{y}, U$ devono essere a loro volta normali: le (2.1.13) poi discendono dal paragrafo 2.4 del Quaderno n. 2 e in particolare dal fatto che $\hat{x}, \hat{y}$ sono stimatori corretti e dalla conoscenza delle loro matrici di covarianza.

Inoltre, proprio in base al principio dei m.q., si può scrivere la decomposizione

$$Y - y = Y - \hat{y} + \hat{y} - y = U + (\hat{y} - y) = U + A(\hat{x} - x)$$



notando che essa corrisponde alla decomposizione

$$\mathcal{R}^n = \mathcal{V}^c + \mathcal{V} \qquad (\dim \mathcal{V}^c = n - m,\ \dim \mathcal{V} = m)$$

dove $\mathcal{V}^c$ è Q^{-1}-ortogonale a $\mathcal{V}$ (infatti
$A^+Q^{-1}U = A^+Q^{-1}(Y_0 - a - A\hat{x}) = A^+Q^{-1}(Y_0 - a) - N\hat{x} = 0$).

Per il Lemma 2.1.1 si ha allora che U e $\hat{y} - y$ sono stocasticamente indipendenti e altrettanto deve essere per U e $\hat{x}$, essendo quest'ultimo fissato in modo univoco da $\hat{y}$: sempre per il Lemma 2.1.1 valgono le (2.1.14) e (2.1.15). Si può osservare che l'indipendenza stocastica di U e $\hat{x}$ è coerente col fatto (cfr. paragrafo 2.4 del Quaderno n. 2) che $C_{U\hat{x}} = 0$.

2.2 Verifica della correttezza del modello deterministico

Ci proponiamo di sottoporre a ipotesi la correttezza del modello deterministico

$$y = Ax + a\ , \tag{2.2.1}$$

in particolare contro l'ipotesi che anziché conoscere il vettore a esatto si conosca erroneamente $a + \delta a$, dove δa rappresenta un vettore costante di errori sistematici o bias.

Osservazione 2.2.1: questo tipo di test è importante quando si sospetti che tra i valori osservati Y_{0i}, qualcuno non segua la legge prevista dal modello (2.2.1) perché nel processo di osservazione si è verificato un evento imprevisto: si ha così per Y_{0i} un valore che sta al di fuori della popolazione prevista (in inglese "outlier").

Osservazione 2.2.2: nel caso che vi sia il sospetto di un errore sistematico nella matrice disegno A, ci si può ancora ridurre al test sul termine noto a, qualora sia nota una stima approssimata attendibile $\tilde{x}$. Infatti in tal caso si può porre

$$x \ = \ \tilde{x} + \xi$$

$$y = (A+\delta A)x + a = A\xi + a + A\tilde{x} + \delta A\tilde{x} + \delta A\xi \qquad (2.2.2)$$

così che, considerando $\delta A\xi$ come infinitesimo del 2° ordine che può essere trascurato, si può interpretare la (2.2.2) come

$$y = A\xi + a' + \delta a \ ,$$

con

$$\begin{aligned} a' &= a + A\tilde{x} \\ \delta a &= \delta A\tilde{x} \ (\underline{\text{costante}}). \end{aligned}$$

Per prima cosa vogliamo vedere quale sia l'effetto di sostituire a con $a+\delta a$ sulla stima di $\hat{\sigma}_0^2$.

In effetti, poiché

$$U = [I - AN^{-1}A^+Q^{-1}](Y_0 - a) = L(Y_0 - a) \ ,$$

sbagliando a di δa si sbaglierà la stima di U di una quantità

$$\delta U = -L\delta a \ . \qquad (2.2.3)$$

Corrispondentemente la stima di $\hat{\sigma}_0^2$ risulterà

$$\begin{aligned} \hat{\sigma}_0^2 &= \frac{1}{n-m}(U+\delta U)^+Q^{-1}(U+\delta U) = \\ &= \frac{1}{n-m}\{U^+Q^{-1}U + 2U^+Q^{-1}\delta U + \delta U^+Q^{-1}\delta U\} \qquad (2.2.4) \end{aligned}$$

anziché avere il valore corretto

$$\hat{\sigma}_0^2 = \frac{1}{n-m} U^+ Q^{-1} U \ . \tag{2.2.5}$$

Poiché la (2.2.5) dà uno stimatore corretto di σ_0^2, si ha dalla (2.2.4)

$$E\{\hat{\sigma}_0^2\} = \sigma_0^2 + \frac{2}{n-m} E\{U^+ Q^{-1} \delta U\} + \frac{1}{n-m} E\{\delta U^+ Q^{-1} \delta U\} \ , \tag{2.2.6}$$

ma dalla (2.2.3) si vede che δU è un vettore costante, così che

$$E\{U^+ Q^{-1} \delta U\} = E\{U^+\} Q^{-1} (-L\delta a) = 0$$

e poiché, essendo Q definita positiva,

$$E\{\delta U^+ Q^{-1} \delta U\} = \delta a^+ L^+ Q^{-1} L \delta a > 0$$

dalla (2.2.6) si deduce che in media

$$E\{\hat{\sigma}_0^2\} > \sigma_0^2 \ . \tag{2.2.7}$$

Si potrebbe anzi dire più precisamente che, quanto più la componente Q^{-1}-ortogonale a V di δa è grande, tanto più si gonfierà la stima di σ_0^2 rispetto al valore corretto.

Naturalmente la componente di δa parallela a V, esprimibile nella forma

$$\delta a_{||} = Ac \ ,$$

avrà come unico effetto di produrre un bias in $\hat{x}$, senza variare il vettore degli scarti U.

Sulla base del risultato (2.1.16) e vista la (2.2.7), si può istituire il seguente test:
fissata l'ipotesi $H_0 : \sigma_0^2 = \overline{\sigma}_0^2$, si calcola la stima $\hat{\sigma}_0^2$ e se H_0 è vera

$$(n-m)\frac{\hat{\sigma}_0^2}{\overline{\sigma}_0^2} = \chi^2_{0,n-m}$$

è un'estrazione da una v.c. χ^2 a $n-m$ gradi di libertà; fissato perciò un livello di significatività α del test ed il relativo valore critico $\overline{\chi}^2_{\alpha,n-m}$ (cfr. fig. 2.2.1) si verifica se

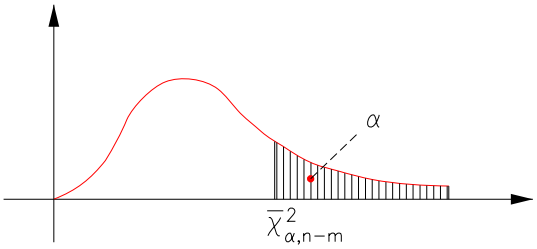


Figura 2.2.1:

$$\begin{cases} \chi^2_{0,n-m} & < \overline{\chi}^2_{\alpha,n-m} & \rightarrow & \text{si accetta} \;\; H_0 \\ \chi^2_{0,n-m} & > \overline{\chi}^2_{\alpha,n-m} & \rightarrow & \text{si rifiuta} \;\; H_0 \; . \end{cases} \tag{2.2.8}$$

Osservazione 2.2.3: come si è già notato il test è efficace contro la presenza in a di un bias δa, Q^{-1}-ortogonale a V, perché la componente parallela a V non modifica U e quindi neanche la stima $\hat{\sigma}^2_0$.

Osservazione 2.2.4: il test sul $\hat{\sigma}^2_0$, qualora H_0 vada rifiutata, non ci dice ancora quali componenti di a, ovvero di Y_0, siano affetti da errori sistematici. Questo è particolarmente importante perché, quando si sospetti che solo in poche equazioni, più spesso in una sola, siano presenti outliers, se queste possono essere identificate è possibile scartarle e ottenere così una stima più attendibile del vettore dei parametri x.

Indichiamo in breve il procedimento adottato nel caso più comune, nel quale si sospetti la presenza di un solo outlier.

In primo luogo si costruisce il vettore degli scarti normalizzati

$$v_i = \frac{u_i}{\sqrt{Q_{ii} - (AN^{-1}A^+)_{ii}}} \;\; : \tag{2.2.9}$$

in base alla (2.1.13), se non ci fossero outliers dovrebbe essere

$$v_i = \mathcal{N}[0, \sigma^2_0] \; , \tag{2.2.10}$$

così che se invece si sospetta che un outlier sia presente, è ragionevole aspettarsi che questo corrisponda al massimo dei v_i.

Per verificare tale ipotesi si riduce il sistema di equazioni d'osservazione eliminando l'equazione sospetta: supponiamo per semplicità che essa sia l'ultima e riscriviamo le (2.2.1) in forma partizionata

$$\begin{cases} y_{(n-1)} = A_{(n-1)}x + a_{(n-1)} \\ y_n = Rx + a_n \end{cases} \tag{2.2.11}$$

con R corrispondente all'ultima riga di A ed a_n all'ultima componente del termine noto in cui si sospetta la presenza dell'outlier.

Basandoci solo sulle prime $n-1$ equazioni (presumibilmente corrette) si stima x ottenendo il vettore

$$\hat{x}_{(n-1)} = (A^{+}_{(n-1)} Q^{-1}_{(n-1)} A^{-1}_{(n-1)})^{-1} A^{+}_{(n-1)} Q^{-1}_{(n-1)} (Y_{0,(n-1)} - a_{(n-1)}) \ , \quad (2.2.12)$$

che non è influenzato dall'eventuale outlier presente in a_n. Se la (2.2.12) è corretta, vale

$$C_{\hat{x}_{(n-1)}\hat{x}_{(n-1)}} = \sigma_0^2 N^{-1}_{(n-1)} \ .$$

Ora conviene porre un'ipotesi semplificativa, supponendo che le componenti di Y_0 siano tra loro incorrelate, così che in particolare Y_{0n} è indipendente dalle altre.

Ciò detto formiamo il vettore

$$Y_{0n} - (R\hat{x}_{(n-1)} + a_n) = W_n \ , \quad (2.2.13)$$

ed applichiamo la legge di propagazione degli errori supponendo che valga l'ipotesi $H_0(a_n$ è giusta$)$. Si ha allora

$$\begin{aligned} E\{W_n\} &= E\{Y_{0n} - (R\hat{x}_{(n-1)} + a_n)\} = 0 \\ \sigma^2(W_n) &= \sigma_0^2 Q_{nn} + R C_{\hat{x}_{(n-1)}\hat{x}_{(n-1)}} R^{+} = \sigma_0^2 (Q_{nn} + R N^{-1}_{(n-1)} R^{+}) \ . \end{aligned}$$

Poiché W_n è normale, essendo funzione lineare di variabili normali, quando vale H_0 si deve avere

$$\frac{W_n}{\sigma_0 \sqrt{Q_{nn} + R N^{-1}_{(n-1)} R^{+}}} = Z \ , \quad (2.2.14)$$

mentre se a_n, ovvero Y_{0n}, contiene un errore ci aspettiamo un valore significativamente diverso da zero di W_n.

Pertanto la (2.2.14) è adatta a verificare H_0, almeno se si suppone di conoscere σ_0 : fissato un livello di significatività α, se risulta

$$\frac{|W_n|}{\sigma_0\sqrt{Q_{nn}+RN_{(n-1)}^{-1}R^+}} \leq Z_{\alpha/2} \tag{2.2.15}$$

H_0 è accettata, in caso contrario si ritiene W_n contenga un outlier e la n-esima equazione viene scartata.

Si osservi che la (2.2.15) corrisponde ad un disegno del test su due code. Infine se invece di un valore noto a priori σ_0 dobbiamo utilizzare il valore stimato $\hat{\sigma}_0$, usando il Teorema 2.1.1. vediamo che

$$(n-1-m)\hat{\sigma}_0^2 = \sigma_0^2\chi_{n-1-m}^2$$

ed inoltre sappiamo che $\hat{\sigma}_0^2$ è indipendente da $\hat{x}_{n-1}$; se poi teniamo conto che $\hat{\sigma}_0^2$ dipende da $Y_{0,n-1}$ che è indipendente da Y_{0n}, dalla definizione (2.2.13) comprendiamo che $\hat{\sigma}_0^2$ è indipendente da W_n. Ma allora si potrà scrivere

$$\frac{W_n}{\hat{\sigma}_0\sqrt{Q_{nn}+RN_{(n-1)}^{-1}R^+}} = t_{n-1-m} \ , \tag{2.2.16}$$

che, sostituendo la (2.2.14), permette di eseguire il test per H_0.

Esempio 2.2.1: sia dato il modello lineare

$$y = Ax + a \qquad y = \begin{vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{vmatrix} \ , \ A = \begin{vmatrix} 1 \\ 2 \\ -1 \\ -2 \end{vmatrix} \ , \ x = [x] \ , \ a = \begin{vmatrix} 2 \\ -1 \\ 3 \\ -1 \end{vmatrix}$$

e si supponga di aver eseguito la osservazione di y,

$$Y_0{}^+ = [3,1 \qquad -2,2 \qquad 1,7 \qquad -2,8] \quad :$$

valga inoltre il modello stocastico $C_{YY} = \sigma_0^2 I$.

Sapendo, sulla base della conoscenza del processo di osservazione, che $\sigma_0 \sim 0,2$, si vuole verificare la correttezza del modello (2.2.16).

La stima di $\hat{x}$ è data da

$$\hat{x} = (10)^{-1}[1 \qquad 2 \qquad -1 \qquad -2] \begin{vmatrix} 1,1 \\ -1,2 \\ -1,3 \\ -1,8 \end{vmatrix} = 0,36$$

ed il corrispondente vettore degli scarti è

$$U^+ = [0,74 \qquad -1,92 \qquad -0,94 \qquad -1,08] \ .$$

La stima di $\hat{\sigma}_0^2$ è

$$\hat{\sigma}_0^2 = \frac{U^+ U}{4-1} = 2,09 \ ,$$

così che se $\sigma_0^2 \sim 0,04$ e se il modello è corretto, si dovrebbe avere

$$3\frac{\hat{\sigma}_0^2}{\sigma_0^2} = 156,75 \qquad \text{estrazione da } \chi^2(\nu = 3 \text{ gradi di libertà}).$$

D'altro canto al livello $\alpha = 1\%$ ed a 3 gradi di libertà

$$\chi^2_{(0,99)} = 11,3$$

perciò l'ipotesi di base va rifiutata.

Per decidere in quale equazione può essere presente un outlier calcoliamo

$$\sigma^2_{v_1} = \sigma^2_0 0,9 \quad ; \quad \sigma^2_{v_2} = \sigma^2_0 0,6 \quad ; \quad \sigma^2_{v_3} = \sigma^2_0 0,9 \quad ; \quad \sigma^2_{v_4} = \sigma^2_0 0,6$$

e quindi gli scarti normalizzati

$$v = \left| \begin{array}{c} v_1/\sqrt{0,9} \\ v_2/\sqrt{0,6} \\ v_3/\sqrt{0,9} \\ v_4/\sqrt{0,6} \end{array} \right| = \left| \begin{array}{r} 0,78 \\ -2,48 \\ -0,99 \\ -1,39 \end{array} \right| \leftarrow \text{ sospetto outlier.}$$

Eliminiamo la II equazione e ripetiamo la compensazione: risulta

$$\hat{x} = 1,00 \qquad U = \left| \begin{array}{r} 0,1 \\ -0,3 \\ 0,2 \end{array} \right| \qquad \hat{\sigma}^2_0 = \frac{U^+ U}{2} = 0,07 \ .$$

Il test su $H_0 : \sigma^2_0 = 0,04$ dà il risultato

$$2\frac{\hat{\sigma}^2_0}{\sigma^2_0} = 3,5$$

contro un valore critico di χ^2 per $\alpha = 1\%$, oppure $\alpha = 5\%$, $\nu = 2$ gradi di libertà

$$\chi^2_{(0,99)} = 9,21 \quad , \quad \chi^2_{(0,95)} = 5,99$$

H_0 va perciò accettata, e pare che l'outlier sia stato eliminato.

Per avere una conferma calcoliamo

$$W = Y_{02} - (2\hat{x} - 1) = -2,2 - (2 \cdot 1 - 1) = -3,2 \ :$$

per la legge di propagazione degli errori

$$\sigma^2(W) = \sigma_0^2 + 4\sigma_{\hat{x}}^2 = \sigma_0^2 + 4\frac{\sigma_0^2}{6} \cong 0,117 \ .$$

Se W fosse a media nulla (H_0)

$$\frac{W}{\sigma(W)} = -9,368$$

sarebbe un'estrazione da una t di Student a 2 gradi di libertà; con $\alpha = 5\%,\ t_{\alpha/2} = 4,30$ e l'ipotesi H_0 va rifiutata a quel livello di significatività. Si noti anche che per $\alpha = 1\%,\ \nu = 2$ si avrebbe $t_{\alpha/2} = 9,92$ per il quale il valore diventa appena accettabile.

2.3 Test sui parametri

Vogliamo risolvere il seguente problema:

> sia $\hat{x}$ il vettore dei parametri stimati in un problema di m.q. e sia $C_{\hat{x}\hat{x}} = \sigma_0^2 N^{-1}$ la corrispondente matrice di covarianza; fatta un'ipotesi $H_0(x = \overline{x})$ sui valori delle componenti di x, vogliamo decidere ad un livello di significatività α assegnato se H_0 è plausibile oppure se il vettore stimato $\hat{x}$ sia significativamente diverso da $\hat{x}$.

Il problema è assai semplice se si suppone di conoscere σ_0^2. Infatti, usando la (2.1.14), se $x = \overline{x}(\dim x = m)$ si ha che (cfr. Osservazione 17.6 nel Quaderno n. 1)

$$(\hat{x} - \overline{x})^+ C_{\hat{x}\hat{x}}^{-1}(\hat{x} - \overline{x} = \sigma_0^{-2}(\hat{x} - \overline{x})^+ N(\hat{x} - \overline{x}) = \chi_m^2 \ : \qquad (2.3.1)$$

la (2.3.1) è perfettamente adatta a valutare l'ipotesi H_0, in quanto se H_0 è giusta, il valore empirico

$$\sigma_0^{-2}(\hat{x}-\overline{x})^+N(\hat{x}-\overline{x})=\chi_0^2$$

è un'estrazione da una χ^2 a m gradi di libertà e quindi può essere confrontata con il valore critico χ^2_α, decidendo che

$$\begin{array}{rcl}\chi_0^2\le\chi_\alpha^2 & \rightarrow & H_0 \text{ è accettata}\\ \chi_0^2>\chi_\alpha^2 & \rightarrow & H_0 \text{ è rifiutata.}\end{array} \tag{2.3.2}$$

In effetti è facile vedere che, se $\overline{x}$ è diverso da x, vero valore medio di $\hat{x}$, la forma quadratica (2.3.1) tende a gonfiarsi in media

$$\begin{array}{rcl}E\{(\hat{x}-\overline{x})^+N(\hat{x}-\overline{x})\} & = & E\{(\hat{x}-x)^+N(\hat{x}-x)\}+\\ & + & 2E\{(\hat{x}-x)^+N(x-\overline{x})\}+\\ & + & E\{(x-\overline{x})^+N(x-\overline{x})\}=\\ & = & E\{(\hat{x}-x)^+N(\hat{x}-x)\}+(\overline{x}-x)^+N(\overline{x}-x)\ge\\ & \ge & E\{(\hat{x}-x)^+N(\hat{x}-x)\}=m\sigma_0^2\ .\end{array}$$

Se invece σ_0^2 è incognito, si potrà sempre usare la (2.1.14) insieme alla (2.1.15) riscritta nella forma

$$\begin{cases}\frac{1}{m}(\hat{x}-x)^+N(\hat{x}-x)=\sigma_0^2\dfrac{\chi_m^2}{m}\\ \frac{1}{n-m}U^+Q^{-1}U=\hat{\sigma}_0^2=\sigma_0^2\dfrac{\chi^2_{n-m}}{n-m}\ :\end{cases} \tag{2.3.3}$$

ricordando che le due forme quadratiche (2.3.3) sono stocasticamente indipendenti, dividendo membro a membro si trova:

$$\frac{(1/m)(\hat{x}-x)^+N(\hat{x}-x)}{\hat{\sigma}_0^2}=F_{m,n-m}\ . \tag{2.3.4}$$

Dunque il valore campionario (empirico) della funzione (2.3.4) può essere confrontato con il valore critico F_α di una F di Fisher a $(m,n-m)$ gradi di libertà: se vale $H_0(x=\overline{x})$

$$F_0 = \frac{1/m(\hat{x} - \overline{x})^+ N(\hat{x} - \overline{x})}{\hat{\sigma}_0^2}$$

deve essere minore di F_α con probabilità $(1 - \alpha)$, mentre se H_0 è falsa, F_0 tende ad aumentare, perciò

$$\begin{cases} F_0 \leq F_\alpha & \rightarrow \text{ accetto } H_0 \\ F_0 > F_\alpha & \rightarrow \text{ rifiuto } H_0 \ . \end{cases} \tag{2.3.5}$$

Osservazione 2.3.1: spesso si richiede di sottoporre a test una parte soltanto del vettore x, cioè una o più componenti.

Sia P la matrice che estrae dal vettore x, le r componenti volute, formando un nuovo vettore ridotto ξ. Ad esempio, se dim $x = 5$ e si vogliono le componenti 2 e 5 solo (dim $\xi = r = 2$), si ha

$$\xi = \begin{vmatrix} x_2 \\ x_5 \end{vmatrix} = \begin{vmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{vmatrix} = Px \ . \tag{2.3.6}$$

Si noti che se $C_{\hat{x}\hat{x}}$ è la matrice di covarianza di $\hat{x}$, $PC_{\hat{x}\hat{x}}P^+$ è la matrice di covarianza di $\hat{\xi}$: nel caso illustrato in (2.3.6) si ha

$$C_{\hat{\xi}\hat{\xi}} = PC_{\hat{x}\hat{x}}P^+ = \begin{vmatrix} C_{22} & C_{25} \\ C_{52} & C_{55} \end{vmatrix} \qquad (C_{ik} = \text{cov}(\hat{x}_i \hat{x}_k)) \ . \tag{2.3.7}$$

Poiché $C_{\hat{x}\hat{x}} = \sigma_0^2 N^{-1}$, per $\hat{\xi} = P\hat{x}$ varrà allora la relazione

$$(1/\sigma_0)(\hat{\xi} - \xi)^+ (PN^{-1}P^+)^{-1}(\hat{\xi} - \xi) = \chi_r^2 \tag{2.3.8}$$

È bene notare che $(PN^{-1}P^+)^{-1}$ è una nuova matrice r-dimensionale, che va appositamente calcolata, e non coincide affatto con PNP^+.

Usando la (2.3.8) e la (2.1.15), posta l'ipotesi $H_0(\xi = \overline{\xi})$, si può calcolare il valore campionario

$$\frac{(1/r)(\hat{\xi}-\overline{\xi})^{+}(PN^{-1}P^{+})^{-1}(\hat{\xi}-\overline{\xi})}{\hat{\sigma}_0^2} = F_0 \ , \tag{2.3.9}$$

che quando H_0 è vera è un'estrazione da una F di Fisher a $(r, n-m)$ gradi di libertà

$$F_0 \sim F_{r,n-m} \ . \tag{2.3.10}$$

Si avrà allora, usando il valore critico F_α, a $(r, n-m)$ gradi di libertà

$$\begin{cases} F_0 \leq F_\alpha & \rightarrow \text{ accetto } H_0 \\ F_0 > F_\alpha & \rightarrow \text{ rifiuto } H_0 \ . \end{cases} \tag{2.3.11}$$

Osservazione 2.3.2: quando il vettore ξ dell'Osservazione 2.3.1 si riduce ad una sola componente ($\xi = x_i$), si può porre direttamente

$$\frac{\hat{\xi}-\overline{\xi}}{\hat{\sigma}_0\sqrt{PN^{-1}P^{+}}} = t_0 \ , \tag{2.3.12}$$

con t_0 estratta da una t di Student a $n-m$ gradi di libertà, se $H_0(\xi = \overline{\xi})$ è vera

$$t_0 \sim t_{n-m} \ .$$

In effetti, quadrando la (2.3.12) e tenendo conto che $t^2_{n-m} = F_{1,n-m}$ si ritrova la (2.3.9). Il test è allora eseguito accettando H_0 in caso che

$$|t| \leq t_\alpha \tag{2.3.13}$$

e rifiutando H_0 in caso contrario.

Osservazione 2.3.3: la (2.3.9) non solo permette di eseguire il test su $H_0(\xi = \overline{\xi})$, ma ci dà anche modo di definire delle regioni di confidenza per ξ. In particolare, fissato il livello di significatività ed il corrispondente valore critico F_α a $(r, n-m)$ *gradi di libertà*, noi chiameremo regione

di confidenza per ξ al livello α l'insieme dei vettori ξ che soddisfano la relazione

$$\frac{(1/r)(\hat{\xi}-\xi)^{+}(PN^{-1}P^{+})^{-1}(\hat{\xi}-\xi)}{\hat{\sigma}_0^2} \leq F_\alpha \ . \tag{2.3.14}$$

Osservazione 2.3.4: nei problemi di controllo, il vettore delle osservazioni viene campionato ripetutamente a tempi diversi; ad esempio, ad un istante t_1 si osserva

$$Y_{0(1)} = \begin{vmatrix} Y_{01(1)} \\ \vdots \\ Y_{0n(1)} \end{vmatrix}$$

e ad un istante t_2 si osserva un nuovo $Y_{0(2)}$.

Supponiamo che, essendo lo schema delle osservazioni invariato, debba ad ogni istante valere lo stesso modello parametrico, eventualmente con diversi valori di x,

$$y = Ax + a \ , \tag{2.3.15}$$

e lo stesso modello stocastico

$$C_{Y_{0(i)}Y_{0(i)}} = \sigma_0^2 Q \ . \tag{2.3.16}$$

Ai due diversi tempi corrispondono due diverse stime di x,

$$\begin{aligned} \hat{x}_1 &= N^{-1}A^{+}Q^{-1}(Y_{01}-a) \\ \hat{x}_2 &= N^{-1}A^{+}Q^{-1}(Y_{02}-a) \ : \end{aligned}$$

si osservi che per le ipotesi fatte le due stime $\hat{x}_1, \hat{x}_2$ sono diverse proprio perché $Y_{01} \neq Y_{02}$.

Ora facciamo l'ipotesi che le osservazioni Y_{01} e Y_{02} siano tra loro stocasticamente indipendenti; ne seguirà che anche $\hat{x}_1$ ed $\hat{x}_2$ sono indipendenti e quindi, sotto ipotesi di nomalità, posto $x_1 = E\{\hat{x}_1\}, x_2 = E\{\hat{x}_2\}$.

$$\hat{x}_1 - \hat{x}_2 = \mathcal{N}[x_1 - x_2,\ 2\sigma_0^2 N^{-1}] \ . \tag{2.3.17}$$

Il problema del controllo, cioè di verificare se il modello è cambiato dal tempo t_1 al tempo t_2, consiste proprio nel sottoporre a verifica l'ipotesi $H_0(x_1 = x_2)$. A tale scopo, accanto alla (2.3.17) possiamo usare la relazione

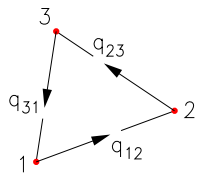
$$(n-m)\{\hat{\sigma}_{01}^2 + \hat{\sigma}_{02}^2\} = \sigma_0^2 \chi^2_{2(n-m)} \ \ : \tag{2.3.18}$$

inoltre la variabile (2.3.18), che è funzione solo degli scarti delle due compensazioni, è indipendente della (2.3.17).

Pertanto, se H_0 è vera, si ha

$$\frac{(1/2m)(\hat{x}_1 - \hat{x}_2)^+ N (\hat{x}_1 - \hat{x}_2)}{(1/2)(\hat{\sigma}_{01}^2 + \hat{\sigma}_{02}^2)} = F_0 \quad , \quad F_0 = F_{m,2(n-m)} \tag{2.3.19}$$

essendo questa un'estrazione da una F di Fisher a $(m, 2(n-m))$ gradi di libertà. Se $F_0 \leq F_\alpha, H_0$ è accettata, in caso contrario H_0 è rifiutata e si ritiene che ci sia stato un cambiamento significativo tra i due tempi.



Esempio 2.3.1: in un problema di controllo si misurano i dislivelli tra 3 punti, dei quali il punto 1 è da ritenersi fisso per motivi fisici, mentre i punti 2 e 3 sono effettivamente da controllare. I valori osservati ai due tempi t_1 e t_2 sono

	t_1	t_2	
$Y_{01} = q_{012}$	1,5734	1,5763	
$Y_{02} = q_{023}$	-0,4205	-0,4236	(in metri)
$Y_{03} = q_{031}$	-1,1526	-1,1525	

Volendo verificare le quote Q_2 e Q_3 conviene scrivere le equazioni in forma parametrica con

$$x = \begin{vmatrix} Q_2 \\ Q_3 \end{vmatrix} \; ; \qquad Y = Ax \; ;$$

$$A = \begin{vmatrix} 1 & 0 \\ -1 & 1 \\ 0 & 1 \end{vmatrix}, \quad N = A^+A = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix}, \quad N^{-1} = \tfrac{1}{3} \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} \quad .$$

Le soluzioni ottenute ai due tempi sono

	t_1	t_2	$\hat{x}(t_2) - \hat{x}(t_1)$
$\hat{x}_1$	1,57330	$1,5762\overline{3}$	$-29,\overline{3} \cdot 10^{-4}$
$\hat{x}_2$	1,15270	$1,1525\overline{6}$	$1,\overline{3} \cdot 10^{-4}$
$\hat{\sigma}_0^2$	$3 \cdot 10^{-8}$	$1,\overline{3} \cdot 10^{-8}$	(1 grado di libertà)

Il valore empirico della F per testare l'ipotesi

$$H_0 : x(t_2) - x(t_1) = 0 \; ; \tag{2.3.20}$$

è dato dalla (2.3.19) e vale nel nostro caso, notando che un fattore 1/2 va semplificato sia al numeratore che al denominatore,

$$F_0 = \frac{901,\overline{3} \cdot 10^{-8}}{4,3 \cdot 10^{-8}} = 208 \; .$$

Se H_0 è vera, F_0 deve essere un'estrazione da una F di Fisher a (2,2) gradi di libertà, ma fissato $\alpha = 5\%$ il corrispondente valore critico è 19, così che H_0 va rifiutata.

A questo punto ci si può chiedere se si sia mosso il punto 2, il punto 3 o entrambi. Per rispondere (con la significatività $\alpha = 5\%$) a questa domanda, possiamo ripetere il test F sulle singole componenti di x.

Punto 2: in questo caso $(PN^{-1}P^{+}) = [2/3]$, così che

$$\begin{aligned} \text{numeratore} &= (1/r)\{\hat{x}_1(t_1) - \hat{x}_1(t_2)\}^{+}(PN^{-1}P^{+})^{-1}\{\hat{x}_1(t_1) - \hat{x}_1(t_2)\} = \\ &= 1290,\overline{6} \cdot 10^{-8} \\ \text{denominatore} &= \hat{\sigma}^2_{01} + \hat{\sigma}^2_{02} = 4,\overline{3} \cdot 10^{-8} \\ F_0 &= 297,85 \end{aligned}$$

che è assai più grande del corrispondente valore critico ($\alpha = 5\%$) per la F a (1,2) gradi di libertà, ovvero a 18,51.

Perciò l'ipotesi $H_0 : x_1(t_1) = x_1(t_2)$ è senz'altro da rifiutare, cioè il punto 2 si è mosso significativamente.

Punto 3: anche in questo caso, per la particolare simmetria della matrice normale, risulta $(PN^{-1}P^{+}) = [2/3]$, così che

$$F_0 = 0,62 \ :$$

poiché tale valore è ben al di sotto del valore critico della F a (1,2) gradi di libertà, H_0 va accettata, cioè il punto 3 non si è significativamente mosso.

2.4 Scelta del modello di regressione lineare

Come già discusso nel paragrafo 8 del Quaderno n. 2, spesso ci si trova a valutare in maniera puramente empirica una legge che lega una (o più) variabile criterio y ad un gruppo di variabili $\underline{t} = (t_1, \ldots, t_p)$: sotto opportune ipotesi la dipendenza può essere linearizzata e, supposto di aver scelto l'origine delle t in modo tale che

$$\overline{t}_k = \frac{1}{n}\sum_{i=1}^{n} t_{ki} = 0 \ , \tag{2.4.1}$$

si ha il modello

$$\begin{cases} y = c_0 + \sum_{k=1}^{p} c_k t_k \\ Y_{0i} = y_i + \varepsilon_i \\ C_{\varepsilon\varepsilon} = \sigma_0^2 I \ . \end{cases} \tag{2.4.2}$$

I parametri incogniti sono $x = c_0 \quad , \quad \underline{x}_0 = \begin{vmatrix} c_1 \\ \vdots \\ c_p \end{vmatrix}$, ed essi vengono stimati tramite il principio dei m.q.

Se il modello (2.4.2) è puramente empirico, le variabili $t_k (k = 1, \ldots, p)$ vengono messe in regressione solo in base a ragionamenti qualitativi di carattere generale sul processo analizzato.

Qualsiasi siano le variabili concomitanti introdotte, è chiaro che per via degli scarti stocastici ε_i si troveranno per i rispettivi coefficienti c_k delle stime diverse da zero: tuttavia resta aperto il problema di definire per quali $\hat{c}_k$ l'ipotesi

$$H_0 : (c_k = E\{\hat{c}_k\} = 0) \tag{2.4.3}$$

possa essere significativamente rigettata.

In effetti, dove H_0 dovesse essere accettata, la dipendenza della variabile criterio da t_k potrebbe essere messa in dubbio.

Pertanto ci proponiamo di selezionare tra la originali $t_1, \ldots, t_p$ un sottoinsieme di variabili concomitanti (ad esempio $t_1, \ldots, t_r$), per cui H_0 va rifiutata, mentre l'aggiunta di una delle escluse (ad esempio $t_{r+1}, \ldots, t_p$) porterebbe ad accettare H_0 stessa.

In questo modo si arriva a definire nel nostro modello empirico un sottomodello

$$y = c_0 + c_1 t_1 + \ldots + c_r t_r$$

in cui entrano solo le variabili essenziali, cioè solo quelle che in base all'analisi dei dati influenzano veramente y.

Per ottenere il risultato presentiamo un metodo detto di selezione all'indietro.

In base a tale metodo si parte calcolando la regressione di y su tutte le variabili $t_1, \ldots, t_p$ ottenendo il vettore $\hat{y}_{(p)}$. Successivamente si prova a porre l'ipotesi, per ogni variabile concomitante,

$$H_0 : (c_k = 0) \qquad k = 1, 2, \ldots, p \ . \tag{2.4.4}$$

Si hanno così p ipotesi, ciascuna delle quali può essere valutata considerando la corrispondente equazione $c_k = 0$ come un vincolo; il test sarà quindi eseguito per mezzo di una F di Fisher.

Ad esempio sia $k = p$, cioè si voglia verificare se

$$c_p = 0 \ . \tag{2.4.5}$$

Si cerca allora il vettore $\hat{y}_{(p-1)}$ che esprime la regressione su $t_1, \ldots, t_{p-1}$, ovvero su tutte le p, ma imponendo il vincolo (2.4.5).

Ricordando la teoria del paragrafo 2.8 del Quaderno n. 2 notiamo che

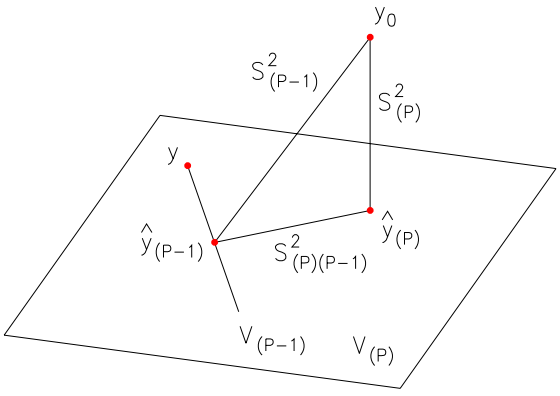


Figura 2.4.1:

$Y_0 - \hat{y}_{(p)}$	rappresenta il residuo non spiegato dalla regressione su p variabili
$Y_0 - \hat{y}_{(p-1)}$	rappresenta il residuo non spiegato dalla regressione su $p-1$ variabili solamente
$\hat{y}_{(p)} - \hat{y}_{(p-1)}$	rappresenta il contributo di spiegazione di Y_0 dovuto all'introduzione di t_p.

Inoltre è [2] (vedi (2.8.14) del Quaderno n. 2)

$$\begin{aligned} |Y_0 - \hat{y}_{(p)}|^2 &= S^2_{(p)} = (n-p-1)\hat{\sigma}_{0(p)} = S^2_G - \sum_{k=1}^{p}\sum_{i=1}^{n} \hat{c}_k t_{ki} Y_{0i} \\ |Y_0 - \hat{y}_{(p-1)}|^2 &= S^2_{(p-1)} = (n-p)\hat{\sigma}^2_{0(p-1)} = \\ &= S^2_G - \sum_{k=1}^{p-1}\sum_{i=1}^{n} \hat{c}_k t_{ki} Y_{0i} \\ |\hat{y}_{(p)} - \hat{y}_{(p-1)}|^2 &= |Y_0 - \hat{y}_{(p-1)}|^2 - |Y_0 - \hat{y}_{(p)}|^2 \ : \end{aligned}$$

all'ultima equazione, se H_0 è vera, cioè se $y = E\{Y_0\} \in V(p-1)$, corrisponde anche la rispettiva relazione distribuzionale

$$\sigma_0^2 \chi_1^2 = \sigma_0^2 \chi^2_{n-p} - \sigma_0^2 \chi^2_{(n-p-1)} \ . \tag{2.4.6}$$

Quindi H_0 può essere testata in base al valore campionario

$$\frac{|\hat{y}_{(p)} - \hat{y}_{(p-1)}|^2}{|Y_0 - \hat{y}_{(p)}|^2/(n-p-1)} = \frac{(n-p)\hat{\sigma}^2_{0(p-1)} - (n-p-1)\hat{\sigma}^2_{0(p)}}{\hat{\sigma}^2_{0(p)}} = F_0 \tag{2.4.7}$$

se H_0 è giusta, F_0 è un'estrazione da una $F_{1,n-p-1}$ e può quindi essere confrontata col corrispondente valore critico F_α , $P\{F_0 \le F_\alpha\} = 1-\alpha$.

[2] si noti che i $\hat{c}_k$ di questa seconda equazione sono diversi dai $\hat{c}_k$ precedenti che derivano da una diversa compensazione, senza il vincolo $c_p = 0$.

Notiamo che il valore critico F_α è fissato una volta che si sia scelto il valore critico α, esendo i gradi di libertà necessariamente $(1, n-p-1)$.

Naturalmente di valori empirici F_0 ne possiamo ottenere p, escludendo una volta t_p, una volta t_{p-1} ecc., finché si è provato ad escludere ognuna delle variabili concomitanti.

Se tutti i vari F_{0k} sono maggiori di F_α, si rifiuta H_0 per tutte le t_k, cioè il modello mostra una dipendenza significativa da tutte le variabili; se invece una o più F_{0k} sono inferiori ad F_α si cerca la minore tra tutte, che è quella che mostra di essre più significativamente prossima a zero, si esclude la variabile corrispondente e si ricomincia il processo di selezione con le $p-1$ variabili rimaste:

$$\begin{cases} F_{0m} = \min\limits_k \{F_{0k}\} \\ F_{0m} > F_\alpha \rightarrow \text{ rifiuto } H_0\text{, tengo il modello} \\ F_{0m} < F_\alpha \rightarrow \text{ elimino } t_m \text{ e ricomincio l'analisi.} \end{cases} \tag{2.4.8}$$

Osservazione 2.4.1: il modello a cui si perviene con questo procedimento non è mai sicuro: in particolare il metodo potrebbe dare risposte sbagliate tanto più quanto una variabile, almeno per i valori che assume in quell'esperimento, è prossima ad essere combinazione lineare delle altre. Ad esempio, se un processo è analizzato in funzione di tempo e temperatura,

$$y = c_0 + c_1 t + c_2 T \ ,$$

ma i valori empirici di T_i sono prossimi ad essere linearmente dipendenti da t_i, cioè

$$T_i \sim a_0 + a_1 t_i \ ,$$

allora l'effetto di t_i e T_i su y si confonde ed è più facile arrivare a scartare una variabile che è invece quella che governa il processo.

Casi di questo tipo sono sempre denunciati da un cattivo condizionamento della matrice C_{tt} e si manifestano in difficoltà numeriche nel calcolo dell'inversa C_{tt}^{-1} (tipicamente il det C_{tt} risulta molto piccolo).

In questi casi è meglio affidarsi all'esperienza che può dire quale variabile sia importante per spiegare y o, quando ciò non sia possibile, acquisire nuovi dati in condizioni diverse in modo che cessi l'accoppiamento lineare tra le variabili indipendenti.

Osservazione 2.4.2: talvolta si considera, al posto dell'indice (2.4.7), l'indice

$$\frac{|\hat{y}(p) - \hat{y}(p-1)|^2}{|Y_0 - \hat{y}(p-1)|^2} = R^2_{(p)(1..p-1)} \tag{2.4.9}$$

che rappresenta evidentemente la diminuzione percentuale del modulo quadrato del vettore degli scarti, dovuta all'introduzione della variabile t_p.

L'indice (2.4.9) è detto *coefficiente di correlazione parziale* di Y_0 con t_p, tolta la dipendenza dalle variabili $t_1, \ldots, t_{p-1}$. È chiaro che un basso valore di $R^2_{(p)(1..p-1)}$ indica che l'aggiunta di t_p non porta una significativa informazione nuova alla spiegazione di Y_0, oltre a quella già fornita dalle variabili $t_1, \ldots, t_{p-1}$.

D'altro canto $F_{0(p)}$ ed $R^2_{(p)(1..p-1)}$ sono legate tra loro dalla relazione algebrica

$$F_0 = (n - p - 1)\frac{R^2}{1 - R^2} \tag{2.4.10}$$

così che ogni test basato su un basso valore di R^2 equivale al test su F_0 già visto.

Osservazione 2.4.3: il procedimento di selezione all'indietro non è l'unico capace di identificare un modello significativo di regressione lineare. Ad esempio si può procedere selezionando in avanti nel seguente modo: si parte calcolando la correlazione di Y_0 con ogni variabile $t_1, \ldots t_2$ tra le possibili variabili concomitanti.

Si sceglie poi, supponiamo sia t_1, la variabile che mostra la correlazione maggiore.

Si prova poi a costruire la regressione con tutte le coppie $(t_1, t_2), (t_1, t_3), \ldots, (t_1, t_p)$ e si calcolano i coefficienti di correlazione par-

ziale dovuti all'introduzione di ognuna delle variabili $t_2, \ldots, t_p$, tolta la dipendenza da t_1: si sceglie la variabile che dà il valore più alto di R^2, supponiamo sia t_2, e si passa quindi ad esaminare la regressione per le triplette di tipo (t_1, t_2, t_k).

Il procedimento si ferma al passo in cui il coefficiente R^2 massimo ed il corrispondente F_0 calcolato con la (2.4.10), non supera il valore critico F_α.

Esempio 2.4.1: riprendiamo l'Esempio 9.1 del Quaderno n. 2 e verifichiamo se entrambe, una o nessuna delle variabili t, T influiscono significativamente sulla quota $Q = y$.

Ricordiamo che posto

$$\begin{aligned} t &= \overline{t} + \tau \\ T &= \overline{T} + \theta \end{aligned}$$

il modello generale di regressione in esame è

$$y = c_0 + c_1\tau + c_2\theta \ .$$

Passo 0 È quello già svolto nel paragrafo 8 del Quaderno n. 2
$\hat{c}_0 = \ 129,6583$
$\hat{c}_1 = \ -0,0215$
$\hat{c}_2 = \ 0,4812$
$S_R^2 = 3 \cdot \hat{\sigma}_0^2 = 0,1172$

Passo 1 a) Eliminando τ $\qquad y = c_0 + c_2\theta$
$\hat{c}_0 = 129,6583$
$\hat{c}_2 = \frac{C_{\theta y}}{C_{\theta\theta}} = 0,4792 \left(= \frac{S_{\theta y}}{S_\theta^2}\right)$
$S_R^2 = 4 \cdot \hat{\sigma}_0^2 = S_G^2 - \hat{c}_2 S_{\theta y} = 0,1474$
$F_0 = \frac{0,1474 - 0,1172}{0,1172/3} = 0,7723$
b) Eliminando θ $\qquad y = c_0 + c_1\tau$
$\hat{c}_0 = 129,6583$
$\hat{c}_1 = \frac{C_{\tau y}}{C_{\tau\tau}} = 0,0713 \qquad \left(= \frac{S_{\tau y}}{S_\tau^2}\right)$
$S_R^2 = 4 \cdot \hat{\sigma}_0^2 = S_G^2 - \hat{S}_{\tau y} = 32,3213$
$F_0 = \frac{32,3213 - 0,1172}{0,1172/3} = 824,3366$.

I valori campionari di F_0 vanno confrontati con il valore critico F_α a (1,3) gradi di libertà.

Preso $\alpha = 5\%$ è $F_\alpha = 10,13$, per cui si ha nel caso a) $F_0 < F_\alpha$, nel caso b) $F_0 > F_\alpha$

quindi è chiaro che occorre accettare H_0 per il caso a) e non nel caso b).

2.5 L'analisi di varianza

In questo paragrafo studiamo l'applicazione del test alle stime dei m.q. per un problema particolare:

> dato un insieme di osservazioni (stocasticamente) indipendenti tratte da distribuzioni normali con ugual varianza e con medie che potrbbero dipendere da uno o più fattori concomitanti, ci si chiede di valutare se tale dipendenza esiste veramente ed eventualmente per quale dei fattori essa sia significativa.

Di solito la trattazione di questo problema viene fatta in modo distinto in base al numero dei fattori che si considerano concomitanti alla formazione dei valori medi: tuttavia la trattazione è metodologicamente identica quando essa viene ridotta ad un problema di m.q. e quando si supponga che le varie cause $A, B, C, \ldots$ agiscano sui valori medi μ in modo puramente additivo, detto anche *senza interazione*.

$$\mu(A, B, C, \ldots) = \mu_A + \mu_B + \mu_C + \ldots \ . \tag{2.5.1}$$

a) Classificazione a una via

Cominciamo col caso più semplice, in cui il fattore concomitante sia uno solo, A, e che, nei dati analizzati, A sia specificato da p possibili valori $A_1, A_2, \ldots, A_p$ [3]

Quando $A = A_1$ si hanno $j = 1, \ldots, n_1$ osservazioni, quando $A = A_2$ si hanno $j = 1, \ldots, n_2$ osservazioni e così via. Le osservazioni possono

[3] Questi *valori* non sono affatto necessariamente numerici: ad esempio A potrebbe essere un colore, e si potrebbe avere A_1 = giallo, A_2 = rosso, A_3 = blu.

allora essere naturalmente rappresentate in funzione di due indici i, j ed il modello sottostante diviene:

$$Y_{0ij} \ , \ \begin{matrix} i = 1, 2, \ldots, p \\ j = 1, 2, \ldots, n_i \end{matrix} \qquad \text{insieme delle osservazioni} \qquad (2.5.2)$$

$$E\{Y_{0ij}\} = \mu_i \qquad \text{modello deterministico} \qquad (2.5.3)$$

$$C_{Y_0 Y_0} = \sigma_0^2 I \qquad \text{modello stocastico.} \qquad (2.5.4)$$

Questo schema può essere trattato con le formule generali dei m.q. sebbene ciò non sia comodo perché le osservazioni, anziché organizzarsi spontaneamente in un vettore, sono piuttosto più semplicemente rappresentate da una tabella (cfr. Tab.2.5.1).

Osservazione 2.5.1: per motivi storici due valori di Y_0 in colonne diverse vengono detti osservazioni con *trattamenti* diversi: due valori di Y_0 sulla stessa colonna, sono detti *replicazioni* di osservazioni con lo stesso trattamento. Si noti inoltre che la tabella dei valori osservati non è una matrice in quanto in generale le colonne hanno lunghezze diverse. Infine in Tab. 2.5.1 si sono aggiunte due righe che indicano medie e momenti semplici del 2° ordine per colonna, adottando per questi la simbologia $\overline{y}_i, M_{2i}$.

$\{Y_{0ij}\}$	$A_i(i=1)$	$A_2(i=2)$	$\cdot$	$A_p(i=p)$	
$j=1$	Y_{011}	Y_{021}	$\cdot$	Y_{0p1}	
$j=2$	Y_{012}	Y_{022}	$\cdot$	Y_{0p2}	
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	
$j=n_1$	Y_{01n_1}	$\cdot$	$\cdot$	$\cdot$	Tab. 2.5.1
$\cdot$		$\cdot$	$\cdot$	$\cdot$	
$j=n_p$		$\cdot$	$\cdot$	Y_{0pn_p}	
$\cdot$		$\cdot$	$\cdot$		
$j=n_2$		Y_{02n_2}	$\cdot$		
$\overline{y}_i$	$\frac{1}{n_1}\sum_{j=1}^{n_1} Y_{01j}^2$	$\frac{1}{n_2}\sum_{j=1}^{n_2} Y_{02j}^2$	$\cdot$	$\frac{1}{n_p}\sum_{j=1}^{n_p} Y_{0pj}$	$\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{p} n_i \overline{y}_i$
M_{2i}	$\frac{1}{n_1}\sum_{j=1}^{n_1} Y_{01j}^2$	$\frac{1}{n_2}\sum_{j=1}^{n_2} Y_{02j}^2$	$\cdot$	$\frac{1}{n_p}\sum_{j=1}^{n_p} Y_{0pj}^2$	$M_2 = \frac{1}{n}\sum_{i=1}^{p} n_i M_{2i}$

Per trovare gli stimatori $\hat{\mu}_i$ dei parametri μ_i, rifacciamoci direttamente al principio del m.q.; a causa dell'ipotesi (2.5.4) esso diviene semplicemente

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{0ij}-\hat{\mu}_i)^2 = \min \ . \tag{2.5.5}$$

Osserviamo anche che in base alla teoria dei m.q., una volta determinati $\hat{\mu}_i$, sarà

$$U^+U = (n-p)\hat{\sigma}_0^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{0ij}-\hat{\mu}_i)^2 \tag{2.5.6}$$

dove $n = \sum_{i=1}^{p} n_i$.

Derivando la (2.5.5) rispetto a $\hat{\mu}_i$ ed uguagliando a zero si trova

$$-2\sum_{j=1}^{n_i}(Y_{0ij}-\hat{\mu}_i) = 0$$

ovvero

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{0ij} = \overline{y}_j \ , \tag{2.5.7}$$

cioè $\hat{\mu}_i$ sono proprio le medie per colonna.

Si può notare che in tal caso

$$\sum_{j=1}^{n_i}(Y_{0ij}-\overline{y}_i)^2 = n_i(M_{2i}-\overline{y}_i^2) = \sum_{j=1}^{n_i} Y_{0ij}^2 - n_i\overline{y}_i^2 \ . \tag{2.5.8}$$

Ma allora dalla (2.5.6) si ha

$$U^+U = (n-p)\hat{\sigma}_0^2 = \sum_{i,j} Y_{0ij}^2 - \sum_{i=1}^{p} n_i \overline{y}_i^2 =$$
$$= nM_2 - \sum_{i=1}^{p} n_i \overline{y}_i^2 \ . \qquad (2.5.9)$$

Inoltre, supponendo la normalità di tutte le variabili, si ha anche

$$U^+U = (n-p)\hat{\sigma}_0^2 = \chi^2_{(n-p)}\sigma_0^2 \ . \qquad (2.5.10)$$

A partire da questo schema si possono porre diversi tipi di ipotesi da verificare in base alle osservazioni.

Ad esempio, ci si potrebbe chiedere se per due trattamenti i valori medi, per esempio $\overline{y}_1, \overline{y}_2$ non indichino che le medie sottostanti μ_1, μ_2 hanno valori diversi: ma questo problema può essere semplicemente risolto sulla base di un test di confronto tra medie di campioni normali con ugual varianza, la cui teoria si trova nel paragrafo 1.6 di questo quaderno. Oppure ci si potrebbe chiedere se un certo trattamento, ad esempio A_1, abbia media diversa da quella di tutti gli altri, cioè se $\mu_1 \neq \mu_2 = \mu_3 \ldots = \mu_p$. Questo problema può essere risolto accorpando tutti i dati dei trattamenti $A_2, A_3, \ldots, A_p$, prendendone la media e poi testando se questa è significativamente diversa da y_1; ci si è così ricondotti al problema del confronto tra due medie.

Diverso è il caso, che qui tratteremo, in cui si voglia verificare se globalmente i trattamenti abbiano un qualche effetto, cioè se
$\mu_1 \neq \mu_2 \neq \ldots \neq \mu_p$, ovvero più in generale se tra le μ_i alcune sono diverse dalle altre. In tal caso possiamo porre l'ipotesi semplice

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p \qquad (2.5.11)$$

e vedere se essa è contraddetta o no dalle stime empiriche $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_p$, ad un livello di significatività prefissato α. Notiamo che la verifica di (2.5.11) può essere vista come la verifica della correttezza di $p-1$ vincoli, sui parametri dello schema di m.q. (2.5.2), (2.5.3), (2.5.4).

$$\begin{array}{c} \mu_1 = \mu_2 \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \mu_{p-1} = \mu_p \end{array} \quad : \tag{2.5.12}$$

si potrà perciò applicare la teoria dei test sui vincoli del paragrafo 2.4.

A questo scopo occorre ricavare la stima di $\hat{\mu} = \hat{\mu}_1 = \hat{\mu}_2 = \ldots = \hat{\mu}_p$, imponendo un tale vincolo nel principio (2.5.5), e successivamente trovare

$$\overline{U}^+ \overline{U} = (n - p + v)\hat{\overline{\sigma}}_0^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{0ij} - \hat{\mu})^2 \ , \tag{2.5.13}$$

dove abbiamo indicato con una soprasegnatura le quantità vincolate e con v il numero dei vincoli $v = p - 1$. Ma se si suppone che tutte le Y_0 abbiano la stessa media, la sua stima, anche di m.q., è semplicemente

$$\hat{\mu} = \overline{\overline{y}} \ , \tag{2.5.14}$$

così che dalla (2.5.13) si trova

$$\begin{aligned} \overline{U}^+ \overline{U} &= (n-1)\hat{\overline{\sigma}}_0^2 = \sum_{i,j} (Y_{0ij} - \overline{\overline{y}})^2 = \\ &= \sum Y_{0ij}^2 - n\overline{\overline{y}}^2 = nM_2 - n\overline{\overline{y}}^2 \ . \end{aligned} \tag{2.5.15}$$

Si noti che la (2.5.15) in fondo non è altro che la stima corretta di una varianza campionaria: infatti risulta anche

$$\overline{U}^+ \overline{U} = (n-1)\hat{\overline{\sigma}}_0^2 = \chi^2_{n-1}\sigma_0^2 \ , \tag{2.5.16}$$

naturalmente se l'ipotesi (2.5.11) è corretta.

Dunque se H_0 è giusta, si ha che (cfr. (2.5.10), (2.5.16))

$$\begin{cases} \overline{U}^{+}\overline{U} - U^{+}U &= \displaystyle\sum_{i=1}^{p} n_i \overline{\overline{y}}_i^2 - n\overline{\overline{\overline{y}}}^2 = \chi^2_{p-1}\sigma_0^2 \\ U^{+}U &= nM_2 - \displaystyle\sum_{i=1}^{p} n_i \overline{\overline{y}}_i^2 = \chi^2_{n-p}\sigma_0^2 \end{cases} ; \qquad (2.5.17)$$

ed ancora, per l'indipendenza delle due χ^2 della (2.5.17),

$$F_0 = \frac{n-p}{p-1} \frac{\displaystyle\sum_{i=1}^{p} n_i \overline{\overline{y}}_i^2 - n\overline{\overline{\overline{y}}}^2}{nM_2 - \displaystyle\sum_{i=1}^{p} n_i \overline{\overline{y}}_i^2}$$

è un'estrazione da una F di Fisher

$$F_0 \sim F_{\nu,\lambda} \qquad \nu \text{ (num.)} = p-1 \quad , \quad \lambda \text{ (denom.)} = n-p \; . \qquad (2.5.18)$$

Concludendo, trovato il valore critico F_α, a $(p-1,\ n-p)$ gradi di libertà, si conclude che

$$\begin{aligned} F_0 \leq F_\alpha \quad &\rightarrow \quad H_0 \text{ è accettata } \rightarrow \\ &\rightarrow \text{ le medie sono uguali } = \\ &= \text{ i trattamenti non hanno effetto} \end{aligned}$$

$$\begin{aligned} F_0 \leq F_\alpha \quad &\rightarrow \quad H_0 \text{ è rifiutata } \rightarrow \\ &\rightarrow \text{ le medie non sono tutte uguali } = \\ &= \text{ i trattamenti hanno effetto} \end{aligned}$$

Esempio 2.5.1: campi di ugual estensione e nella stessa zona sono trattati con tre diversi concimi A_1, A_2, A_3, producendo Y_0 quintali di frumento per anno. Si osservano le produzioni di quattro anni (cfr. Tabella 2.5.2) e ci si chiede se i risultati sono tali da poter affermare che i concimi hanno effetto, al livello di significatività $\alpha = 5\%$.

Se H_0 è vera, allora

$$F_0 = \frac{12-3}{2}\frac{4 \cdot 7205 - 12 \cdot 2401}{12 \cdot 2402,1\overline{6} - 4 \cdot 7205} = 6$$

è un'estrazione da una F a

$$\text{g.l. numeratore } = 2 \quad , \quad \text{g.l. denominatore } = 9 \ :$$

con $\alpha = 5\%$ il corrispondente valore critico è

$$F_\alpha = 4,26$$

cioè

$$F_0 > F_\alpha$$

il che ci porta a concludere che H_0 è falsa e che quindi i trattamenti hanno effetto.

Anni	Trattamenti			
	A_1	A_2	A_3	
1°	48	47	49	
2°	49	49	51	
3°	50	48	50	
4°	49	48	50	
$\overline{y}_i$	49	48	50	$\overline{\overline{y}} = 49$
M_{2i}	2401,5	2304,5	2500,5	$M_2 = 2402,1\overline{6}$

Tab. 2.5.2

b) Classificazione a due vie

Studiamo il caso i cui i risultati di osservazioni Y_0 siano classificati secondo i valori argomentali assunti da due fattori concomitanti A, B

$$A = A_1, A_2, \ldots, A_q$$
$$B = B_1, B_2, \ldots, B_p \ .$$

Supponiamo per semplicità che per ogni "cella", ovvero per ogni coppia di valori $(A_i,\ B_j)$, vi sia una sola osservazione Y_{0ij}. Notiamo anche che lo schema che elaboreremo qui di seguito vale pure nel caso che ogni cella (i, j) contenga r repliche, purché tale numero sia lo stesso per tutte le celle: si prenderà allora la media in ogni cella e si passerà a considerare questa come osservazione Y_{0ij}. Supponiamo ora che le osservazioni seguano il modello

$$\begin{cases} E\{Y_{0ij}\} = \alpha_i + \beta_j & (i = 1, \ldots, q) \\ & (j = 1, \ldots, p) \\ C_{Y_0 Y_0} = \sigma_0^2 I \ . \end{cases} \tag{2.5.19}$$

Notiamo che la prima delle (2.5.19) corrisponde alla (2.5.1) ed in particolare sottintende l'ipotesi che non vi siano interazioni non lineari tra i fattori A e B.

Tra i vari problemi che si potrebbero analizzare, vogliamo qui considerare la verifica dell'ipotesi che A abbia influenza sulle osservazioni: questo al solito viene fatto ponendo come ipotesi fondamentale H_0 l'opposto, cioè che A non abbia alcun effetto

$$H_0 = \{\alpha_1 = \alpha_2 = \ldots = \alpha_q\} \tag{2.5.20}$$

e si va poi a vedere se i vincoli (2.5.20) sono accettati al livello di significatività α prefissato.

Osservazione 2.5.2: prima di proseguire nel trattamento analitico occorre fare una precisazione: il modello deterministico in (2.5.19) è sovraparametrizzato. Infatti, sia $\hat{\alpha}_i, \hat{\beta}_j$ un insieme di stime di m.q., allora è

ovvio che altrettanto saranno $\hat{\alpha}_i + c, \hat{\beta}_j - c$ per ogni costante c. Ciò significa che se tentiamo di determinare tutti gli α e i β da (2.5.19), troveremo necessariamente un sistema normale singolare. Ciò tuttavia ci interessa poco, poiché in realtà ciò di cui abbiamo bisogno per costruire il test sul vincolo (2.5.20) sono gli scarti delle equazioni, ovvero $U_{ij} = Y_{0ij} - \hat{\alpha}_i - \hat{\beta}_j$, che restano gli stessi qualsiasi siano le particolari stime di α e β.

Ne deriva che eseguiremo i conti della compensazione imponendo un vincolo arbitrario che ci permetta di selezionare una particolare soluzione $\hat{\alpha}, \hat{\beta}$: più precisamente imporremo che

$$\sum_{j=1}^{p} \hat{\beta}_j = 0 \ . \tag{2.5.21}$$

Prima di passare alla compensazione, conveniamo di usare la seguente simbologia standard nella letteratura che tratta analisi di varianza

$$\begin{array}{lll} \overline{y}_{i\cdot} = (1/p)\sum_{j=1}^{p} Y_{0ij} & = & \text{medie per riga} \\ \overline{y}_{\cdot j} = (1/q)\sum_{i=1}^{q} Y_{0ij} & = & \text{medie per colonna} \\ \overline{\overline{y}} = (1/q)\sum_{i=1}^{q} \overline{y}_{i\cdot} = \frac{1}{p}\sum_{j=1}^{p} \overline{y}_{\cdot j} & = & \text{media generale} \ : \end{array}$$

inoltre indicheremo come al solito con M_2 il momento semplice totale di ordine 2

$$M_2 = \frac{1}{pq} \sum_{i=1}^{q} \sum_{j=1}^{p} Y_{0ij}^2 \ .$$

La situazione è riassunta nella Tabella 2.5.3.

	B_1	B_2	...	B_p	
A_1	Y_{011}	Y_{012}	...	Y_{01p}	$\overline{y}_{1\cdot}$
A_2	Y_{021}	Y_{022}	...	Y_{02p}	$\overline{y}_{2\cdot}$
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
A_q	Y_{0q1}	Y_{0q2}	...	Y_{0qp}	$\overline{y}_{q\cdot}$
	$\overline{y}_{\cdot 1}$	$\overline{y}_{\cdot 2}$	...	$\overline{y}_{\cdot p}$	$\overline{\overline{y}}$

Tab. 2.5.3

Ciò detto passiamo a stimare α_i, β_j, minimizzando la somma di quadrati

$$U^+U = \sum_{i=1}^{q}\sum_{j=1}^{p}(Y_{0ij} - \hat{\alpha}_i - \hat{\beta}_j)^2 \ . \tag{2.5.22}$$

È facile vedere che il sistema normale corrispondente è

$$\begin{cases} p\hat{\alpha}_i + \sum_{j=1}^{p}\hat{\beta}_j = p\overline{y}_{i\cdot} \\ \sum_{i=1}^{q}\hat{\alpha}_i + q\hat{\beta}_j = q\overline{y}_{\cdot j} \end{cases} \ . \tag{2.5.23}$$

Si può così verificare che, come già previsto nella Osservazione 2.5.2, $(\alpha_i = c,\ \beta_j = -c)$ costituisce una soluzione non nulla del sistema omogeneo associato, cioè che (2.5.23) è singolare. Scegliamo una soluzione particolare imponendo la (2.5.21), il che ci dà

$$\begin{cases} \hat{\alpha}_i = \overline{y}_{i\cdot} \\ \hat{\beta}_j = \overline{y}_{\cdot j} - \overline{\overline{y}} \ . \end{cases} \tag{2.5.24}$$

Occorre ora calcolare (2.5.22):

$$\begin{aligned} U^+U &= \sum_{i,j}[(Y_{0ij} - \overline{y}_{i\cdot}) - (\overline{y}_{\cdot j} - \overline{\overline{y}})]^2 = \\ &= \sum_{i,j}(Y_{0ij} - \overline{y}_{i\cdot})^2 - 2\sum_{i,j}(Y_{0ij} - \overline{y}_{i\cdot})(\overline{y}_{\cdot j} - \overline{\overline{y}}) + \\ &+ \sum_{i,j}(\overline{y}_{\cdot j} - \overline{\overline{y}})^2 = \\ &= \sum_{i,j}(Y_{0ij} - \overline{y}_{i\cdot})^2 - q\sum_{j}(\overline{y}_{\cdot j} - \overline{\overline{y}})^2 = \\ &= \sum_{i,j}Y_{0ij}^2 - p\sum_{i=1}^{q}\overline{y}_{i\cdot}^2 - q\sum_{j=1}^{p}\overline{y}_{\cdot j}^2 + qp\overline{\overline{y}}^2 = \end{aligned} \tag{2.5.25}$$

$$= \quad qpM_2 - p\sum_{i=1}^{q}\overline{y}_{i\cdot}^2 - q\sum_{j=1}^{p}\overline{y}_{\cdot j}^2 + qp\overline{\overline{y}}^2 \ .$$

Notiamo ancora che, avendo determinato $p+q-1$ parametri (la condizione (2.5.21) infatti abbassa il numero dei parametri determinati), si ha anche il risultato distribuzionale

$$\begin{cases} U^+U = \sigma_0^2\chi_\nu^2 \\ \nu = pq - (p+q-1) = (p-1)(q-1) \ . \end{cases} \tag{2.5.26}$$

Ora occorre rifare le stime sotto il vincolo (2.5.20), ovvero

$$\alpha_1 = \alpha_2 = \ldots = \alpha_q = \alpha \ :$$

ma in tal caso il modello deterministico è semplicemente

$$E\{Y_{0ij}\} = \alpha + \beta_j \qquad (\sum \beta_j = 0) \ . \tag{2.5.27}$$

Come si vede si hanno p parametri β col vincolo $\sum \beta_j = 0$, e l'ulteriore parametro indipendente α, e ciò equivale a p parametri senza alcun vincolo, definiti da

$$\gamma_j = \alpha + \beta_j \ .$$

In questo caso lo schema (2.5.27) torna ad essere quello della classificazione ad una via per cui sappiamo già che

$$\hat{\gamma}_j = \hat{\alpha} + \hat{\beta}_j = y_{\cdot j} \tag{2.5.28}$$

e che (cfr. (2.5.9))

$$\begin{aligned} \overline{U}^+\overline{U} \ &= \ qpM_2 - q\sum_{j=1}^{p}\overline{y}_{\cdot j}^2 = \\ &= \ \chi^2_{pq-p}\sigma_0^2 = \chi^2_{p(q-1)}\sigma_0^2 \ . \end{aligned} \tag{2.5.29}$$

Se l'ipotesi H_0 è corretta, cioè se valgono i $q-1$ vincoli (2.5.20), e quindi A non influenza le osservazioni, si deve avere [4]

$$F_0 = (p-1)\frac{p\sum_{i=1}^{q}\overline{y}_{i\cdot}^2 - qp\overline{\overline{y}}^2}{pqM_2 - p\sum_{i=1}^{q}\overline{y}_{i\cdot}^2 - q\sum_{j=1}^{p}\overline{y}_{\cdot j}^2 + pq\overline{\overline{y}}^2} \tag{2.5.30}$$

con

$$F_0 \sim F_{\nu,\lambda} \qquad \nu \text{ (num.)} = q-1,\ \lambda \text{ (den.)} = (p-1)(q-1)\ . \tag{2.5.31}$$

In conclusione, trovato il valore critico F_α a $[(q-1),(p-1)(q-1)]$ gradi di libertà, H_0 è accettata o rifiutata a seconda che $F_0 \le F_\alpha$ oppure $F_0 > F_\alpha$.

Esempio 2.5.2: quattro diverse varietà di grano $B_1,\ B_2,\ B_3,\ B_4$ sono impiegate in cinque terreni a diversa composizione. Si vuole valutare, con la significatività del 5%, se la produzione (riportata in tabella in quintali di grano per ettaro) è influenzata del tipo terreno o dalla semente usata.

	Qualità di grano				
Terreni	B_1	B_2	B_3	B_4	$\overline{y}_{i\cdot}$
A_1	32,3	33,3	30,8	29,3	31,425
A_2	34,0	33,0	34,3	26,0	31,825
A_3	34,3	36,3	35,3	29,8	33,925
A_4	35,0	36,8	32,3	28,0	33,025
A_5	36,5	34,5	35,8	28,8	33,900
$\overline{y}_{\cdot j}$	34,42	34,78	33,70	28,38	$32{,}82 = \overline{\overline{y}}$

Tab. 2.5.4

Si pone ora

$$H_0(\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5)\ ;$$

[4] Si osservi infatti che g.l. (num.) $= p(q-1) - (p-1)(q-1) = q-1$, mentre g.l. (denom.) $= (p-1)(q-1)$.

se vale questa ipotesi si deve avere

$$F_0 = 3 \cdot \frac{21564,51 - 21543,05}{21725,22 - 21564,51 - 21677,50 + 21543,05} = 2,45 \ ,$$

estrazione da una F a (4,12) gradi di libertà: il valore critico è $F_{0,05} = 3,26$ e quindi H_0 va accettata, cioè non si ha una significativa dipendenza dai terreni (fattore A).

Per valutare la dipendenza da B poniamo

$$H_0(\beta_1 = \beta_2 = \beta_3 = \beta_4) \ :$$

se è giusta H_0 si dovrà avere

$$F_0 = 4 \cdot \frac{21677,50 - 21543,05}{21725,22 - 21564,51 - 21677,50 + 21543,05} = 20,48 \ ,$$

estrazione da una F a (3,12) gradi di libertà: il corrispondente valore critico $F_{0,05}$ essendo 3,49 , si vede che H_0 va rifiutata, cioè si ha una dipendenza significativa dalle varietà di grano impiegate.