la distribuzione di Student ad N gradi di libertà è a sua volta distribuito con una densità di probabilità data da $F(t^2; 1, N)$.

Per terminare, quando i due parametri M ed N (da cui la funzione di frequenza di Fisher (12.19) dipende) vengono resi arbitrariamente grandi, essa tende ad una distribuzione normale; ma la convergenza è lenta, e l'approssimazione normale alla distribuzione di Fisher si può pensare in pratica usabile quando sia M che N sono superiori a 50.

12.6.1 Confronto tra varianze

Supponiamo di avere a disposizione due campioni di misure, che ipotizziamo provenire da due differenti popolazioni che seguano delle distribuzioni normali.

Siano M ed N le dimensioni di tali campioni, e siano ${\sigma_1}^2$ e ${\sigma_2}^2$ le varianze delle rispettive popolazioni di provenienza; indichiamo poi con ${s_1}^2$ ed ${s_2}^2$ le due stime delle varianze delle popolazioni ricavate dai campioni. Vogliamo ora capire come si può verificare l'ipotesi statistica che le due popolazioni abbiano *la stessa varianza*, ossia che $\sigma_1 = \sigma_2$.

Ora sappiamo già dalla equazione (12.8) che le due variabili casuali

$$X = (M-1)\,\frac{{s_1}^2}{{\sigma_1}^2} \qquad \text{e} \qquad Y = (N-1)\,\frac{{s_2}^2}{{\sigma_2}^2}$$

sono entrambe distribuite come il χ^2, con $M-1$ ed $N-1$ gradi di libertà rispettivamente; quindi la quantità

$$w \;=\; \frac{X}{M-1}\,\frac{N-1}{Y} \;=\; \frac{{s_1}^2}{{\sigma_1}^2}\,\frac{{\sigma_2}^2}{{s_2}^2}$$

ha densità di probabilità data dalla funzione di Fisher con $M-1$ ed $N-1$ gradi di libertà.

Assunta a priori vera l'ipotesi statistica $\sigma_1 = \sigma_2$, la variabile casuale

$$w = \frac{{s_1}^2}{{s_2}^2}$$

ha densità di probabilità data dalla funzione di Fisher prima menzionata, $F(w; M-1, N-1)$; per cui, fissato un livello di confidenza al di là del quale rigettare l'ipotesi, e ricavato dalle apposite tabelle[7] il valore W che lascia alla propria sinistra, al di sotto della funzione $F(w; M-1, N-1)$, un'area pari al livello di confidenza prescelto, si può escludere che i due campioni provengano da popolazioni con la stessa varianza se $w > W$.

[7] Per un livello di confidenza pari a 0.95 o 0.99, e per alcuni valori dei due parametri M ed N, ci si può riferire ancora alle tabelle dell'appendice G; in esse si assume che sia $s_1 > s_2$, e quindi $w > 1$.

12.7 Il metodo di Kolmogorov e Smirnov

Il *test di Kolmogorov e Smirnov* è un metodo di analisi statistica che permette di confrontare tra loro un campione di dati ed una distribuzione teorica (oppure due campioni di dati) allo scopo di verificare l'ipotesi statistica che la popolazione da cui i dati provengono sia quella in esame (oppure l'ipotesi che entrambi i campioni provengano dalla stessa popolazione).

Una caratteristica interessante di questo metodo è che esso non richiede la preventiva, e più o meno arbitraria, suddivisione dei dati in classi di frequenza; definendo queste ultime in modo diverso si ottengono ovviamente, dal metodo del χ^2, differenti risultati per gli stessi campioni.

Il test di Kolmogorov e Smirnov si basa infatti sulla *frequenza cumulativa relativa* dei dati, introdotta nel paragrafo 4.1 a pagina 33; e sull'analogo concetto di *funzione di distribuzione* di una variabile continua definito nel paragrafo 6.1 a pagina 68. Per la compatibilità tra un campione ed una ipotetica legge che si ritiene possa descriverne la popolazione di provenienza, e collegata ad una funzione di distribuzione $\Phi(x)$, bisogna confrontare la frequenza cumulativa relativa $F(x)$ del campione con $\Phi(x)$ per ricavare *il valore assoluto del massimo scarto tra esse*,

$$\delta = \max\Big\{\,|\,F(x) - \Phi(x)\,|\,\Big\}\ .$$

Si può dimostrare che, se l'ipotesi da verificare fosse vera, la probabilità di ottenere casualmente un valore di δ non inferiore ad una prefissata quantità (positiva) δ_0 sarebbe data da

$$\Pr\left(\delta \geq \delta_0\right) = F_{\mathrm{KS}}\left(\delta_0'\right)$$

ove F_{KS} è la serie

$$F_{\mathrm{KS}}(x) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2\,k^2x^2} \tag{12.20}$$

e δ_0' vale

$$\delta_0' = \left(\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}}\right)\delta_0\ . \tag{12.21}$$

La legge ora enunciata è approssimata, ma il test di Kolmogorov e Smirnov può essere usato già per dimensioni del campione N uguali a 5. Attenzione però che, se qualche parametro da cui la distribuzione teorica dipende è stato stimato sulla base dei dati, l'integrale della densità di probabilità

per la variabile δ di Kolmogorov e Smirnov *non segue più* la legge (12.20): non solo, ma non è più possibile ricavare teoricamente una funzione che ne descriva il comportamento in generale (in questi casi, nella pratica, la distribuzione di δ viene studiata usando metodi di Montecarlo).

Se si vogliono invece confrontare tra loro due campioni indipendenti per verificarne la compatibilità, bisogna ricavare dai dati il massimo scarto (in valore assoluto), δ, tra le due frequenze cumulative relative; e ricavare ancora dalla (12.20) la probabilità che questo possa essere avvenuto (ammessa vera l'ipotesi) per motivi puramente casuali. L'unica differenza è che la funzione (12.20) va calcolata in un'ascissa δ'_0 data dalla (12.21), nella quale N vale

$$N = \frac{1}{\frac{1}{N_1} + \frac{1}{N_2}} = \frac{N_1 N_2}{N_1 + N_2}$$

(N_1 ed N_2 sono le dimensioni dei due campioni).

Oltre al già citato vantaggio di non richiedere la creazione di più o meno arbitrarie classi di frequenza per raggrupparvi i dati, un'altra caratteristica utile del test di Kolmogorov e Smirnov è quella di essere, entro certi limiti, *indipendente dalla variabile usata* nella misura: se al posto di x si usasse, per caratterizzare il campione, $\ln(x)$ o $\sqrt{x}$, il massimo scarto tra frequenza cumulativa e funzione di distribuzione rimarrebbe invariato.

Un altrettanto ovvio svantaggio è collegato al fatto che per valori molto piccoli (o molto grandi) della variabile casuale usata, qualsiasi essa sia, *tutte* le funzioni di distribuzione e tutte le frequenze cumulative *hanno lo stesso valore* (0, o 1 rispettivamente). Per questo motivo il test di Kolmogorov e Smirnov è assai sensibile a differenze nella zona centrale dei dati (attorno al valore medio), mentre non è affatto efficace per discriminare tra due distribuzioni che differiscano significativamente tra loro solo nelle code; ad esempio che abbiano lo stesso valore medio e differente ampiezza.

Capitolo 13

La verifica delle ipotesi (II)

Nel precedente capitolo 12 abbiamo esaminato varie tecniche che ci permettono di decidere se una caratteristica del processo fisico che ha prodotto un campione di dati è o non è confermata dai dati stessi; tutte queste tecniche non sono che casi particolari di una teoria generale, di cui ora ci occuperemo, senza però scendere in profondità nei dettagli.

In sostanza, nei vari casi del capitolo 12, abbiamo formulato una certa ipotesi H_0 sulla natura di un fenomeno casuale; e, ammesso per assurdo che questa ipotesi fosse vera, abbiamo associato un ben definito valore della densità di probabilità ad ogni punto E dello spazio S degli eventi.

Se indichiamo con K un valore (arbitrariamente scelto) della probabilità, *livello di confidenza* nel linguaggio statistico, abbiamo in sostanza diviso S in due sottoinsiemi esclusivi ed esaurienti: uno $\mathcal{R}$ di eventi con probabilità complessiva $1 - K$, ed uno $\mathcal{A} = S - \mathcal{R}$ di eventi con probabilità complessiva K.

Per verificare l'ipotesi H_0 occorre scegliere a priori un valore di K da assumere come il confine che separi, da una parte, eventi che riteniamo ragionevole si possano presentare nell'ambito di pure fluttuazioni casuali se è vera H_0; e, dall'altra, eventi così improbabili (sempre ammesso che H_0 sia vera) da far sì che la loro effettiva realizzazione debba implicare la falsità dell'ipotesi.

Normalmente si sceglie $K = 0.95$ o $K = 0.997$, i valori della probabilità che corrispondono a scarti di due o tre errori quadratici medi per la distribuzione di Gauss, anche se altri valori (come ad esempio $K = 0.999$ o $K = 0.99$) sono abbastanza comuni; e, una volta fatto questo, si *rigetta l'ipotesi* H_0 se il

dato a disposizione (un evento E ottenuto dall'effettivo studio del fenomeno in esame) appartiene ad $\mathcal{R}$; e la si accetta se appartiene ad $\mathcal{A}$.

In realtà nella pratica si presenta in generale la necessità di discriminare tra *due* ipotesi, sempre mutuamente esclusive, che indicheremo con i simboli H_0 ed H_a e che, usando la terminologia della statistica, si chiamano rispettivamente *ipotesi nulla* ed *ipotesi alternativa*; i casi precedenti corrispondono al caso particolare in cui l'ipotesi alternativa coincida con il *non realizzarsi* di H_0.

Ipotesi nulla ed ipotesi alternativa possono essere entrambe eventi semplici, oppure composti (ossia somma logica di più eventualità semplici); e lo scopo di questo capitolo è quello di mostrare dei criteri sulla base dei quali si possa opportunamente definire nello spazio degli eventi una *regione di rigetto* $\mathcal{R}$ per l'ipotesi nulla (e, in corrispondenza, ovviamente, una regione $\mathcal{A} = S - \mathcal{R}$ nella quale tale ipotesi viene accettata).

È chiaro che si corre sempre il rischio di sbagliare: o rigettando erroneamente ipotesi in realtà vere (*errori di prima specie*) o accettando invece ipotesi in realtà false (*errori di seconda specie*); e che, allargando o restringendo la regione di rigetto, si può diminuire la probabilità di uno di questi due tipi di errori solo per aumentare la probabilità di quelli dell'altra categoria. Se indichiamo con P_I e P_{II} le probabilità degli errori di prima e seconda specie rispettivamente, sulla base della definizione risulta

$$P_I = \Pr(E \in \mathcal{R}|H_0) \qquad \text{e} \qquad P_{II} = \Pr(E \in \mathcal{A}|H_a) \ .$$

Quello che abbiamo finora chiamato "livello di confidenza" non è altro che $1 - P_I$; P_I viene anche indicato col simbolo α e chiamato *significanza* del criterio adottato. Infine, la probabilità di *non* commettere un errore di seconda specie, ovvero la probabilità di rigettare H_0 quando l'ipotesi nulla è falsa (e quindi quella alternativa è vera) si indica col simbolo β e si chiama *potenza* del criterio adottato; essa vale quindi

$$\beta = \Pr(E \in \mathcal{R}|H_a) = 1 - P_{II} \ .$$

Per fare un esempio concreto, il fisico si trova spesso ad esaminare "eventi" sperimentali e deve decidere se essi sono del tipo desiderato (segnale) o no (fondo): in questo caso l'ipotesi nulla H_0 consiste nell'appartenenza di un evento al segnale, mentre l'ipotesi alternativa H_a corrisponde invece all'appartenenza dello stesso evento al fondo; che in genere non è l'intero insieme di eventi complementare all'ipotesi nulla, $\overline{H}_0$, ma si sa restringere ad una classe ben definita di fenomeni.

Gli errori di prima specie consistono in questo caso nello scartare eventi buoni (errori di *impoverimento* del segnale), e quelli di seconda specie nell'introduzione nel segnale di eventi di fondo (errori di *contaminazione*).

I criteri da seguire per definire una regione $\mathcal{R}$ nella quale rigettare H_0 sono dettati dalle caratteristiche del processo di generazione: se gli eventi di fondo sono preponderanti rispetto al segnale, ad esempio, bisognerà evitare gli errori di seconda specie per quanto possibile; anche al prezzo di scartare in questo modo una parte consistente del segnale.

Estendendo al caso generale il metodo seguito nei vari casi del capitolo 12 e prima delineato, se si è in grado di associare ad ogni punto dello spazio degli eventi *due* valori della probabilità (o della densità di probabilità nel caso di variabili continue), sia ammessa vera l'ipotesi nulla che ammessa invece vera l'ipotesi alternativa, si può pensare di usare *il loro rapporto* per definire la regione di rigetto.

Limitandoci al caso delle variabili continue, insomma, dopo aver definito una nuova variabile casuale λ attraverso la

$$\lambda = \frac{\mathcal{L}(\boldsymbol{x}|H_0)}{\mathcal{L}(\boldsymbol{x}|H_a)} \ ,$$

possiamo scegliere arbitrariamente un numero reale k e decidere di accettare l'ipotesi H_0 se $\lambda \geq k$ o di rifiutarla se $\lambda < k$; in definitiva ad ogni k ammissibile è associata una differente regione di rigetto $\mathcal{R}_k$ definita da

$$\mathcal{R}_k \equiv \left\{\lambda = \frac{\mathcal{L}(\boldsymbol{x}|H_0)}{\mathcal{L}(\boldsymbol{x}|H_a)} < k\right\} \ .$$

$\mathcal{L}$, nelle espressioni precedenti, è la funzione di verosimiglianza; che rappresenta appunto la densità di probabilità corrispondente all'ottenere (sotto una certa ipotesi) un campione di N valori $x_1, x_2, \ldots, x_N$ (qui indicato sinteticamente come un vettore $\boldsymbol{x}$ a N componenti). Ma in base a quale criterio dobbiamo scegliere k?

13.1 Un primo esempio

Cominciamo con un esempio didattico: supponiamo che i valori x_i si sappiano provenienti da una popolazione normale $N(x;\mu,\sigma)$ di varianza σ^2 nota: e che il nostro scopo consista nel discriminare tra due possibili valori μ_1 e μ_2 per μ; valori che, senza perdere in generalità, supponiamo siano 0 e 1 (potendosi sempre effettuare un opportuno cambiamento di variabile casuale che ci porti in questa situazione). Riassumendo: siano

$$\begin{cases} x \sim N(x;\mu,\sigma) & (\text{con } \sigma > 0 \text{ noto}) \\ H_0 \equiv \{\mu = 0\} \\ H_a \equiv \{\mu = 1\} \end{cases}$$

le nostre ipotesi.

La densità di probabilità della x vale

$$N(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

e, quindi, la funzione di verosimiglianza ed il suo logaritmo valgono

$$\mathcal{L}(\boldsymbol{x};\mu,\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^N}\prod_{i=1}^{N} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

e, rispettivamente,

$$\begin{aligned}\ln\mathcal{L}(\boldsymbol{x};\mu,\sigma) &= -N\,\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 \\ &= -N\,\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left({x_i}^2 - 2\mu x_i + \mu^2\right)\ ;\end{aligned}$$

per cui

$$\ln\mathcal{L}(\boldsymbol{x};\mu,\sigma) = -N\,\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}{x_i}^2 - 2N\bar{x}\mu + N\mu^2\right)\ . \qquad (13.1)$$

Dalla (13.1) si ricava immediatamente

$$\ln\lambda \;=\; \ln\mathcal{L}(\boldsymbol{x};0,\sigma) - \ln\mathcal{L}(\boldsymbol{x};1,\sigma) \;=\; \frac{N}{2\sigma^2}\,(1-2\bar{x})$$

e la regione di rigetto $\mathcal{R}_k$ è definita dalla

$$\mathcal{R}_k \;\equiv\; \left\{\, \ln\lambda = \frac{N}{2\sigma^2}\,(1-2\bar{x}) < \ln k \,\right\}$$

da cui consegue, con facili passaggi,

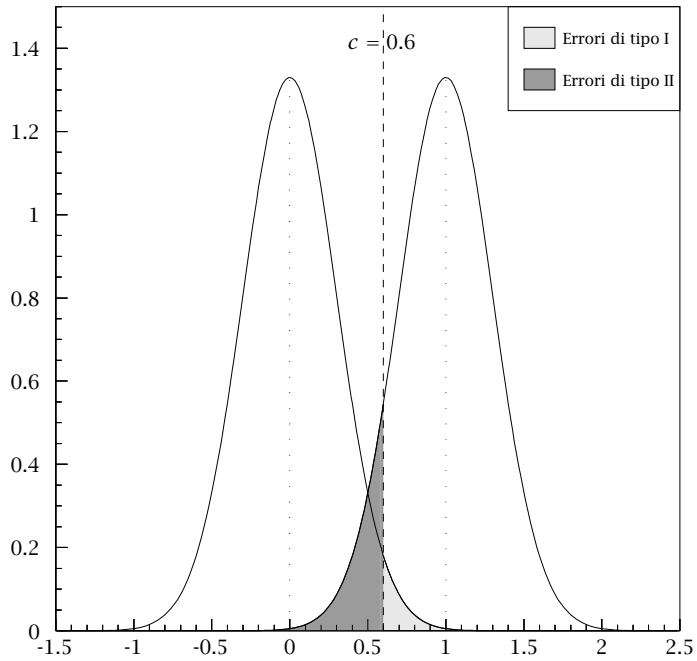
$$\mathcal{R}_k \;\equiv\; \left\{\, \bar{x} > \frac{N - 2\sigma^2\ln k}{2N} = c \,\right\}\ ;$$

ed insomma H_0 va rigettata se la media aritmetica del campione $\bar{x}$ risulta superiore a c; ed accettata altrimenti.

Come si può scegliere un valore opportuno di k (e quindi di c)? Gli errori di prima specie (si faccia riferimento anche alla figura 13a) hanno probabilità

$$P_I \;=\; 1-\alpha \;=\; \Pr(\bar{x} > c\,|H_0) \;=\; \int_c^{+\infty} N\left(x;0,\frac{\sigma}{\sqrt{N}}\right)\mathrm{d}x \qquad (13.2)$$

FIGURA 13a - L'esempio del paragrafo 13.1, con delineate (in corrispondenza ad un particolare valore di c) le probabilità degli errori di prima e seconda specie; le due curve sono $N(0,\sigma/\sqrt{N})$ e $N(1,\sigma/\sqrt{N})$.



e quelli di seconda specie

$$P_{II} \;=\; 1-\beta \;=\; \Pr(\bar{x} < c|H_a) \;=\; \int_{-\infty}^{c} N\left(x;1,\frac{\sigma}{\sqrt{N}}\right)\mathrm{d}x \qquad (13.3)$$

per cui si hanno svariate possibilità: ad esempio, se interessa contenere gli errori di prima specie e la dimensione del campione è nota, si fissa un valore opportunamente grande per α e dalla (13.2) si ricava c; o, se interessa contenere gli errori di seconda specie e la dimensione del campione è nota, si fissa β e si ricava c dalla (13.3); o, infine, se si vogliono contenere gli errori di entrambi i tipi, si fissano sia α che β e si ricava la dimensione minima del campione necessaria per raggiungere lo scopo utilizzando entrambe le equazioni (13.2) e (13.3).

13.2 Il lemma di Neyman–Pearson

L'essere ricorsi per la definizione della regione di rigetto $\mathcal{R}$ al calcolo del *rapporto* delle funzioni di verosimiglianza non è stato casuale; esiste infatti un teorema (il cosiddetto *lemma di Neyman–Pearson*) il quale afferma che

> *Se si ha a disposizione un campione di N valori indipendenti x_i da utilizzare per discriminare tra un'ipotesi nulla ed un'ipotesi alternativa entrambe semplici, e se è richiesto un livello fisso α di significanza, la massima potenza del test (ovvero la minima probabilità di errori di seconda specie) si raggiunge definendo la regione di rigetto $\mathcal{R}_k$ attraverso una relazione del tipo*

$$\mathcal{R}_k \;\equiv\; \left\{\, \lambda = \frac{\mathcal{L}(\boldsymbol{x}|H_0)}{\mathcal{L}(\boldsymbol{x}|H_a)} < k \,\right\} \quad . \qquad (13.4)$$

Infatti, indicando con $f = f(x;\theta)$ la densità di probabilità della variabile x (che supponiamo dipenda da un solo parametro θ), siano $H_0 \equiv \{\theta = \theta_0\}$ e $H_a \equiv \{\theta = \theta_a\}$ le due ipotesi (semplici) tra cui decidere; la funzione di verosimiglianza vale, come sappiamo,

$$\mathcal{L}(\boldsymbol{x};\theta) = \prod_{i=1}^{N} f(x_i;\theta) \quad .$$

Indichiamo con $\mathfrak{R}$ l'insieme di tutte le regioni $\mathcal{R}$ per le quali risulti

$$P_I \;=\; \int_{\mathcal{R}} \mathcal{L}(\boldsymbol{x};\theta_0)\,\mathrm{d}\boldsymbol{x} \;=\; 1-\alpha \qquad (13.5)$$

con α costante prefissata (nella (13.5) abbiamo sinteticamente indicato con $\mathrm{d}\boldsymbol{x}$ il prodotto $\mathrm{d}x_1\,\mathrm{d}x_2\cdots\mathrm{d}x_N$). Vogliamo trovare quale di queste regioni rende massima

$$\beta \;=\; 1-P_{II} \;=\; \int_{\mathcal{R}} \mathcal{L}(\boldsymbol{x};\theta_a)\,\mathrm{d}\boldsymbol{x} \quad .$$

Ora, per una qualsiasi regione $\mathcal{R} \neq \mathcal{R}_k$, valgono sia la

$$\mathcal{R}_k \;=\; (\mathcal{R}_k \cap \mathcal{R}) \cup (\mathcal{R}_k \cap \overline{\mathcal{R}})$$

che la

$$\mathcal{R} \;=\; (\mathcal{R} \cap \mathcal{R}_k) \cup (\mathcal{R} \cap \overline{\mathcal{R}}_k) \;;$$

e quindi, per una qualsiasi funzione $\phi(\boldsymbol{x})$, risulta sia

$$\int_{\mathcal{R}_k} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \;=\; \int_{\mathcal{R}_k\cap\mathcal{R}} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_k\cap\overline{\mathcal{R}}} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

che

$$\int_{\mathcal{R}} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \;=\; \int_{\mathcal{R}\cap\mathcal{R}_k} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}\cap\overline{\mathcal{R}}_k} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

e, sottraendo membro a membro,

$$\int_{\mathcal{R}_k} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \;=\; \int_{\mathcal{R}_k\cap\overline{\mathcal{R}}} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}\cap\overline{\mathcal{R}}_k} \phi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad . \qquad (13.6)$$

Applicando la (13.6) alla funzione $\mathcal{L}(\boldsymbol{x}|\theta_a)$ otteniamo:

$$\begin{aligned}\int_{\mathcal{R}_k} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} \;&= \\ &= \int_{\mathcal{R}_k\cap\overline{\mathcal{R}}} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}\cap\overline{\mathcal{R}}_k} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} \;; \qquad (13.7)\end{aligned}$$

ma, nel primo integrale del secondo membro, essendo la regione di integrazione contenuta in $\mathcal{R}_k$, deve valere la (13.4); e quindi risultare ovunque

$$\mathcal{L}(\boldsymbol{x}|\theta_a) \;>\; \frac{1}{k}\cdot\mathcal{L}(\boldsymbol{x}|\theta_0)$$

mentre, per lo stesso motivo, nel secondo integrale

$$\mathcal{L}(\boldsymbol{x}|\theta_a) \;\leq\; \frac{1}{k}\cdot\mathcal{L}(\boldsymbol{x}|\theta_0)$$

e quindi la (13.7) implica che

$$\int_{\mathcal{R}_k} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} > \\ > \frac{1}{k}\cdot\left[\int_{\mathcal{R}_k\cap\overline{\mathcal{R}}} \mathcal{L}(\boldsymbol{x}|\theta_0)\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}\cap\overline{\mathcal{R}}_k} \mathcal{L}(\boldsymbol{x}|\theta_0)\,\mathrm{d}\boldsymbol{x}\right] .$$

Ricordando la (13.6),

$$\int_{\mathcal{R}_k} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} > \frac{1}{k}\cdot\left[\int_{\mathcal{R}_k} \mathcal{L}(\boldsymbol{x}|\theta_0)\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}} \mathcal{L}(\boldsymbol{x}|\theta_0)\,\mathrm{d}\boldsymbol{x}\right]$$

e, se $\mathcal{R} \in \mathfrak{R}$ e quindi soddisfa anch'essa alla (13.5),

$$\int_{\mathcal{R}_k} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{R}} \mathcal{L}(\boldsymbol{x}|\theta_a)\,\mathrm{d}\boldsymbol{x} > \frac{1}{k}\cdot[P_I - P_I] = 0$$

che era la nostra tesi.

13.3 Tests di massima potenza uniforme

Consideriamo ora un esempio del tipo di quello del paragrafo 13.1; e sia sempre disponibile un campione di N misure indipendenti derivante da una popolazione normale di varianza nota. Assumiamo ancora come ipotesi nulla quella che la popolazione abbia un certo valore medio, che supponiamo essere 0, ma sostituiamo alla vecchia ipotesi alternativa H_a una nuova ipotesi *composta*; ovvero quella che il valore medio della popolazione sia positivo:

$$\begin{cases} x \sim N(x;\mu,\sigma) & (\text{con } \sigma > 0 \text{ noto}) \\ H_0 \equiv \{\mu = 0\} \\ H_a \equiv \{\mu > 0\} \end{cases}$$

(l'ipotesi alternativa è dunque somma logica di infinite ipotesi semplici del tipo $\mu = \mu_a$ con $\mu_a > 0$).

Dalla (13.1) ricaviamo immediatamente le

$$\mathcal{L}(\boldsymbol{x};0,\sigma) = -N\,\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{N} {x_i}^2$$

e

$$\mathcal{L}(\boldsymbol{x};\mu_a,\sigma) = -N\,\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{N} {x_i}^2 - 2N\bar{x}\mu_a + N{\mu_a}^2\right)$$

(sempre con $\mu_a > 0$); e, sostituendole nella (13.4), che definisce la generica regione di rigetto $\mathcal{R}_k$, otteniamo

$$\ln\lambda = \ln\mathcal{L}(\boldsymbol{x};0,\sigma) - \mathcal{L}(\boldsymbol{x};\mu_a,\sigma) = \frac{N\mu_a}{2\sigma^2}(\mu_a - 2\bar{x}) < \ln k$$

equivalente alla

$$\mathcal{R}_k \equiv \left\{ \bar{x} > \frac{N{\mu_a}^2 - 2\sigma^2\ln k}{2N\mu_a} = c \right\} .$$

Si rigetta quindi H_0 se la media aritmetica del campione è superiore a c e la si accetta altrimenti: la probabilità di commettere errori di prima specie vale

$$P_I = 1 - \alpha = \int_c^{+\infty} N\left(x;0,\frac{\sigma}{\sqrt{N}}\right)\mathrm{d}x$$

ed è ben definita; ma, al contrario, la probabilità di commettere errori di seconda specie dipende dal particolare valore di μ_a, e non può quindi essere calcolata.

Se interessa solo contenere gli errori di prima specie e la dimensione del campione è nota, si fissa α e si ricava il corrispondente valore di c dall'equazione precedente; altrimenti occorre fare delle ulteriori ipotesi sulla funzione di frequenza dei differenti valori di μ_a, e, ad esempio, calcolare la probabilità degli errori di seconda specie con tecniche di Montecarlo.

In ogni caso, però, osserviamo che la regione di rigetto è sempre dello stesso tipo (13.4) per *qualsiasi* $\mu_a > 0$; e quindi un confronto separato tra H_0 ed ognuna delle differenti ipotesi semplici che costituiscono H_a è comunque del tipo per cui il lemma di Neyman–Pearson garantisce la massima potenza.

Tests di questo tipo, per i quali *la significanza è costante e la potenza è massima per ognuno dei casi semplici che costituiscono l'ipotesi alternativa*, si dicono "tests di massima potenza uniforme".

13.4 Il rapporto delle massime verosimiglianze

Nel caso generale in cui sia l'ipotesi nulla che quella alternativa siano composte, la situazione è più complicata: non esiste normalmente un test di massima potenza uniforme, e, tra i vari criteri possibili per decidere tra le due ipotesi, bisogna capire quali abbiano caratteristiche (significanza e potenza) adeguate; un metodo adatto a costruire una regione di rigetto dotata asintoticamente (per grandi campioni) di caratteristiche, appunto, desiderabili, è quello seguente (*metodo del rapporto delle massime verosimiglianze*).

Sia una variabile casuale x, la cui densità di probabilità supponiamo sia una funzione $f(x;\theta_1,\theta_2,\ldots,\theta_M)$ dipendente da M parametri: indicando sinteticamente la M-pla dei valori dei parametri come un vettore $\boldsymbol{\theta}$ in uno spazio a M dimensioni (*spazio dei parametri*), consista H_0 nell'essere $\boldsymbol{\theta}$ compreso all'interno di una certa regione Ω_0 di tale spazio; mentre H_a consista nell'appartenere $\boldsymbol{\theta}$ alla regione Ω_a complementare a Ω_0: $\Omega_a \equiv \overline{H}_0$, così che $(\Omega_0 \cup \Omega_a)$ coincida con l'intero spazio dei parametri S.

In particolare, Ω_0 può estendersi, in alcune delle dimensioni dello spazio dei parametri, da $-\infty$ a $+\infty$; e, in tal caso, il vincolo sulle θ_i cui corrisponde l'ipotesi nulla riguarderà un numero di parametri minore di M.

Scritta la funzione di verosimiglianza,

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i;\boldsymbol{\theta}) \tag{13.8}$$

indichiamo con $\mathcal{L}(\hat{S})$ il suo massimo valore nell'intero spazio dei parametri; e con $\mathcal{L}(\hat{R})$ il massimo valore assunto sempre della (13.8), ma con i parametri vincolati a trovarsi nella regione Ω_0 (quindi limitatamente a quei casi nei quali H_0 è vera). Il rapporto

$$\lambda = \frac{\mathcal{L}(\hat{R})}{\mathcal{L}(\hat{S})} \tag{13.9}$$

deve essere un numero appartenente all'intervallo $[0,1]$; se si fissa un arbitrario valore k $(0 < k < 1)$, esso definisce una generica regione di rigetto, $\mathcal{R}_k$, attraverso la

$$\mathcal{R}_k \equiv \left\{ \lambda = \frac{\mathcal{L}(\hat{R})}{\mathcal{L}(\hat{S})} < k \right\}$$

(ovvero si accetta H_0 quando $\lambda \geq k$ e la si rigetta quando $\lambda < k$). Nel caso si sappia determinare la densità di probabilità di λ condizionata all'assunzione che H_0 sia vera, $g(\lambda|H_0)$, la probabilità di un errore di prima specie è data ovviamente da

$$P_I = \alpha = \Pr(\lambda \in [0,k]|H_0) = \int_0^k g(\lambda|H_0)\,\mathrm{d}\lambda \ .$$

L'importanza del metodo sta nel fatto che si può dimostrare il seguente

TEOREMA: *se l'ipotesi nulla H_0 consiste nell'appartenenza di un insieme di $P \leq M$ dei parametri θ_i ad una determinata regione Ω_0, e se l'ipotesi alternativa H_a consiste nel fatto che essi non vi appartengano ($H_a \equiv \overline{H}_0$), allora $-2\ln\lambda$, ove λ è definito dalla* (13.9)*, ha densità di probabilità che, ammessa vera l'ipotesi nulla, converge in probabilità (all'aumentare di N) alla distribuzione del χ^2 a P gradi di libertà.*

che, dicendoci quale è (almeno nel limite di grandi campioni) la forma di $g(\lambda|H_0)$, ci mette comunque in grado di calcolare la significanza del test.

Illustriamo il metodo con un esempio: disponendo ancora di un campione di N determinazioni indipendenti, provenienti da una popolazione normale di varianza nota, vogliamo applicarlo per discriminare tra l'ipotesi nulla che il valore medio abbia valore 0 ($H_0 \equiv \{\mu = 0\}$) e quella che esso abbia valore differente ($H_a \equiv \{\mu \neq 0\}$).

Il logaritmo della funzione di verosimiglianza è ancora dato dalla (13.1); e già sappiamo, dal paragrafo 11.3, che $\mathcal{L}$ assume il suo massimo valore quando $\mu = \bar{x}$, per cui

$$\ln\mathcal{L}(\hat{S}) = -N\,\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{N} {x_i}^2 - N\bar{x}^2\right) \ .$$

Inoltre Ω_0 si riduce ad un unico punto, $\mu = 0$; per cui

$$\ln\mathcal{L}(\hat{R}) = -N\,\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{N} {x_i}^2 \ .$$

Dalla (13.9) si ricava

$$\ln\lambda = \ln\mathcal{L}(\hat{R}) - \ln\mathcal{L}(\hat{S}) = -\frac{1}{2\sigma^2}N\bar{x}^2$$

e la regione di rigetto è definita dalla $\ln\lambda < \ln k$; ovvero (ricordando che $\ln k < 0$) da

$$\mathcal{R}_k \equiv \left\{ \bar{x}^2 > -\frac{2\sigma^2\ln k}{N} \right\}$$

e, posto

$$c = \sigma\sqrt{-\frac{2\ln k}{N}}$$

si accetterà H_0 se $|\bar{x}| \leq c$ (e la si rigetterà se $|\bar{x}| > c$).

In questo caso il teorema precedentemente citato afferma che

$$-2\ln\lambda = \frac{\bar{x}^2}{\frac{\sigma^2}{N}}$$

è distribuito asintoticamente come il χ^2 ad un grado di libertà (cosa che del resto già sapevamo, vista l'espressione di $-2\ln\lambda$); per cui, indicando

con $F(t;N)$ la densità di probabilità della distribuzione del χ^2 a N gradi di libertà, avremo

$$P_I \;=\; \alpha \;=\; \int_0^k g(\lambda|H_0)\,\mathrm{d}\lambda \;=\; \int_{-2\ln k}^{+\infty} F(t;1)\,\mathrm{d}t$$

della quale ci possiamo servire per ricavare k se vogliamo che la significanza del test abbia un certo valore: ad esempio un livello di confidenza del 95% corrisponde ad $\alpha = 0.05$ e, dalle tabelle della distribuzione del χ^2, ricaviamo

$$-2\ln k = 3.84 \qquad \text{e quindi} \qquad c = 1.96\,\frac{\sigma}{\sqrt{N}}\ .$$

Anche senza dover ricorrere al teorema sul comportamento asintotico di $-2\ln\lambda$, allo stesso risultato si può pervenire per altra via: in questo caso si conosce infatti esattamente α, che vale

$$P_I \;=\; \alpha \;=\; \Pr\bigl(|\bar{x}| > c\,|\,H_0\bigr) \;=\; 2\int_c^{+\infty} N\left(t;0,\frac{\sigma}{\sqrt{N}}\right)\mathrm{d}t$$

e, dalle tabelle della distribuzione normale standardizzata, si ricava che un'area two-tailed del 5% corrisponde ad un valore assoluto dello scarto normalizzato $t_0 = 1.96$; per cui, ancora, si ricaverebbe $|\bar{x}| > 1.96(\sigma/\sqrt{N})$ come test per un livello di confidenza del 95%.

13.5 Applicazione: ipotesi sulle probabilità

Nel paragrafo 11.5 abbiamo preso in considerazione il caso di un evento casuale che si può manifestare in un numero finito M di modalità, aventi ognuna probabilità incognita p_i; la stima di massima verosimiglianza delle p_i è data dal rapporto tra la frequenza assoluta di ogni modalità, n_i, ed il numero totale di prove, N.

Vogliamo ora applicare il metodo del rapporto delle massime verosimiglianze per discriminare, sulla base di un campione di determinazioni indipendenti, l'ipotesi nulla che le probabilità abbiano valori noti a priori e l'ipotesi alternativa complementare, $H_a \equiv \overline{H}_0$:

$$\begin{cases} H_0 \;\equiv\; \{p_i = \pi_i\} & (\forall\, i \in \{1,2,\ldots,M\}) \\ H_a \;\equiv\; \{p_i \neq \pi_i\} & (\exists i \in \{1,2,\ldots,M\}) \end{cases}$$

Ricordiamo che la funzione di verosimiglianza, a meno di un fattore moltiplicativo costante, è data da

$$\mathcal{L}(\boldsymbol{n};\boldsymbol{p}) = \prod_{i=1}^{M} {p_i}^{n_i}$$

e che, essendo la stima di massima verosimiglianza data da

$$\hat{p}_i = \frac{n_i}{N}$$

il massimo assoluto di $\mathcal{L}$ è

$$\mathcal{L}(\hat{S}) \;=\; \prod_{i=1}^{M}\left(\frac{n_i}{N}\right)^{n_i} \;=\; \frac{1}{N^N}\prod_{i=1}^{M} {n_i}^{n_i}\ .$$

Inoltre, nell'unico punto dello spazio dei parametri che corrisponde ad H_0,

$$\mathcal{L}(\hat{R}) = \prod_{i=1}^{M} {\pi_i}^{n_i}$$

per cui

$$\lambda \;=\; \frac{\mathcal{L}(\hat{R})}{\mathcal{L}(\hat{S})} \;=\; N^N \prod_{i=1}^{M}\left(\frac{\pi_i}{n_i}\right)^{n_i}$$

dalla quale si può, come sappiamo, derivare una generica regione di rigetto attraverso la consueta $\mathcal{R}_k \equiv \{\lambda < k\}$.

$$-2\,\ln\lambda \;=\; -2\left[N\,\ln N + \sum_{i=1}^{M} n_i\,(\ln\pi_i - \ln n_i)\right]$$

è inoltre asintoticamente distribuita come il χ^2 a $M-1$ gradi di libertà (c'è un vincolo: che le n_i abbiano somma N), e questo può servire a scegliere un k opportuno (nota la dimensione del campione) una volta fissata α.

Il criterio di verifica dell'ipotesi dato in precedenza consisteva nel calcolo del valore della variabile casuale

$$X = \sum_{i=1}^{M} \frac{(n_i - N\pi_i)^2}{N\pi_i}$$

e nel suo successivo confronto con la distribuzione del χ^2 a $M-1$ gradi di libertà; lo studio del rapporto delle massime verosimiglianze porta dunque ad un criterio *differente* e, senza sapere nulla della probabilità di commettere errori di seconda specie, non è possibile dire quale dei due risulti migliore (a parità di significanza).

13.6 Applicazione: valore medio di una popolazione normale

Ancora un esempio: sia una popolazione normale $N(x;\mu,\sigma)$ dalla quale vengano ottenuti N valori indipendenti x_i, ma questa volta *la varianza σ sia ignota*; vogliamo discriminare, sulla base del campione, tra l'ipotesi nulla che il valore medio della popolazione abbia un valore prefissato e l'ipotesi alternativa complementare,

$$\begin{cases} H_0 \equiv \{\mu = \mu_0\} \\ H_a \equiv \{\mu \neq \mu_0\} \end{cases}$$

Il logaritmo della funzione di verosimiglianza è

$$\ln \mathcal{L}(\boldsymbol{x};\mu,\sigma) = -N \ln\sigma - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 \qquad (13.10)$$

ed essendo le stime di massima verosimiglianza date, come avevamo trovato nel paragrafo 11.5, da

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{e} \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$$

ne deriva, sostituendo nella (13.10), che

$$\ln \mathcal{L}(\hat{S}) = -\frac{N}{2}\ln\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right] + \frac{N}{2}\ln N - \frac{N}{2}\ln(2\pi) - \frac{N}{2} \ .$$

D'altra parte, ammessa vera H_0, abbiamo che

$$\ln \mathcal{L}(\boldsymbol{x}|H_0) = -N\ln\sigma - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu_0)^2$$

e, derivando rispetto a σ,

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\ln \mathcal{L}(\boldsymbol{x}|H_0) = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu_0)^2 \ .$$

Annullando la derivata prima, si trova che l'unico estremante di $\mathcal{L}(\boldsymbol{x}|H_0)$ si ha per

$$\sigma_0 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_0)^2$$

mentre la derivata seconda vale

$$\frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\ln \mathcal{L}(\boldsymbol{x}|H_0) = \frac{N}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{N}(x_i-\mu_0)^2$$

e, calcolata per $\sigma = \sigma_0$,

$$\left.\frac{\mathrm{d}^2(\ln\mathcal{L})}{\mathrm{d}\sigma^2}\right|_{\sigma=\sigma_0} = -\frac{2N^2}{\sum_i (x_i-\mu_0)^2} < 0$$

per cui l'estremante è effettivamente un massimo. Sostituendo,

$$\ln \mathcal{L}(\hat{R}) = -\frac{N}{2}\ln\left[\sum_{i=1}^{N}(x_i-\mu_0)^2\right] + \frac{N}{2}\ln N - \frac{N}{2}\ln(2\pi) - \frac{N}{2}$$

$$\ln\lambda = \ln \mathcal{L}(\hat{R}) - \ln \mathcal{L}(\hat{S}) = -\frac{N}{2}\left\{\ln\left[\sum_{i=1}^{N}(x_i-\mu_0)^2\right] - \ln\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]\right\}$$

ed infine

$$\begin{aligned} \ln\lambda &= -\frac{N}{2}\ln\left[\frac{\sum_i (x_i-\mu_0)^2}{\sum_i (x_i-\bar{x})^2}\right] \\ &= -\frac{N}{2}\ln\left[1 + \frac{N(\bar{x}-\mu_0)^2}{\sum_i (x_i-\bar{x})^2}\right] \\ &= -\frac{N}{2}\ln\left(1 + \frac{t^2}{N-1}\right) \end{aligned}$$

tenendo conto dapprima del fatto che $\sum_i (x_i-\mu_0)^2 = \sum_i (x_i-\bar{x})^2 + N(\bar{x}-\mu_0)^2$, e definendo poi una nuova variabile casuale

$$t = (\bar{x}-\mu_0)\sqrt{\frac{N(N-1)}{\sum_i (x_i-\bar{x})^2}} = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{N}}} \ .$$

Un qualunque metodo per il rigetto di H_0 definito confrontando λ con un prefissato valore k si traduce, in sostanza, in un corrispondente confronto da eseguire per t:

$$\mathcal{R}_k \equiv \{\ln\lambda < \ln k\}$$

che porta alla

$$-\frac{N}{2}\ln\left(1+\frac{t^2}{N-1}\right) \;<\; \ln k$$

ed alla condizione

$$t^2 \;>\; (N-1)\left(k^{-\frac{2}{N}}-1\right) \;;$$

ovvero si rigetta l'ipotesi nulla se $|t|$ è maggiore di un certo t_0 (derivabile dall'equazione precedente), e la si accetta altrimenti.

Ma t (vedi anche l'equazione (12.17)) segue la distribuzione di Student a $N-1$ gradi di libertà, e quindi accettare o rigettare H_0 sotto queste ipotesi si riduce ad un test relativo a quella distribuzione: come già si era concluso nel capitolo 12. Il livello di significanza α è legato a t_0 dalla

$$\frac{\alpha}{2} = \int_{t_0}^{+\infty} F(t;N-1)\,\mathrm{d}t$$

(indicando con $F(t;N)$ la funzione di frequenza di Student a N gradi di libertà), tenendo conto che abbiamo a che fare con un two-tailed test ($\mathcal{R}_k \equiv \{|t| > t_0\}$).

Insomma non c'è differenza, in questo caso, tra quanto esposto nel capitolo precedente e la teoria generale discussa in quello presente: nel senso che i due criteri di verifica dell'ipotesi portano per questo problema allo stesso metodo di decisione (ma, come abbiamo visto nel paragrafo precedente, non è sempre così).

Appendice A

Cenni di calcolo combinatorio

Il *calcolo combinatorio* è una branca della matematica orientata alla discussione ed allo sviluppo di formule che permettano di ottenere il numero di casi distinti che si possono presentare in un esperimento, od il numero di elementi che compongono un insieme, senza ricorrere alla loro enumerazione esplicita.

Il calcolo combinatorio trova importanti applicazioni nella teoria della probabilità e nella statistica: alcune formule, specificatamente quelle per le permutazioni e le combinazioni, vengono usate nel corso del testo; qui se ne dà una breve giustificazione.

A.1 Il lemma fondamentale del calcolo combinatorio

LEMMA FONDAMENTALE DEL CALCOLO COMBINATORIO: *dati due insiemi I_1 ed I_2, composti da N_1 ed N_2 elementi distinti rispettivamente, l'insieme $I = I_1 \otimes I_2$ di tutte le coppie ordinate che si possono costruire associando un elemento di I_1 con un elemento di I_2 è composto da $N_1 \cdot N_2$ elementi.*

Questo lemma si può immediatamente generalizzare (per induzione completa) a K insiemi $I_1, \ldots, I_K$ composti da $N_1, \ldots, N_K$ elementi distinti rispettivamente: l'insieme $I = I_1 \otimes I_2 \otimes \cdots \otimes I_K$, costituito da tutte le possibili associazioni ordinate di K elementi ognuno dei quali provenga da un differente insieme I_j, con $j = 1, \ldots, K$, è composto da $N_1 \cdot N_2 \cdots N_K$ elementi.

A.2 Fattoriale di un numero intero

Si definisce come *fattoriale* di un numero intero positivo N, e si indica con il simbolo $N!$, il prodotto dei primi N numeri interi:

$$N! = 1 \cdot 2 \cdot 3 \cdots N \ ;$$

per motivi che appariranno chiari più avanti[1], si definisce poi il fattoriale di zero come $0! = 1$.

A.3 Disposizioni

Se N e K sono due numeri interi positivi tali che sia $K \leq N$, si definisce come numero delle disposizioni di N oggetti di classe K (che si indica con il simbolo D_K^N) il numero dei gruppi *distinti* di K oggetti che è possibile formare a partire dagli N originali; definendo come *distinti* due gruppi se essi differiscono o per qualche elemento o per l'ordine.

Come esempio, le disposizioni di classe 2 che si possono formare con le 21 lettere dell'alfabeto italiano sono le seguenti:

$$\begin{cases} & AB & AC & AD & \cdots & AV & AZ \\ BA & & BC & BD & \cdots & BV & BZ \\ & & & & \cdots & & \\ ZA & ZB & ZC & ZD & \cdots & ZV & \end{cases}$$

Il valore di D_K^N si può facilmente trovare sfruttando il lemma fondamentale del calcolo combinatorio: il primo elemento di una disposizione si può infatti scegliere in N modi distinti, il secondo in $N-1$, e così via. Di conseguenza D_K^N è il prodotto di K numeri interi decrescenti a partire da N:

$$D_K^N \ = \ N \cdot (N-1) \cdot (N-2) \cdots (N-K+1) \ = \ \frac{N!}{(N-K)!} \qquad \text{(A.1)}$$

(nel caso dell'esempio fatto, le disposizioni sono $D_2^{21} = 21 \cdot 20 = 420$; nella tabella in cui sono state elencate vi sono 21 righe di 20 elementi ciascuna).

L'espressione (A.1) è verificata anche se $K = N$, però purché (come prima detto) si ponga $0! = 1$.

[1] La "definizione" $0! = 1$ non è così arbitraria come può sembrare: in realtà si comincia definendo una certa funzione di variabile complessa $\Gamma(z)$ che, quando l'argomento z è un numero intero positivo, coincide con il suo fattoriale; e per la quale si vede che $\Gamma(0) = 1$.

A.4 Permutazioni

Se N è un numero intero positivo, si definisce come numero delle permutazioni di N oggetti, e si indica con P_N, il numero di maniere distinte in cui si possono ordinare gli N oggetti stessi. Evidentemente risulta

$$P_N \ \equiv \ D_N^N \ = \ N! \ .$$

A.5 Permutazioni con ripetizione

Se gli N oggetti che si hanno a disposizione sono tali da poter essere divisi in M gruppi (composti da $N_1, N_2, \ldots, N_M$ oggetti rispettivamente; ovviamente $N_1 + N_2 + \cdots + N_M = N$), tali che gli oggetti in ognuno di questi gruppi siano *indistinguibili* tra loro, il numero di permutazioni che con essi si possono realizzare è inferiore a P_N; più precisamente, visto che gli oggetti di ogni gruppo si possono scambiare tra loro in qualsiasi modo senza per questo dare luogo a una sequenza distinta, il numero di *permutazioni con ripetizione* è dato da

$$\frac{N!}{N_1! \cdot N_2! \cdots N_M!} \ . \qquad \text{(A.2)}$$

A.6 Combinazioni

Se N e K sono due numeri interi positivi tali che sia $K \leq N$, si definisce come numero delle combinazioni di classe K di N oggetti il numero dei sottoinsiemi *distinti* composti da K oggetti che è possibile formare a partire dagli N originali; definendo come *distinti* due sottoinsiemi se essi differiscono per qualche elemento. Il numero delle combinazioni di classe K di N oggetti si indica con uno dei due simboli

$$C_K^N \qquad \text{o} \qquad \binom{N}{K}$$

(l'ultimo dei quali si chiama *coefficiente binomiale*).

Consideriamo l'insieme composto da tutte le disposizioni di classe K di N oggetti, e pensiamo di raggruppare i suoi elementi in sottoinsiemi in modo

che ciascuno di essi contenga tutte e sole quelle disposizioni che differiscano esclusivamente per l'ordine ma siano composte dagli stessi oggetti; ovviamente il numero di questi sottoinsiemi è C_K^N: ed ognuno di essi contiene un numero di elementi che è P_K.

Da qui ricaviamo

$$C_K^N \equiv \binom{N}{K} = \frac{D_K^N}{P_K} = \frac{N \cdot (N-1) \cdots (N-K+1)}{K \cdot (K-1) \cdots 1} = \frac{N!}{K!\,(N-K)!} \qquad \text{(A.3)}$$

O, in altre parole, il numero di combinazioni di classe K di N oggetti è uguale al rapporto tra il prodotto di K numeri interi decrescenti a partire da N ed il prodotto di K numeri interi crescenti a partire dall'unità.

Si dimostrano poi facilmente, a partire dalla definizione, due importanti proprietà dei coefficienti binomiali:

$$\binom{N}{K} = \binom{N}{N-K}$$

e

$$\binom{N+1}{K} = \binom{N}{K-1} + \binom{N}{K} \ .$$

È da osservare che, così come sono stati ricavati (dalla definizione delle possibili combinazioni di N oggetti), i coefficienti binomiali hanno senso solo se N e K sono numeri interi; ed inoltre se risulta sia $N > 0$ che $0 \le K \le N$. La definizione (A.3) può comunque essere estesa a valori interi qualunque, ed anche a valori reali di N — ma questo esula dal nostro interesse.

A.7 Partizioni ordinate

Consideriamo un insieme di N oggetti; vogliamo calcolare il numero di maniere in cui essi possono essere divisi in M gruppi che siano composti da $N_1, N_2, \ldots, N_M$ oggetti rispettivamente (essendo $N_1 + N_2 + \cdots + N_M = N$).

Gli N_1 oggetti che compongono il primo gruppo possono essere scelti in $C_{N_1}^N$ modi differenti; quelli del secondo gruppo in $C_{N_2}^{N-N_1}$ modi; e così via. Per il lemma fondamentale del calcolo combinatorio, il numero delle *partizioni*

ordinate deve essere uguale a

$$\binom{N}{N_1}\binom{N-N_1}{N_2}\binom{N-N_1-N_2}{N_3}\cdots\binom{N-N_1-\cdots-N_{M-1}}{N_M} =$$

$$= \frac{N!}{N_1!\,(N-N_1)!} \cdot \frac{(N-N_1)!}{N_2!\,(N-N_1-N_2)!} \cdots \frac{(N-N_1-\cdots-N_{M-1})!}{N_M!\,(N-N_1-\cdots-N_M)!} =$$

$$= \frac{N!}{N_1!\,N_2!\cdots N_M!}$$

(sfruttando il fatto che tutti i numeratori dei termini dal secondo in poi si semplificano con uno dei fattori del denominatore del termine precedente; inoltre, nell'ultimo termine, $N-N_1-\cdots-N_M \equiv 0$). Si può notare che l'ultimo termine della prima espressione, essendo $N - N_1 - \cdots - N_{M-1} = N_M$, vale sempre uno; cosa non sorprendente visto che, quando i primi $M-1$ gruppi sono stati scelti, anche l'ultimo risulta univocamente determinato.

Insomma il numero delle partizioni ordinate è uguale al numero delle permutazioni con ripetizione di N oggetti raggruppabili in M insiemi, composti rispettivamente da $N_1, N_2, \ldots, N_M$ oggetti indistinguibili tra loro, dato dalla formula (A.2)

Appendice B

L'errore della varianza

Può a volte essere utile valutare l'errore della stima della varianza ricavata da un campione di dati sperimentali. Facendo un esempio concreto, supponiamo di disporre di un ampio insieme di valutazioni della stessa grandezza fisica: $N \cdot M$ misure ripetute $x_1, x_2, \ldots, x_{N \cdot M}$. Dividiamo questi valori in M sottoinsiemi costituiti da N dati ciascuno, e per ognuno di questi M sottocampioni calcoliamo la media aritmetica dei dati; otterremo così M medie parziali, che indicheremo con i simboli $\bar{x}_1, \ldots, \bar{x}_M$.

Lo scopo di queste operazioni può essere quello di verificare che le medie di questi sottocampioni sono distribuite su un intervallo di valori più ristretto di quello su cui si distribuisce l'insieme dei dati originali: in sostanza, per verificare che le medie di N dati hanno errore quadratico medio inferiore a quello dei dati di partenza.

L'errore delle medie dei sottocampioni può essere stimato sperimentalmente calcolandone la varianza:

$$\sigma_{\bar{x}}{}^2 = \frac{1}{M-1} \sum_{i=1}^{M} \left(\bar{x}_i - \langle \bar{x} \rangle \right)^2 \qquad \textit{(sperimentale)}$$

intendendo con $\langle \bar{x} \rangle$ la media delle M medie parziali, che coinciderà necessariamente con la media complessiva dell'intero campione di $N \cdot M$ dati.

Questo valore può essere poi confrontato con quello previsto dalla teoria per la varianza della media di un gruppo di dati, allo scopo di verificare in pratica l'adeguatezza della teoria stessa; tale previsione teorica è come sappiamo data dal rapporto tra la varianza di ognuno dei dati che

contribuiscono alla media ed il numero dei dati stessi:

$$\sigma_{\bar{x}}{}^2 = \frac{\sigma^2}{N} \qquad \textit{(teorico)} \ .$$

Come stima di σ si può usare l'errore quadratico medio dell'insieme di tutti gli $N \cdot M$ dati; ma, naturalmente, perché il confronto tra questi due numeri abbia un significato, *occorre conoscere gli errori* da cui sia la valutazione sperimentale che la previsione teorica di $\sigma_{\bar{x}}$ sono affette.

Consideriamo (come già fatto precedentemente) una popolazione a media zero per semplificare i calcoli:

$$E(x) \equiv x^* = 0 \ ;$$

i risultati si potranno in seguito facilmente estendere ad una popolazione qualsiasi, tenendo presente il teorema di pagina 52 ed i ragionamenti conseguenti. La varianza di una qualsiasi variabile casuale x, indicata di seguito come $\mathrm{Var}(x)$, si può scrivere come

$$\mathrm{Var}(x) = E(x^2) - [E(x)]^2$$

e, usando questa formula per calcolare la varianza della varianza di un campione di N misure s^2, avremo

$$\mathrm{Var}(s^2) = E(s^4) - \left[E(s^2)\right]^2 \ .$$

Ora

$$\begin{aligned} s^4 &= \left[\frac{\sum_i x_i{}^2}{N} - \left(\frac{\sum_i x_i}{N} \right)^2 \right]^2 \\ &= \frac{1}{N^2} \left(\sum_i x_i{}^2 \right)^2 - \frac{2}{N^3} \left(\sum_i x_i{}^2 \right) \left(\sum_i x_i \right)^2 + \frac{1}{N^4} \left(\sum_i x_i \right)^4 \ . \end{aligned}$$

Sviluppiamo uno per volta i tre termini a secondo membro; per il primo risulta

$$\begin{aligned} \left(\sum_i x_i{}^2 \right)^2 &= \left(\sum_i x_i{}^2 \right) \left(\sum_j x_j{}^2 \right) \\ &= \sum_i \left(x_i{}^2 \sum_{j=i} x_j{}^2 \right) + \sum_i \left(x_i{}^2 \sum_{j \neq i} x_j{}^2 \right) \\ &= \sum_i x_i{}^4 + \sum_{\substack{i,j \\ j \neq i}} x_i{}^2 x_j{}^2 \\ &= \sum_i x_i{}^4 + 2 \sum_{\substack{i,j \\ j<i}} x_i{}^2 x_j{}^2 \ . \end{aligned}$$

La prima sommatoria comprende N addendi distinti; la seconda è estesa a tutte le possibili *combinazioni* dei valori distinti di i e j presi a due a due: è costituita quindi da

$$C_2^N = \frac{N\,(N-1)}{2}$$

addendi distinti.

Il fattore 2 che compare davanti ad essa è dovuto al fatto che una coppia di valori degli indici si presentava nella sommatoria su $i \neq j$ una volta come $x_i{}^2 x_j{}^2$ e un'altra come $x_j{}^2 x_i{}^2$, termini diversi per l'ordine ma con lo stesso valore. In definitiva, passando ai valori medi e tenendo conto dell'indipendenza statistica di x_i e x_j quando è $i \neq j$, risulta

$$E\left\{ \left(\sum_i x_i{}^2 \right)^2 \right\} = N\,E(x^4) + N\,(N-1)\,\left[E(x^2)\right]^2 \ .$$

Con simili passaggi, si ricava per il secondo termine

$$\begin{aligned} \left(\sum_i x_i{}^2 \right) \left(\sum_j x_j \right)^2 &= \left(\sum_i x_i{}^2 \right) \left(\sum_j x_j{}^2 + \sum_{\substack{j,k \\ j \neq k}} x_j x_k \right) \\ &= \sum_i x_i{}^4 + \sum_{\substack{i,j \\ i \neq j}} x_i{}^2 x_j{}^2 + \sum_{\substack{i,j \\ i \neq j}} x_i{}^3 x_j + \sum_{\substack{i,j,k \\ i \neq j \neq k}} x_i{}^2 x_j\, x_k \end{aligned}$$

dove gli indici aventi simboli diversi si intendono avere anche valori sempre diversi tra loro nelle sommatorie.

Il valore medio del terzo e del quarto termine si annulla essendo $E(x) = 0$; inoltre gli addendi nella prima sommatoria sono in numero di N e quelli nella seconda in numero di $N\,(N-1)/2$ e vanno moltiplicati per un fattore 2. Pertanto anche

$$E\left\{ \left(\sum_i x_i^2 \right) \left(\sum_i x_i \right)^2 \right\} = N\,E(x^4) + N\,(N-1)\,\left[E(x^2)\right]^2 \ .$$

Infine avremo, con la medesima convenzione sugli indici,

$$\begin{aligned} \left(\sum_i x_i \right)^4 &= \left(\sum_i x_i \right) \left(\sum_j x_j \right) \left(\sum_k x_k \right) \left(\sum_l x_l \right) \\ &= \sum_i x_i{}^4 + \sum_{\substack{i,j \\ i \neq j}} x_i{}^3 x_j + \sum_{\substack{i,j \\ i \neq j}} x_i{}^2 x_j{}^2 + \sum_{\substack{i,j,k \\ i \neq j \neq k}} x_i{}^2 x_j\, x_k + \sum_{\substack{i,j,k,l \\ i \neq j \neq k \neq l}} x_i\, x_j\, x_k\, x_l \ . \end{aligned}$$

I valori medi del secondo, quarto e quinto termine (che contengono potenze dispari delle x) sono nulli. Gli addendi nella prima sommatoria sono

in numero di N; nella terza vi sono $N\,(N-1)/2$ termini distinti: ma ciascuno appare in 6 modi diversi solo per l'ordine, corrispondenti al numero C_2^4 di combinazioni dei quattro indici originari i, j, k ed l presi a due a due. Allora

$$E\left(\sum_i x_i\right)^4 = N\,E(x^4) + 3\,N\,(N-1)\,\left[E(x^2)\right]^2 \;\;;$$

e, riprendendo la formule di partenza,

$$E(s^4) = \frac{(N-1)^2}{N^3}\,E(x^4) + \frac{(N-1)(N^2-2N+3)}{N^3}\,\left[E(x^2)\right]^2 \;\;.$$

Per il valore medio di s^2, già sappiamo come risulti per la varianza del campione

$$E(s^2) = \sigma^2 - {\sigma_{\bar{x}}}^2$$

inoltre

$$\sigma^2 \;=\; E\left\{(x-x^*)^2\right\} \;=\; E(x^2)$$

(essendo $x^* = 0$) e

$${\sigma_{\bar{x}}}^2 \;=\; E\left\{(\bar{x}-x^*)^2\right\} \;=\; \frac{\sigma^2}{N}$$

da cui abbiamo ottenuto a suo tempo la

$$E(s^2) \;=\; \frac{N-1}{N}\,\sigma^2 \;=\; \frac{N-1}{N}\,E(x^2) \;\;.$$

Per la varianza di s^2, che vogliamo determinare:

$$\begin{aligned}
\mathrm{Var}(s^2) &= E(s^4) - \left[E(s^2)\right]^2 \\
&= \frac{(N-1)^2}{N^3}\,E(x^4) + \\
&\quad + \left[\frac{(N-1)(N^2-2N+3)}{N^3} - \frac{(N-1)^2}{N^2}\right]\left[E(x^2)\right]^2 \\
&= \frac{(N-1)^2}{N^3}\,E(x^4) \;-\; \frac{(N-1)(N-3)}{N^3}\,\left[E(x^2)\right]^2 \\
&= \frac{N-1}{N^3}\,\left\{(N-1)\,E(x^4) \;-\; (N-3)\,\left[E(x^2)\right]^2\right\} \;\;.
\end{aligned}$$

Questa relazione ha validità generale. *Nel caso poi che la popolazione ubbidisca alla legge normale*, potremo calcolare il valore medio di x^4 usando la forma analitica della funzione di Gauss: per distribuzioni normali qualsiasi, i momenti di ordine pari rispetto alla media sono dati dalla formula (8.5), che qui ricordiamo:

$$\mu_{2k} \;=\; E\left\{\left[x-E(x)\right]^{2k}\right\} \;=\; \frac{(2k)!}{2^k\,k!}\,{\mu_2}^k \;=\; \frac{(2k)!}{2^k\,k!}\,\sigma^{2k} \;\;.$$

Per la varianza di s^2 se ne ricava

$$E(x^4) = 3\sigma^4$$

e, sostituendo,

$$\mathrm{Var}(s^2) \;=\; \frac{2(N-1)}{N^2}\,\left[E(x^2)\right]^2 \;=\; \frac{2(N-1)}{N^2}\,\sigma^4 \;\;;$$

insomma *l'errore quadratico medio della varianza s^2 del campione* vale

$$\sigma_{s^2} = \frac{\sqrt{2(N-1)}}{N}\,\sigma^2 \;\;.$$

La varianza, invece, della stima della varianza della popolazione

$$\sigma^2 = \frac{N}{N-1}\,s^2$$

vale

$$\mathrm{Var}(\sigma^2) \;=\; \left(\frac{N}{N-1}\right)^2\,\mathrm{Var}(s^2) \;=\; \frac{2}{N-1}\,\sigma^4 \;\;;$$

ed infine *l'errore quadratico medio della stima della varianza della popolazione* ricavata dal campione è

$$\boxed{\sigma_{\sigma^2} = \sqrt{\frac{2}{N-1}}\,\sigma^2}$$

Sottolineiamo ancora come queste formule che permettono di calcolare, per una popolazione *avente distribuzione normale*, gli errori quadratici medi sia della varianza di un campione di N misure che della stima della varianza della popolazione ricavata da un campione di N misure, siano *esatte*.

Se si vuole invece calcolare l'errore da attribuire agli *errori quadratici medi*, cioè alle quantità s e σ radici quadrate delle varianze di cui sopra, non è possibile dare delle formule esatte: la ragione ultima è che il valore medio di s non può essere espresso in forma semplice in termini di grandezze caratteristiche della popolazione.

Per questo motivo *è sempre meglio riferirsi ad errori di varianze* piuttosto che ad errori di scarti quadratici medi; comunque, in prima approssimazione, l'errore di σ si può ricavare da quello su σ^2 usando la formula di propagazione:

$$\mathrm{Var}(\sigma) \;\approx\; \left(\frac{1}{\frac{\mathrm{d}(\sigma^2)}{\mathrm{d}\sigma}}\right)^2 \mathrm{Var}(\sigma^2) \;=\; \frac{1}{4\,\sigma^2}\,\mathrm{Var}(\sigma^2) \;=\; \frac{\sigma^2}{2\,(N-1)} \;;$$

cioè

$$\boxed{\sigma_\sigma \approx \frac{\sigma}{\sqrt{2\,(N-1)}}} \qquad \text{(B.1)}$$

(il fatto che questa formula sia approssimata risulta chiaramente se si considera che la relazione tra σ^2 e σ è non lineare).

Una conseguenza dell'equazione (B.1) è che l'errore relativo di σ dipende *solo dal numero di misure*; diminuisce poi all'aumentare di esso, ma questa diminuzione è inversamente proporzionale alla radice quadrata di N e risulta perciò lenta.

In altre parole, per diminuire l'errore relativo di σ di un ordine di grandezza occorre aumentare il numero delle misure di *due* ordini di grandezza; σ_σ/σ è (circa) il 25% per 10 misure, il 7% per 100 misure ed il 2% per 1000 misure effettuate: e questo è sostanzialmente il motivo per cui, di norma, si scrive l'errore quadratico medio *dando per esso una sola cifra significativa.*

Due cifre significative *reali* per σ corrisponderebbero infatti ad un suo errore relativo compreso tra il 5% (se la prima cifra significativa di σ è 1, ad esempio $\sigma = 10 \pm 0.5$) e lo 0.5% ($\sigma = 99 \pm 0.5$); e presupporrebbero quindi che siano state effettuate almeno 200 misure nel caso più favorevole e quasi 20˙000 in quello più sfavorevole.

Appendice C

Covarianza e correlazione

C.1 La covarianza

Per due variabili casuali x ed y si definisce la *covarianza*, che si indica con uno dei due simboli $\mathrm{Cov}(x,y)$ o K_{xy}, nel seguente modo:

$$\begin{aligned}\mathrm{Cov}(x,y) &= E\Big\{[x - E(x)]\,[y - E(y)]\Big\} \\ &= E(xy) - E(x)\cdot E(y) \;\;.\end{aligned}$$

Per provare l'equivalenza delle due forme, basta osservare che[1]

$$\begin{aligned}\mathrm{Cov}(x,y) &= E\Big\{[x - E(x)]\,[y - E(y)]\Big\} \\ &= \textstyle\sum_{ij} P_{ij}\,[x_i - E(x)]\,[y_j - E(y)] \\ &= \textstyle\sum_{ij} P_{ij}\,x_i\,y_j \;-\; E(x)\sum_{ij} P_{ij}\,y_j \;-\; E(y)\sum_{ij} P_{ij}\,x_i \;+ \\ &\quad + E(x)\,E(y)\sum_{ij} P_{ij} \\ &= E(xy) \;-\; E(x)\sum_j q_j\,y_j \;-\; E(y)\sum_i p_i\,x_i \;+\; E(x)\,E(y) \\ &= E(xy) \;-\; E(x)\cdot E(y)\end{aligned}$$

ricordando alcune relazioni già ricavate nel capitolo 5, e valide per variabili casuali *qualunque*: in particolare, anche *non* statisticamente indipendenti.

[1] Nel seguito useremo per la varianza, per le probabilità di ottenere i vari valori x_i o y_j e così via, le stesse notazioni già introdotte nel capitolo 5.

È chiaro come per variabili *statisticamente indipendenti* la covarianza sia nulla: infatti per esse vale la

$$E(xy) \;=\; \sum_{ij} P_{ij}\, x_i\, y_j \;=\; \sum_{ij} p_i\, q_j\, x_i\, y_j \;=\; E(x)\cdot E(y) \;\;.$$

Non è però vero l'inverso: consideriamo ad esempio le due variabili casuali x ed $y = x^2$, ovviamente dipendenti l'una dall'altra: la loro covarianza vale

$$\mathrm{Cov}(x,y) \;=\; E(xy) - E(x)\cdot E(y) \;=\; E(x^3) - E(x)\cdot E(x^2)$$

ed è chiaramente nulla per qualunque variabile casuale x con distribuzione simmetrica rispetto allo zero; quindi l'annullarsi della covarianza è condizione *necessaria ma non sufficiente* per l'indipendenza statistica di due variabili casuali.

Possiamo ora calcolare la varianza delle combinazioni lineari di due variabili casuali qualunque, estendendo la formula già trovata nel capitolo 5 nel caso particolare di variabili statisticamente indipendenti; partendo ancora da due variabili x e y con media zero per semplificare i calcoli, per la loro combinazione lineare $z = ax + by$ valgono le:

$$\begin{aligned}
E(z) \;&=\; a\,E(x) + b\,E(y) \;=\; 0 \\
\mathrm{Cov}(x,y) \;&=\; E(xy) - E(x)\cdot E(y) \;=\; E(xy) \\
\mathrm{Var}(z) \;&=\; E\left\{\left[z - E(z)\right]^2\right\} \\
&=\; E(z^2) \\
&=\; E\left[(ax+by)^2\right] \\
&=\; \textstyle\sum_{ij} P_{ij}(a\,x_i + b\,y_j)^2 \\
&=\; a^2 \textstyle\sum_{ij} P_{ij}\, x_i^{\,2} \;+\; b^2 \sum_{ij} P_{ij}\, y_j^{\,2} \;+\; 2ab \sum_{ij} P_{ij}\, x_i\, y_j \\
&=\; a^2 \textstyle\sum_{i} p_i x_i^{\,2} \;+\; b^2 \sum_{j} q_j y_j^{\,2} \;+\; 2ab\, E(xy)
\end{aligned}$$

ed infine

$$\mathrm{Var}(z) \;=\; a^2\,\mathrm{Var}(x) + b^2\,\mathrm{Var}(y) + 2ab\,\mathrm{Cov}(x,y) \;\;. \qquad \text{(C.1)}$$

Questa si estende immediatamente a variabili casuali con media qualsiasi: introducendo ancora le variabili ausiliarie

$$\xi = x - E(x) \qquad \text{ed} \qquad \eta = y - E(y)$$

per le quali già sappiamo che vale la

$$E(\xi) \;=\; E(\eta) \;=\; 0$$

con le

$$\mathrm{Var}(\xi) = \mathrm{Var}(x) \qquad \text{e} \qquad \mathrm{Var}(\eta) = \mathrm{Var}(y) \;\;;$$

basta osservare infatti che vale anche la

$$\mathrm{Cov}(x,y) \;=\; E\left\{\left[x - E(x)\right]\left[y - E(y)\right]\right\} \;=\; \mathrm{Cov}(\xi,\eta) \;\;.$$

La (C.1) si può poi generalizzare, per induzione completa, ad una variabile z definita come combinazione lineare di un numero qualsiasi N di variabili casuali: si trova che, se

$$z = \sum_{i=1}^{N} a_i\, x_i$$

risulta

$$\mathrm{Var}(z) = \sum_i a_i^{\,2}\,\mathrm{Var}(x_i) + \sum_{\substack{i,j \\ j>i}} 2\,a_i\,a_j\,\mathrm{Cov}(x_i,x_j) \;\;. \qquad \text{(C.2)}$$

Per esprimere in modo compatto la (C.2), si ricorre in genere ad una notazione che usa la cosiddetta *matrice delle covarianze* delle variabili x; ovverosia una matrice quadrata $\boldsymbol{V}$ di ordine N, in cui il generico elemento V_{ij} è uguale alla covarianza delle variabili casuali x_i e x_j:

$$V_{ij} \;=\; \mathrm{Cov}(x_i,x_j) \;=\; E(x_i\cdot x_j) - E(x_i)\cdot E(x_j) \;\;. \qquad \text{(C.3)}$$

La matrice è ovviamente *simmetrica* ($V_{ij} = V_{ji}$); e, in particolare, gli elementi diagonali V_{ii} valgono

$$V_{ii} \;=\; E(x_i^{\,2}) - \left[E(x_i)\right]^2 \;\equiv\; \mathrm{Var}(x_i) \;\;.$$

Consideriamo poi le a_i come le N componenti di un *vettore* $\boldsymbol{A}$ di dimensione N (che possiamo concepire come una matrice rettangolare di N righe ed una colonna); ed introduciamo la matrice *trasposta* di $\boldsymbol{A}$, $\widetilde{\boldsymbol{A}}$, che è una matrice rettangolare di una riga ed N colonne i cui elementi valgono

$$\widetilde{A}_i = A_i \;\;.$$

Possiamo allora scrivere la (C.2) nella forma

$$\mathrm{Var}(z) = \sum_{i,j} \widetilde{A}_i \, V_{ij} \, A_j$$

(la simmetria di $\boldsymbol{V}$ e quella tra $\widetilde{\boldsymbol{A}}$ ed $\boldsymbol{A}$ produce, nello sviluppo delle sommatorie, il fattore 2 che moltiplica le covarianze); o anche, ricordando le regole del prodotto tra matrici,

$$\boxed{\mathrm{Var}(z) \;=\; \widetilde{\boldsymbol{A}}\,\boldsymbol{V}\,\boldsymbol{A}}$$

Si può poi facilmente dimostrare il seguente teorema, che ci sarà utile più avanti:

> TEOREMA: *due differenti combinazioni lineari delle stesse variabili casuali sono sempre correlate.*

Infatti, dette A e B le due combinazioni lineari:

$$A = \sum_{i=1}^{N} a_i \, x_i \qquad \Rightarrow \qquad E(A) = \sum_{i=1}^{N} a_i \, E(x_i)$$

$$B = \sum_{j=1}^{N} b_j \, x_j \qquad \Rightarrow \qquad E(B) = \sum_{j=1}^{N} b_j \, E(x_j)$$

abbiamo che la covarianza di A e B vale

$$\begin{aligned}
\mathrm{Cov}(A,B) &= E\Big\{\big[A - E(A)\big]\,\big[B - E(B)\big]\Big\} \\
&= E\left\{\sum_{i,j} a_i \, b_j \,\big[x_i - E(x_i)\big]\,\big[x_j - E(x_j)\big]\right\} \\
&= \sum_{i,j} a_i \, b_j \, E\Big\{\big[x_i - E(x_i)\big]\,\big[x_j - E(x_j)\big]\Big\} \\
&= \sum_{i} a_i \, b_i \; \mathrm{Var}(x_i) + \sum_{\substack{i,j \\ i \neq j}} a_i \, b_j \; \mathrm{Cov}(x_i, x_j)
\end{aligned}$$

e non è in genere nulla. In forma matriciale e con ovvio significato dei simboli,

$$\mathrm{Cov}(A,B) = \widetilde{\boldsymbol{A}}\,\boldsymbol{V}\,\boldsymbol{B} \;\; .$$

È da notare come A e B siano di norma sempre correlate *anche se le variabili di partenza x_i sono tutte tra loro statisticamente indipendenti*: in questo caso infatti tutti i termini non diagonali della matrice delle covarianze si annullano, e risulta

$$\mathrm{Cov}(A,B) = \sum_{i=1}^{N} a_i \, b_i \; \mathrm{Var}(x_i) \;\; . \tag{C.4}$$

C.2 La correlazione lineare

Per due variabili casuali qualunque si definisce poi il *coefficiente di correlazione lineare* $\mathrm{Corr}(x,y)$ (anche indicato col simbolo r_{xy}, o semplicemente come r) nel modo seguente:

$$r_{xy} \;\equiv\; \mathrm{Corr}(x,y) \;=\; \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\;\mathrm{Var}(y)}} \;=\; \frac{\mathrm{Cov}(x,y)}{\sigma_x \, \sigma_y} \;\; .$$

Il coefficiente di correlazione di due variabili è ovviamente adimensionale; è nullo quando le variabili stesse sono statisticamente indipendenti (visto che è zero la loro covarianza); ed è comunque compreso tra i due limiti -1 e $+1$. Che valga quest'ultima proprietà si può dimostrare calcolando dapprima la varianza di una variabile casuale ausiliaria z definita attraverso la relazione $z = \sigma_y \, x - \sigma_x \, y$, ed osservando che essa deve essere una quantità non negativa:

$$\begin{aligned}
\mathrm{Var}(z) &= \sigma_y{}^2 \; \mathrm{Var}(x) + \sigma_x{}^2 \; \mathrm{Var}(y) - 2 \, \sigma_x \, \sigma_y \; \mathrm{Cov}(x,y) \\
&= 2 \; \mathrm{Var}(x) \; \mathrm{Var}(y) - 2 \, \sigma_x \, \sigma_y \; \mathrm{Cov}(x,y) \\
&\geq 0 \;\; ;
\end{aligned}$$

da cui

$$\mathrm{Corr}(x,y) \leq 1 \;\; .$$

Poi, compiendo analoghi passaggi su un'altra variabile definita stavolta come $z = \sigma_y \, x + \sigma_x \, y$, si troverebbe che deve essere anche $\mathrm{Corr}(x,y) \geq -1$.

Se il coefficiente di correlazione lineare raggiunge uno dei due valori estremi ± 1, risulta $\mathrm{Var}(z) = 0$; e dunque deve essere

$$z \;=\; \sigma_y \, x \mp \sigma_x \, y \;=\; \text{costante}$$

cioè x ed y devono essere legati da una relazione funzionale *di tipo lineare*.

Vale anche l'inverso: partendo infatti dall'ipotesi che le due variabili siano legate da una relazione lineare data da $y = a + bx$, con b finito e non nullo, ne consegue che:

$$\begin{aligned}
E(y) &= a + b\,E(x) \\
\mathrm{Var}(y) &= b^2\,\mathrm{Var}(x) \\
E(xy) &= E(a\,x + b\,x^2) \\
&= a\,E(x) + b\,E(x^2) \\
\mathrm{Cov}(x,y) &= E(xy) - E(x)\cdot E(y) \\
&= a\,E(x) + b\,E(x^2) - E(x)[a + b\,E(x)] \\
&= b\left\{E(x^2) - [E(x)]^2\right\} \\
&= b\cdot\mathrm{Var}(x) \\
\mathrm{Corr}(x,y) &= \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\;\mathrm{Var}(y)}} \\
&= \frac{b\;\mathrm{Var}(x)}{\sqrt{b^2[\mathrm{Var}(x)]^2}} \\
&= \frac{b}{|b|} \\
&= \pm 1 \ .
\end{aligned}$$

Il segno del coefficiente di correlazione è quello del coefficiente angolare della retta. Sono da notare due cose: innanzi tutto il rapporto $b/|b|$ perde significato quando $b = 0$ o quando $b = \infty$, cioè quando la retta è parallela ad uno degli assi coordinati: in questi casi (x = costante o y = costante) una delle due grandezze non è in realtà una variabile casuale, e l'altra è dunque indipendente da essa; è facile vedere che tanto il coefficiente di correlazione tra x e y quanto la covarianza valgono zero, essendo $E(xy) \equiv E(x)\cdot E(y)$ in questo caso.

Anche quando esiste una relazione funzionale esatta tra x e y, se questa non è rappresentata da una funzione lineare il coefficiente di correlazione non raggiunge i valori estremi ± 1; per questa ragione appunto esso si chiama più propriamente "coefficiente di correlazione *lineare*".

C.3 Propagazione degli errori per variabili correlate

Vediamo ora come si può ricavare una formula di propagazione per gli errori (da usare in luogo dell'equazione (10.2) che abbiamo incontrato a pagina 164) se le grandezze fisiche misurate direttamente non sono tra loro statisticamente indipendenti; nel corso di questo paragrafo continueremo ad usare la notazione già introdotta nel capitolo 10.

Consideriamo una funzione F di N variabili, $F = F(x_1, x_2, \ldots, x_N)$; ed ammettiamo che sia lecito svilupparla in serie di Taylor nell'intorno del punto $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N)$ trascurando i termini di ordine superiore al primo (questo avviene, come sappiamo, o se gli errori di misura sono piccoli o se F è lineare rispetto a tutte le variabili). Tenendo presente il teorema di pagina 52, ed applicando alla formula dello sviluppo

$$F(x_1, x_2, \ldots, x_N) \;\approx\; F(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N) \;+\; \sum_{i=1}^{N} \frac{\partial F}{\partial x_i}\,(x_i - \bar{x}_i)$$

l'equazione (C.2), otteniamo

$$\mathrm{Var}(F) \;\approx\; \sum_i \left(\frac{\partial F}{\partial x_i}\right)^2 \mathrm{Var}(x_i) + 2\sum_{\substack{i,j \\ j>i}} \frac{\partial F}{\partial x_i}\,\frac{\partial F}{\partial x_j}\;\mathrm{Cov}(x_i, x_j) \ . \qquad \text{(C.5)}$$

Per esprimere in modo compatto la (C.5), si può ricorrere ancora alla matrice delle covarianze $\boldsymbol{V}$ delle variabili x_i; ricordandone la definizione (data dall'equazione (C.3) a pagina 257) ed introducendo poi un vettore $\boldsymbol{F}$ di dimensione N di componenti

$$F_i \;=\; \frac{\partial F}{\partial x_i}$$

ed il suo trasposto $\widetilde{\boldsymbol{F}}$, la (C.5) si può riscrivere nella forma

$$\mathrm{Var}(F) \;=\; \sum_{i,j} \widetilde{F}_i\, V_{ij}\, F_j$$

ossia

$$\boxed{\mathrm{Var}(F) \;=\; \widetilde{\boldsymbol{F}}\,\boldsymbol{V}\,\boldsymbol{F}}$$

C.4 Applicazioni all'interpolazione lineare

Riprendiamo adesso il problema dell'interpolazione lineare, già discusso nel capitolo 11: si sia cioè compiuto un numero N di misure indipendenti di coppie di valori di due grandezze fisiche x e y, tra le quali si ipotizza che esista una relazione funzionale di tipo lineare data da $y = a + bx$. Supponiamo inoltre che siano valide le ipotesi esposte nel paragrafo 11.4.1; in particolare che le x_i siano prive di errore, e che le y_i siano affette da errori normali e tutti uguali tra loro.

C.4.1 Riscrittura delle equazioni dei minimi quadrati

Sebbene i valori della x siano scelti dallo sperimentatore e privi di errore, e non siano pertanto variabili casuali in senso stretto; e sebbene la variabilità delle y sia dovuta non solo agli errori casuali di misura ma anche alla variazione della x, introduciamo ugualmente (in maniera *puramente formale*) le medie e le varianze degli N valori x_i e y_i, date dalle espressioni

$$\bar{x} = \frac{\sum_i x_i}{N} \qquad \text{e} \qquad \mathrm{Var}(x) = \frac{\sum_i (x_i - \bar{x})^2}{N} = \frac{\sum_i {x_i}^2}{N} - \bar{x}^2$$

(e simili per la y); e la covarianza di x e y, data dalla

$$\mathrm{Cov}(x, y) = \frac{\sum_i x_i y_i}{N} - \bar{x}\bar{y} \ .$$

Queste grandezze permettono di riscrivere le equazioni (11.9) risolutive del problema dell'interpolazione lineare per un insieme di dati, che abbiamo già incontrato a pagina 181, nella forma

$$\begin{cases} a & + & b\,\bar{x} & = & \bar{y} \\ a\,\bar{x} & + & b\,[\mathrm{Var}(x) + \bar{x}^2] & = & \mathrm{Cov}(x, y) + \bar{x}\bar{y} \end{cases}$$

La prima equazione intanto implica che la retta interpolante deve passare per il punto $(\bar{x}, \bar{y})$ le cui coordinate sono le medie dei valori misurati delle due variabili in gioco; poi, ricavando da essa $a = \bar{y} - b\bar{x}$ e sostituendo nella seconda equazione, dopo aver semplificato alcuni termini si ottiene la soluzione per l'altra incognita:

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} \equiv \mathrm{Corr}(x, y)\sqrt{\frac{\mathrm{Var}(y)}{\mathrm{Var}(x)}} \qquad \text{(C.6)}$$

e la retta interpolante ha quindi equazione

$$y = a + bx = (\bar{y} - b\bar{x}) + bx$$

o anche

$$(y - \bar{y}) = b\,(x - \bar{x})$$

(in cui b ha il valore (C.6)). Introduciamo ora le due variabili casuali ausiliarie $\xi = x - \bar{x}$ e $\eta = y - \bar{y}$, per le quali valgono le

$$\bar{\xi} = 0 \qquad \text{e} \qquad \mathrm{Var}(\xi) = \mathrm{Var}(x)$$

(con le analoghe per η ed y), ed inoltre la

$$\mathrm{Cov}(\xi, \eta) = \mathrm{Cov}(x, y)$$

ed indichiamo poi con $\hat{y}_i$ il valore della y sulla retta interpolante in corrispondenza dell'ascissa x_i:

$$\hat{y}_i = a + bx_i = \bar{y} + b\,(x_i - \bar{x}) \qquad \text{(C.7)}$$

e con δ_i la differenza $\hat{y}_i - y_i$. Le differenze δ_i prendono il nome di *residui*, e di essi ci occuperemo ancora più avanti; risulta che

$$\begin{aligned}
\sum_i {\delta_i}^2 &= \sum_i \left\{[\bar{y} + b\,(x_i - \bar{x})] - y_i\right\}^2 \\
&= \sum_i (b\xi_i - \eta_i)^2 \\
&= b^2 \sum_i {\xi_i}^2 + \sum_i {\eta_i}^2 - 2b \sum_i \xi_i \eta_i \\
&= N\,b^2\ \mathrm{Var}(\xi) + N\ \mathrm{Var}(\eta) - 2Nb\ \mathrm{Cov}(\xi, \eta) \\
&= N\,b^2\ \mathrm{Var}(x) + N\ \mathrm{Var}(y) - 2Nb\ \mathrm{Cov}(x, y) \\
&= N\ \mathrm{Var}(x)\left[\frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}\right]^2 + N\ \mathrm{Var}(y) - 2N\,\frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}\ \mathrm{Cov}(x, y) \\
&= N\left\{\mathrm{Var}(y) - \frac{[\mathrm{Cov}(x, y)]^2}{\mathrm{Var}(x)}\right\} \\
&= N\ \mathrm{Var}(y)\,(1 - r^2)
\end{aligned}$$

in cui r è il coefficiente di correlazione lineare calcolato usando, sempre solo formalmente, i campioni dei valori misurati delle x e delle y.

Visto che quest'ultimo, nel calcolo dell'interpolazione lineare fatto con le calcolatrici da tasca, viene in genere dato come sottoprodotto dell'algoritmo,